Wissenschaftliches Gutachten

Statistisch-methodische Bewertung von Modellierungsstudien zu den Effekten einer Zuckersteuer

Dr. Katharina Schüller,

Accredited European Statistician (AEUStat)

Wissenschaftliche Beratung:

Prof. Dr. Ralf Münnich

Prof. Dr. Danny Pfeffermann

Statistisch-methodische Bewertung von Modellierungsstudien zu den Effekten einer Zuckersteuer

Dr. Katharina Schüller, AEUStat[1]

Wissenschaftliche Beratung:
Prof. Dr. Ralf Münnich[2], Prof. Dr. Danny Pfeffermann[3]

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

München

Februar 2025

[1]Geschäftsführung STAT-UP GmbH, Vorstandsmitglied Deutsche Statistische Gesellschaft
[2]Professur für Wirtschafts- und Sozialstatistik der Universität Trier
[3]Professur für Sozialstatistik der Universität Southampton
[4]Mitarbeiter STAT-UP GmbH

STAT-UP

Statistical Consulting & Data Science GmbH

Augustenstr. 5

D - 80333 München

Geschäftsf.: Dr. Katharina Schüller, AEUStat

Telefon: +49 (89) 34 077 451

Telefax: +49 (89) 34 077 453

E-Mail: info@stat-up.com

Web: www.stat-up.com

Hinweis: Zur besseren Lesbarkeit wird in diesem Gutachten das generische Maskulinum verwendet. Die in dieser Arbeit verwendeten Personenbezeichnungen beziehen sich – sofern nicht anders kenntlich gemacht – auf alle Geschlechter.

Executive summary

Das vorliegende Gutachten analysiert und bewertet eine Auswahl[5] von Modellierungsstudien zu den Effekten einer Zuckersteuer aus statistisch-methodischer Perspektive. Modellierungsstudien stellen ein wichtiges Instrument dar, um (potenzielle) Auswirkungen regulatorischer Maßnahmen abzuschätzen. Sie können theoretisch geeignet sein, eine Evidenzbasis für Regulierung zu schaffen – allerdings nur, sofern sie auf präzisen Datengrundlagen sowie belastbaren und sorgfältig validierten Annahmen beruhen.

> Das Gutachten kommt zu dem Schluss, dass keine der evaluierten Modellierungsstudien zu den Effekten einer Zuckersteuer aus statistisch-methodischer Perspektive den hohen Qualitätsstandards genügt, die für evidenzbasierte politische Entscheidungen erforderlich sind.

Ein wesentlicher Grund dafür ist, dass Datengrundlagen von hinreichender Qualität nicht vollumfänglich vorhanden sind. In Ermangelung ausreichend detaillierter Datensätze müssen zahlreiche Annahmen getroffen werden, welche die Realität nicht immer adäquat abbilden. Dadurch sind die Ergebnisse der untersuchten Studien mit erheblichen Unsicherheiten behaftet.

Zentrale Erkenntnisse zu Modellierungsstudien

Auf Grundlage der bestehenden Forschung zu Modellierungsstudien und der in diesem Gutachten erfolgten Evaluationen der Einzelstudien lassen sich folgende zentrale Erkenntnisse zu allgemeinen methodischen Limitationen von Modellierungsstudien ableiten:

1. Modellierungsstudien zur Untersuchung (hypothetischer) regulatorischer Maßnahmen, bspw. die Einführung einer Zuckersteuer in Deutschland, sind kein eigenständiger Beleg kausaler Zusammenhänge; vielmehr wird das Vorliegen von Kausalwirkungen zwischen Einfluss- und Zielgrößen bei den verwendeten Modellierungsverfahren ex ante angenommen.

2. Nicht nur das Vorliegen kausaler Wirkmechanismen, sondern auch deren Quantifizierung basiert auf Annahmen, bspw. der Übertragbarkeit früherer empirischer Untersuchungen. Solche Annahmen sind zwangsläufig mit Unsicherheiten der Modellierungsergebnisse verbunden. Jede Entscheidung für eine bestimmte Annahme ist gleichzeitig eine Entscheidung gegen alternative Möglichkeiten. Modellierungsergebnisse werden daher maßgeblich von den zugrunde liegenden Annahmen beeinflusst.

[5] Die Auswahl umfasst folgende Studien:

- Emmert-Fees et al. (2023), Deutschland: Projected health and economic impacts of sugar-sweetened beverage taxation in Germany: A cross-validation modelling study.
- Schwendicke und Stolpe (2017), Deutschland: 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 modelling study.
- Gračner et al. (2022), Mexiko: Changes in weight-related outcomes among adolescents following consumer price increases of taxed sugar-sweetened beverages.
- Basto-Abreu et al. (2019), Mexiko: Cost-effectiveness of the sugar-sweetened beverage excise tax in Mexico.

3. Die Aussagekraft von Modellierungsstudien ist stark von der Verfügbarkeit und Qualität der benötigten empirischen Eingangsdaten abhängig. Modellierungsstudien beziehen diese in der Regel aus externen Quellen und setzen, häufig ohne explizite Prüfung, deren Repräsentativität und Genauigkeit voraus. Beides ist oftmals nicht gegeben.

4. Ungenaue, verzerrte und/oder nicht repräsentative Daten sowie falsche oder zu stark vereinfachende Annahmen über die zugrunde liegenden Wirkmechanismen führen zu Modellergebnissen, die zwar innerhalb des Modells konsistent sein mögen, aber die Realität nicht korrekt abbilden.

5. Selbst in Fällen, in denen die Annahmen dem Grunde nach zutreffen und die Eingangsdaten repräsentativ sind, bleiben die Ergebnisse von Modellierungsstudien immer mit Unsicherheiten behaftet, die erheblich sein können und bei Fragestellungen der öffentlichen Gesundheit häufig ignoriert werden.

Um Fehlschlüsse mit potenziell weitreichenden Konsequenzen zu verhindern, müssen diese methodischen Einschränkungen von Modellierungsstudien und die daraus resultierenden Unsicherheiten in ihren Ergebnissen unmissverständlich offengelegt und in öffentlichen sowie politischen Debatten zwingend berücksichtigt werden.

Zentrale Limitationen der evaluierten Studien

Die zentralen methodischen Limitationen von Modellierungsstudien werden nachfolgend anhand der sechs evaluierten Studien zu den Effekten einer Zuckersteuer verdeutlicht. Die Analyse erfolgt in drei Bewertungskategorien: In der Kategorie *Datengrundlage* wird die Qualität der verwendeten Daten als fundamentales Rohmaterial von Modellierungsstudien bewertet; in der Kategorie *Modellspezifikation* werden die eingesetzten Modelle als analytische Werkzeuge zur Datenverarbeitung untersucht; die Kategorie *Ergebniskommunikation* widmet sich der Darstellung und Einordnung der Modellergebnisse.

Beurteilung der Datengrundlage

Die Stichproben in den evaluierten Studien repräsentieren die interessierende Population nicht immer angemessen. Zwei Studien basieren auf realen Stichproben, die strukturell von der interessierenden Gesamtpopulation abweichen: Bei Rogers, Cummins et al. sind übergewichtige und adipöse Mädchen unterrepräsentiert; Gračner et al. wiederum untersuchen ausschließlich Jugendliche, die in städtischen Gebieten leben und bei einem bestimmten Unternehmen krankenversichert sind. Die Ergebnisse beider Studien sind daher nicht verlässlich. Die übrigen vier Studien verwenden synthetische Stichproben, abgeleitet aus aktuellen amtlichen Bevölkerungsstatistiken (Emmert-Fees et al.; Cobiac et al.; Basto-Abreu et al., Schwendicke und Stolpe), was eine grundsätzliche strukturelle Übereinstimmung mit der interessierenden Gesamtpopulation nahelegt. Allerdings stützt sich die Analyse von Schwendicke und Stolpe aus dem Jahr 2017 auf Bevölkerungsdaten von 2012, die aufgrund demografischer Veränderungen, etwa durch Migration, von der heutigen deutschen Bevölkerung abweichen können.

Die weiteren Eingangsdaten weisen in allen evaluierten Studien teils erhebliche qualitative Mängel auf, die zu verzerrten Modellparametern und damit unzuverlässigen Schlussfolgerungen führen. Häufig sind die Eingangsdaten nur eingeschränkt auf den Anwendungskontext der Studie **übertragbar**, etwa weil sie **aus anderen Ländern** stammen, die aufgrund länderspezifischer Unterschiede nicht zur untersuchten

Stichprobe passen (Emmert-Fees et al.: Prävalenzen von koronaren Herzerkrankungen und Schlaganfällen aus Großbritannien; Schwendicke und Stolpe: Preiselastizitäten aus den USA; Basto-Abreu et al.: Umrechnungsfaktoren von Konsum- in Gewichtsänderungen von Kindern aus den Niederlanden). Zudem basieren manche Datensätze auf **spezifischen Subpopulationen** und sind daher nicht auf die Gesamtbevölkerung verallgemeinerbar. Dennoch verwenden Basto-Abreu et al. pauschal Daten von Lehrerinnen, um Konsum- in Gewichtsänderungen bei Erwachsenen allgemein umzurechnen. Oft sind Daten außerdem nicht in ausreichender **Detailtiefe** verfügbar, um die Heterogenität der Bevölkerung differenziert abzubilden. Bspw. nehmen Basto-Abreu et al. pauschalisierte Konsumveränderungen über alle Alters- und Geschlechtsgruppen hinweg an. Darüber hinaus bilden die verwendeten Daten die **Realität** nicht immer adäquat ab. So beschränken sich Einkaufsdaten zuckergesüßter Getränke in mehreren Studien ausschließlich auf den stationären Einzelhandel, während gastronomische Vertriebsstätten unberücksichtigt bleiben (z. B. bei Cobiac et al., Gračner et al. und Basto-Abreu et al.), was zu verzerrten Ergebnissen führt. Außerdem sind die Daten teilweise **veraltet**. So stammen die Konsumdaten bei Schwendicke und Stolpe sowie teilweise auch bei Emmert-Fees et al. aus der Nationalen Verzehrsstudie II aus den Jahren 2005-2007; die Daten zu Krankheitskosten bei Emmert-Fees et al. reichen teilweise sogar bis 1999 zurück. Zudem basieren die Daten teils auf **kleinen Stichproben**, was ihre Zuverlässigkeit einschränkt. So basiert etwa die Schätzung der Produktivitätseinbußen aufgrund von Schlaganfällen bei Emmert-Fees et al. auf nur 151 Patienten. Ein weiteres Problem ist die geringere Evidenzbasis **nicht-experimentell erhobener Daten**, auf die viele Studien mangels Verfügbarkeit experimenteller Daten angewiesen sind (z. B. Umrechnungsfaktoren von Konsum in Gewichtsänderungen bei Emmert-Fees et al. und Basto-Abreu et al.). Werden Daten **aus Metastudien** genutzt (z. B. Gesundheitsdaten bei Cobiac et al.), bleibt die Qualität der zugrunde liegenden Primärdaten oft unklar, insbesondere in Bezug auf mögliche Verzerrungen und ihre Passung zum Studienkontext. Daten basierend auf **Selbstauskünften** (z. B. Schwendicke und Stolpe: Körperindizes; Cobiac et al.: Getränkeeinkäufe und Gesundheitsdaten; Basto-Abreu et al.: Basiskonsum) sind fehleranfällig und können Unsicherheiten in den Ergebnissen zur Folge haben. Darüber hinaus sind Datensätze teilweise **unvollständig**, bei Emmert-Fees et al. etwa fehlen Risikofaktoren für koronare Herzkrankheiten und Schlaganfälle. Verfahren zum Ausgleich fehlender Daten (Emmert-Fees et al.: Schätzung; Basto-Abreu et al.: Extrapolation sowie Ergänzung mit Daten aus anderen Studien) gewährleisten nicht immer die Validität. Schließlich sind einige Datensätze untereinander **inkompatibel** (z. B. Konsumdaten bei Emmert-Fees et al.; Konsum- und Körperdaten bei Schwendicke und Stolpe), was zu Abweichungen in den Ergebnissen, etwa bei spezifischen Altersgruppen, führen kann. Zusammenfassend basieren alle Studien auf teilweise verzerrten Parameter. Eine Modellierung auf einer derart fragilen Grundlage – Emmert-Fees et al. bspw. verwendet zur Parametrisierung des Modells Daten aus 36 verschiedenen Quellen – lässt keine belastbaren Schlussfolgerungen zu.

Beurteilung der Modellspezifikation

Die eingesetzten Modellierungsverfahren eignen sich nicht zum Nachweis von Kausalität und bieten entsprechend keine Grundlage, um die Steuer *ursächlich* für die simulierten Effekte verantwortlich zu machen. Die verwendeten Modellierungsverfahren (Mikrosimulation, Monte-Carlo-Simulation, Kohortensimulation, Zeitreihenanalyse und multivariate Regressionsanalyse) sind zwar grundsätzlich geeignet, um reale Situationen mit hypothetischen Szenarien zu vergleichen. Allerdings wird Kausalität bei allen Verfahren ex ante bei der Modellkonstruktion unterstellt, jedoch nicht durch das Modell selbst nachgewiesen. Ob ein

kausaler Zusammenhang für bestimmte Modellannahmen zuvor belegt wurde, wird in keiner der Studien explizit thematisiert.

Die häufig unzureichende Operationalisierung der Einflussgrößen führt zu einer starken Pauschalisierung der Ergebnisse. Bspw. wird Konsum häufig vereinfachenderweise pauschal mit Kauf gleichgesetzt (Cobiac et al.; Basto-Abreu et al.). Zudem bleibt die Heterogenität des Konsumverhaltens hinsichtlich alters- und geschlechtsspezifischer Unterschiede häufig unberücksichtigt (Emmert-Fees et al.; Schwendicke und Stolpe; Cobiac et al.; Basto-Abreu et al.).

Die vereinfachenden Annahmen zur Beziehung zwischen Einfluss- und Zielgrößen in allen evaluierten Studien lassen darauf schließen, dass die Modellergebnisse realen Zusammenhänge nicht präzise abbilden. Die Zusammenhänge werden meist linear und ohne hinreichende empirische Belege modelliert. So wird etwa ein direkter Zusammenhang zwischen Konsum- und Gewichtsveränderungen angenommen (Schwendicke und Stolpe; Cobiac et al.), wobei physiologische Unterschiede, körperliche Aktivität und kalorische Kompensation weitgehend ignoriert werden, was der multifaktoriellen Entstehung von Übergewicht und Adipositas widerspricht. Studien, die gesundheitsbezogene Effekte modellieren (Emmert-Fees et al.; Schwendicke und Stolpe; Cobiac et al.; Basto-Abreu et al.), treffen Annahmen zur zeitlichen Wirksamkeit der Steuer, berücksichtigen aber weder Schwankungen noch langfristige Anpassungseffekte im Konsumverhalten. Auch die übrigen Studien treffen stark vereinfachende Annahmen: Gračner et al. unterstellen ohne empirische Belege einen linearen Zusammenhang zwischen Preisänderungen und Körpermaßen; Rogers, Cummins et al. setzen einen linear fortgesetzten Trend der Adipositasprävalenz voraus. Die realen Zusammenhänge werden dadurch nicht präzise abgebildet.

Alle evaluierten Studien berücksichtigen relevante Einflussfaktoren nur unzureichend, was das Risiko fehlerhafter oder stark verallgemeinerter Schlussfolgerungen über die Steuereffekte birgt. Substitutionseffekte und der sozioökonomische Status werden in keiner Studie angemessen berücksichtigt. Ebenso werden parallele Maßnahmen wie Aufklärungskampagnen oder politische Reformen vernachlässigt. Ein Beispiel ist die Studie von Basto-Abreu et al., welche die Auswirkungen der Besteuerung zuckergesüßter Getränke in Mexiko untersucht. Ihre Simulationsergebnisse basieren auf der Annahme, dass der Konsum dieser Getränke durch die Steuer sinkt. Allerdings zeigen andere Untersuchungen, dass sich unter Einbeziehung weiterer Lebensmittel der Gesamtkalorien-Einkauf in Mexiko nicht signifikant veränderte – was die Verlässlichkeit dieser Ergebnisse von Basto-Abreu et al. erheblich einschränkt.

Beurteilung der Ergebniskommunikation

Die Variabilität der Ergebnisse wird nicht immer transparent dargestellt und eine unzureichende Kontextualisierung kann zu Überinterpretationen der simulierten Effekte führen. Konfidenzintervalle (z. B. bei Emmert-Fees et al. oder Rogers, Cummins et al.) erfassen nur zufällige, aber keine nicht-zufälligen Fehler, sodass sie die Variabilität der simulierten Ergebnisse nur bedingt abbilden. In den evaluierten Studien fehlt teilweise eine angemessene Kontextualisierung der Effekte. Bspw. geben Emmert-Fees et al. die geschätzte Gesundheitskostenreduktion in den drei Szenarien mit etwa drei Milliarden Euro an. Dies entspricht weniger als 0,04 % der Gesamtkosten – eine Information, die von den Studienautoren nicht kommuniziert wird, wodurch der Effekt größer erscheint als er anteilig tatsächlich ist. Ähnlich tritt diese Problematik auch in den Studien von

Schwendicke und Stolpe, Rogers, Cummins et al., Cobiac et al. sowie Basto-Abreu et al. auf.

Unsicherheiten und mögliche Verzerrungen der simulierten Ergebnisse werden in allen Studien insgesamt nur unzureichend adressiert. Systematische Analysen der (globalen) Unsicherheiten in den Modellierungsergebnissen fehlen weitgehend und beschränken sich meist auf Sensitivitätsanalysen. Diese betrachten in allen Studien lediglich Variationen einzelner Parameter, ohne zentrale Modellannahmen wie wesentliche Einflussfaktoren oder die Annahme linearer Beziehungen zu variieren. Sie sind daher nur bedingt aussagekräftig und aufgrund der mangelhaften Datengrundlage in allen Studien ohnehin fragwürdig.

Die Modellergebnisse werden oft ohne ausreichende Berücksichtigung von Limitationen und Unsicherheiten eingeordnet, wodurch ihre Verlässlichkeit fraglich bleibt und teils übergeneralisierten Schlussfolgerungen begünstigt werden. Zwar werden Studienlimitationen meist breit thematisiert, jedoch oft nur oberflächlich behandelt und selten in Bezug auf ihre potenziellen Auswirkungen auf die Ergebnisse reflektiert. So stellt bspw. Emmert-Fees et al. die gestaffelte Herstellerabgabe als optimale Steuerform dar, lässt dabei jedoch die in einer Sensitivitätsanalyse aufgezeigten Unsicherheiten unberücksichtigt. Rogers, Cummins et al. sowie Gračner et al. weisen Effekte lediglich für bestimmte Subgruppen, etwa Mädchen, aus, verallgemeinern diese jedoch auch auf Jungen. Besonders problematisch sind fehlerhafte kausale Schlussfolgerungen: Schwendicke und Stolpe, Cobiac et al. sowie Basto-Abreu et al. folgern, dass eine Steuer Adipositas reduzieren würde, obwohl dies nicht durch die verwendeten Modelle belegt wird. Solche irreführenden Schlussfolgerungen sind besonders kritisch, da sie häufig von Medien aufgegriffen und zur Argumentation für regulatorische Maßnahmen herangezogen werden.

Inhaltsverzeichnis

1 Einleitung

Problemstellung

In Diskussionen über die Einführung einer *Zuckersteuer* in Deutschland (z. B. Appelhans, 2024; Heitmann, 2024; Magoley, 2024; Werny, 2024) wird unter anderem auf Modellierungsstudien verwiesen, die potenzielle Effekte einer solchen Steuer in Deutschland simulieren oder die Wirksamkeit der Steuer in Ländern evaluieren, welche die Maßnahme bereits eingeführt haben. Es wurde bislang jedoch nicht systematisch untersucht, inwieweit diese Studien methodisch den erforderlichen Qualitätsstandards entsprechen, um als Evidenzgrundlage für fundierte Diskussionen oder politische Entscheidungen dienen zu können. Das vorliegende Gutachten will diese Lücke schließen. Es befasst sich hierzu konkret mit einer Auswahl von Studien zu möglichen Effekten einer Steuer auf zuckergesüßte Getränke, die in der öffentlichen Debatte teils erhebliche Medienresonanz erfahren haben, und bewertet diese aus statistisch-methodischer Perspektive.

Hintergrund

Die Frage „Was wäre, wenn ...?“ zählt zu den grundlegenden Überlegungen bei der Analyse potenzieller Auswirkungen von Maßnahmen oder Entscheidungen. Um solche Effekte abschätzen zu können, kommen Modellierungsstudien zum Einsatz (Jahn et al., 2022). Diese sind ein etabliertes Instrument, um hypothetische Entwicklungen und ihre potenziellen Folgen zu veranschaulichen. So wird im Rahmen sogenannter *kontrafaktischer* Analysen eine beobachtete Entwicklung einer hypothetischen gegenübergestellt. Der Vergleich beider Situationen ermöglicht unter gewissen Voraussetzungen eine Beurteilung der Auswirkungen der implementierten Maßnahme.

Modellierungsstudien werden insbesondere genutzt, um die potenziellen Auswirkungen politischer Maßnahmen zu analysieren (Münnich et al., 2021), bspw. durch Simulationen zukünftiger gesellschaftlicher Entwicklungen unter verschiedenen Szenarien (Schmaus, 2023). Im Gesundheitsbereich wird damit etwa die Wirksamkeit ernährungspolitischer Maßnahmen zur Vorbeugung gewisser Krankheiten wie Diabetes untersucht (Mertens et al., 2022). Modellierungsstudien werden daher von Entscheidungsträgern als wichtige Informationsquelle und somit als Grundlage für evidenzbasierte politische Entscheidungsfindungen herangezogen (Jahn et al., 2022) – unabhängig davon, ob dies ursprünglich von den Forschenden intendiert war oder nicht. Dabei wird jedoch nicht immer die statistisch-methodische Qualität der Studien (z. B. die Angemessenheit der Datengrundlage) überprüft.

Sollen Studien maßgeblich zur politischen Entscheidungsfindung beitragen, müssen sie auf allen Ebenen besonders hohen Qualitätsstandards genügen: Sie müssen sich an strengsten Kriterien hinsichtlich der Datenqualität, der Eignung der Methodik und der Einhaltung wissenschaftsethischer Grundsätze messen lassen. Bisher erfüllen Studien, die zur politischen Entscheidungsfindung herangezogen werden, oftmals nicht solche Erfordernisse.

Studien mit besonders hohen Qualitätsstandards erfordern unter anderem eine hohe Genauigkeit der Ergebnisse (Baker et al., 2013; Europäische Union, 2020; Münnich, 2023). Das bedeutet, dass die mithilfe eines Modells geschätzten Parameter möglichst nahe an den tatsächlichen, in der Regel unbekannten Werten liegen (Europäische Union, 2020). Genauigkeit in den Ergebnissen von Modellierungsstudien lässt sich aber nur erzielen, wenn sowohl die Modellannahmen über die realen Wirkzusammenhänge korrekt sind als auch Genauigkeit in den Modellparametern vorliegt:

- Allen Modellierungsstudien liegt ein (oftmals mathematisches) Modell zugrunde, das eine komplexe Realsituation vereinfacht abbildet (Arnold et al., 2019; Bungartz et al., 2013). Da eine vollständig realitätsgetreue Abbildung nicht möglich ist, wird die Realsituation auf jene Faktoren reduziert, die für die betreffende Fragestellung als relevant erachtet werden (Weyer & Roos, 2017). Hierfür werden an verschiedenen Stellen **Annahmen** getroffen, bspw. darüber, welche Faktoren berücksichtigt werden und wie diese miteinander in Beziehung stehen oder aufeinander einwirken. Aufgrund dieser Annahmen sind die Ergebnisse von Modellierungsstudien stets mit Unsicherheit behaftet (Emmert-Fees et al., 2024).

- Unsicherheit in Modellierungsergebnissen resultiert darüber hinaus aus den gewählten Modellparametern, die verschiedenen externen Datenquellen entliehen werden (Jahn et al., 2022) und in das Modell einfließen (Emmert-Fees et al., 2024). Die **Qualität der Eingangsdaten** bestimmt entsprechend die Qualität der Ergebnisse (Li & O'Donoghue, 2012). Hochwertige Daten sind somit eine unverzichtbare Voraussetzung für evidenzbasierte Politikgestaltung (Jahn et al., 2022; Münnich, im Druck; United Nations, 1954).

Modellierungsstudien befassen sich nur selten explizit mit der Frage, wie zugrunde liegende Annahmen oder die Qualität der Daten ihre Ergebnisse beeinflussen können. Ein Positivbeispiel hierzu ist die kürzlich veröffentlichte Modellierungsstudie von Emmert-Fees et al. (2024), die ausführlich und gezielt untersucht, wie stark die simulierten Effekte einer potenziellen Steuer auf zuckergesüßte Getränke von den berücksichtigten Faktoren und gewählten Modellparametern abhängen[6]: Die prognostizierten Veränderungen des Körpergewichts sowie die Schätzungen zu gesundheitlichen und gesundheitsökonomischen Kennzahlen variierten je nach Szenario erheblich. Dies verdeutlicht, dass (1) die Ergebnisse von Modellierungsstudien kritisch auf ihre Abhängigkeit von Annahmen und (2) diese Annahmen auf ihre Belastbarkeit hin geprüft werden sollten – insbesondere dann, wenn sie als Grundlage für regulatorische Maßnahmen dienen sollen.

Die Erstellung einer für die Begründung politischer Maßnahmen geeigneten Modellierungsstudie erfordert also hochwertige und für den Anwendungskontext relevante Daten, realitätsgetreue Annahmen zur Modellierung der Beziehungen zwischen diesen Daten sowie geeignete statistische Verfahren. Nur auf dieser Grundlage können Ergebnisse erzielt werden, die in ihrer Gesamtheit überzeugen. Ungenaue, verzerrte und/oder nicht repräsentative Daten sowie falsche Annahmen über die zugrunde liegenden Wirkmechanismen führen zu Modellergebnissen, die zwar innerhalb des Modells konsistent sein mögen, aber die Realität nicht korrekt abbilden.

Inhalte und Aufbau des Gutachtens

Zahlreiche Modellierungsstudien beschäftigen sich mit den Effekten einer Steuer auf zuckergesüßte Getränke (englisch: *sugar-sweetened-beverages*; kurz: *SSBs*). Das vorliegende Gutachten befasst sich mit einer Auswahl solcher Studien, die in der öffentlichen Debatte teils erhebliche mediale Aufmerksamkeit erfahren haben, und bewertet diese aus statistisch-methodischer Perspektive. Hierfür werden die Datengrundlage, die Spezifikation der Modelle sowie die Kommunikation der Ergebnisse umfassend analysiert.

[6] Während die im Rahmen dieses Gutachtens evaluierte Studie von Emmert-Fees et al. (2023) die Effekte einer potenziellen Steuer auf zuckergesüßte Getränke in Deutschland unter festgelegten Annahmen simuliert (Fokus: Inhalt), untersucht die Folgestudie von Emmert-Fees et al. (2024) in einem vergleichbaren Kontext explizit, wie sich Variationen dieser Annahmen auf die Ergebnisse auswirken (Fokus: Methodik). Die Erkenntnisse der neueren Studie dienen insbesondere in Kapitel 2 und Kapitel 3 als zentrale Argumente für die statistisch-methodische Perspektive, die dieses Gutachten einnimmt. Aufgrund der unterschiedlichen Erscheinungsjahre der Publikationen ist jederzeit eindeutig nachvollziehbar, auf welche der beiden Studien sich die Ausführungen in diesem Gutachten beziehen.

Insgesamt wurden dafür sechs Modellierungsstudien betrachtet: Zentral war die Evaluation der Studie von Emmert-Fees et al. (2023), welche die Effekte verschiedener Varianten einer Besteuerung zuckergesüßter Getränke auf gesundheitliche sowie gesundheitspolitische Aspekte in Deutschland untersuchte und daher in der öffentlichen Debatte rund um die potenzielle Einführung einer solchen Steuer in Deutschland häufig zitiert wird. Ergänzend wurde eine weitere Studie für Deutschland evaluiert (Schwendicke & Stolpe, 2017). Zudem wurden Studien aus Ländern herangezogen, die bereits eine solche Steuer implementiert haben. Zwei dieser Studien (Rogers, Cummins et al., 2023, und Cobiac et al., 2024) untersuchten die Auswirkungen der Steuer in England, wo die Hersteller im Rahmen einer gestaffelten Verbrauchssteuer abhängig vom Zuckergehalt ihrer Getränke unterschiedlich hohe Abgaben zahlen müssen. Weitere zwei Studien befassten sich mit den Auswirkungen der Steuer in Mexiko (Gračner et al., 2022, und Basto-Abreu et al., 2019), wo die Steuer in Form einer pauschalen Verbrauchsteuer realisiert wurde.

Das Gutachten gliedert sich wie folgt: Zunächst erfolgt eine detaillierte Betrachtung von Modellierungsstudien einschließlich ihrer grundsätzlichen Potenziale und Grenzen (Kapitel 2). Darauf basierend werden Beurteilungskriterien zur Evaluation von Modellierungsstudien abgeleitet (Kapitel 3), die anschließend zur Evaluation der sechs ausgewählten Studien angewendet werden (Kapitel 4). Abschließend werden die Ergebnisse der Einzelevaluationen zusammengefasst und diskutiert (Kapitel 5).

2 Modellierungsstudien

Mithilfe von Modellierungsstudien lassen sich hypothetische und tatsächliche Szenarien einander gegenüberstellen. Modellierungsstudien werden daher bspw. angewendet, um die erwarteten Auswirkungen einer potenziellen Besteuerung von Produkten abzuschätzen (Emmert-Fees et al., 2024; Mertens et al., 2022). Die Ergebnisse dieser Modellierungen werden häufig – ob von den Studienautoren intendiert oder nicht – von politischen Entscheidungsträgern als Grundlage für die Diskussion über solche Maßnahmen herangezogen. Dabei ist ein Verständnis der allgemeinen Potenziale und Grenzen von Modellierungsstudien (Abschnitt 2.1) unumgänglich, um zu beurteilen, welche Aussagen sie über Zusammenhänge, Ursachen und Wirkungen der analysierten Thematik ermöglichen und welche nicht (Abschnitt 2.2).

2.1 Potenziale und Grenzen

Die Potenziale und Grenzen von Modellierungsstudien lassen sich wie folgt zusammenfassen:

Potenziale

- **Beitrag zur Entscheidungsfindung**: Modellierungsstudien können zur evidenzbasierten Entscheidungsfindung beitragen (Jahn et al., 2022). Im Idealfall kombinieren sie die bestmöglichen verfügbaren Daten in einem mathematischen Modell, um hypothetische Szenarien zu simulieren (Emmert-Fees et al., 2024). Dabei ist aber zu beachten, dass selbst die bestmöglichen verfügbaren Daten von geringer Qualität sein können, was den Beitrag der darauf basierenden Ergebnisse zur Entscheidungsfindung limitiert (siehe dazu genauere Ausführungen im nachfolgenden Abschnitt *Grenzen*). Was „bestmöglich“ im jeweiligen Kontext bedeutet, sollte jedenfalls immer sorgfältig geprüft und transparent kommuniziert werden. Unter der Voraussetzung hochwertiger Daten dienen Modellierungsstudien als wertvolles Instrument zur systematischen Befassung mit „Was-wäre-wenn“-Fragen sowie zur Analyse potenzieller Auswirkungen verschiedener politischer Maßnahmen dar (Emmert-Fees et al., 2024; Li & O'Donoghue, 2012).

- **Ex-ante-Analysen**: Modellierungsstudien analysieren die potenziellen Auswirkungen einer Maßnahme, bevor die Maßnahme tatsächlich umgesetzt wird (Mertens et al., 2022). So können Abschätzungen der erwarteten Effektstärken oder von Kosten-Nutzen-Relationen bereits ex ante vorgenommen werden.

- **Hoher Detailgrad**: Abhängig von der gewählten Methodik analysieren Modellierungsstudien nicht nur das mögliche Verhalten der Gesamtpopulation oder ihrer Subgruppen (Makroebene), sondern auch das mögliche Verhalten einzelner Individuen und deren Interaktionen (Mikroebene; Schmaus, 2023).

- **Berücksichtigung von Komplexität**: Grundsätzlich lassen sich beliebig komplexe Sachverhalte modellieren, wodurch eine umfassende Annäherung an die Realität möglich wird (Weyer & Roos, 2017).

Grenzen

- **Datenverfügbarkeit**: Eine in der Praxis nicht zu unterschätzende Einschränkung von Modellierungsstudien liegt in der mangelnden Verfügbarkeit geeigneter Daten zur Initialisierung und Parametrisierung der Modelle (Emmert-Fees et al., 2024; Weyer & Roos, 2017). Modelle integrieren im Idealfall die bestmöglichen verfügbaren Daten (Emmert-Fees et al., 2024). Fehlt es jedoch an geeigneten, d. h. repräsentativen und

genauen Daten, so hat dies gravierende Auswirkungen auf die Aussagekraft der Modellierungsergebnisse (Mertens et al., 2022). In ernährungswissenschaftlichen Anwendungsbereichen fehlen häufig hinreichende Daten auf Produktebene oder für spezifische Bevölkerungsgruppen, obwohl diesbezüglich Varianz besteht[7]. Daher muss hilfsweise auf Annahmen zurückgegriffen werden, was zu Verzerrungen oder Unschärfen in den Ergebnissen führen kann.

- **Datenqualität**: Qualitativ hochwertige Daten bilden eine wesentliche Grundlage für Forschung und evidenzbasierte Politik (Jahn et al., 2022). Eine mangelnde Datenqualität stellt Forschende jedoch oft vor große Herausforderungen (Emmert-Fees et al., 2024). Da die Qualität der Eingangsdaten einen direkten Einfluss auf die Qualität der Ergebnisse von Modellierungsstudien hat (Li & O'Donoghue, 2012), hängt die Aussagekraft von Modellierungsergebnissen maßgeblich von der Wahl der Datenquellen und den Unsicherheiten in den Daten ab (Mertens et al., 2022). Sollen Daten als Grundlage für Gesetzgebungsverfahren dienen, müssen besonders strenge Anforderungen an ihre Qualität angelegt werden (Münnich, im Druck).

  Eingangsdaten, die durch Befragungen von Individuen erhoben wurden, können in vielfacher Weise fehlerbehaftet sein. Mögliche Fehler, die bei der Konzeption, Erhebung, Verarbeitung und Analyse der Daten auftreten können, führen zu Abweichungen der aus Befragungsdaten geschätzten Werte (z. B. Mittelwerte) von den tatsächlichen Werten (Biemer, 2010). Darunter fallen sowohl zufällige Fehler wie der *Stichprobenfehler*[8] als auch nichtzufällige Fehler wie *Deckungsfehler*, *Nonresponse-Fehler* (d. h. Fehler aufgrund von Antwortausfällen), *Messfehler* oder *Verarbeitungsfehler* (Lohr, 2021). Wurden die Eingangsdaten bspw. nicht *probabilistisch* (d. h. rein zufällig) erhoben oder unterscheiden sich die Antwortverweigerer systematisch von den Antwortenden, ist die Generalisierbarkeit der Daten problematisch (Münnich, 2020). In solchen Fällen treffen die vom Modell vorhergesagten Resultate für die untersuchte Grundgesamtheit möglicherweise nicht zu. Modellierungsergebnisse, die auf Befragungsdaten basieren, sollten daher mit entsprechender Vorsicht interpretiert werden (Jahn et al., 2022).

  Modellierungsstudien beziehen ihre Eingangsdaten in der Regel aus externen Quellen und setzen häufig deren Repräsentativität und Genauigkeit voraus. Der Verzicht auf eine kritische Diskussion der Datengrundlagen suggeriert allerdings eine vermeintliche Sicherheit in der Gültigkeit der Ergebnisse, obwohl das Risiko von Verzerrungen besteht. Datenerhebungen zu ernährungswissenschaftlichen Fragestellungen sind bspw. besonders anfällig für Messfehler aufgrund von Falschangaben (Emmert-Fees et al., 2024): Individuen neigen dazu, ihren Konsum vermeintlich bedenklicher Produkte im Sinne sozial erwünschten Antwortverhaltens nicht wahrheitsgemäß anzugeben. Dies kann zu einer Untererfassung des Konsums solcher Produkte führen (Emmert-Fees et al., 2024; Mertens et al., 2022). Werden solche Daten in Modellierungsstudien verwendet, kann dies die Ergebnisse erheblich verzerren (Emmert-Fees et al., 2024).

  Daten, die in einem bestimmten Anwendungskontext als hochwertig gelten, müssen nicht zwangsläufig dieselbe Qualität in einem anderen Kontext aufweisen. Bspw. können repräsentative Daten zum Basiskonsum mexikanischer Jugendlicher zwar für eine Studie zu dieser Zielgruppe geeignet sein, sind jedoch ungeeignet, wenn sie auf deutsche Jugendliche übertragen werden, denn körperliche Voraussetzungen und soziokulturelle

[7] Bspw. variiert der Konsum zuckergesüßter Getränke je nach Alter, Geschlecht und spezifischem Produkt (Emmert-Fees et al., 2024).
[8] Der Stichprobenfehler entsteht, weil anstelle der vollständigen Grundgesamtheit nur eine zufällig ausgewählte Stichprobe erhoben wird (Faulbaum, 2022; Lohr, 2021).

Rahmenbedingungen können zwischen den beiden Ländern erheblich variieren.

- **Annahmen**: Prognosen und Projektionen beruhen auf expliziten oder impliziten Annahmen (Ernst et al., 2023), da bei der Erstellung von Modellen an zahlreichen Stellen Entscheidungen getroffen werden müssen (Schmaus, 2023). Jede Entscheidung für eine bestimmte Annahme ist gleichzeitig eine Entscheidung gegen alternative Möglichkeiten und bringt dadurch zwangsläufig Unsicherheit in das Modell. Modellierungsergebnisse werden daher maßgeblich von den zugrunde liegenden Annahmen beeinflusst (Mertens et al., 2022). In der methodenfokussierten Studie von Emmert-Fees et al. (2024) wurde bspw. gezeigt, dass die prognostizierte Veränderung des Körpergewichts im Zusammenhang mit dem Konsumrückgang zuckergesüßter Getränke maßgeblich von den getätigten Annahmen abhängt – etwa davon, ob eine kalorische Substitution durch andere Lebensmittel oder Getränke berücksichtigt wird oder ob davon ausgegangen wird, dass keine solche Substitution erfolgt.

- **Hoher Aufwand**: Modellierungsstudien sind häufig mit einem hohen Aufwand für die Abbildung realer Sachverhalte in Form von Modellen verbunden. Mit zunehmender Komplexität des Gesamtmodells steigt auch die Anzahl potenzieller Einflussgrößen, deren Schätzung jeweils mit Unsicherheit behaftet ist und die deshalb die Unsicherheit der Modellergebnisse erhöhen können (Schmaus, 2023). Außerdem wird die mögliche Komplexität der Modellierung durch die Rechenleistung des eingesetzten Computers sowie die verfügbare Zeit für Implementierung und Analyse begrenzt (Weyer & Roos, 2017). Die Abwägung zwischen Aufwand und Nutzen gestaltet sich dabei oft als schwierig, denn auch eine Reduktion der Modellkomplexität durch vereinfachende Annahmen hat Unsicherheiten in den Modellergebnissen zur Folge, wie die methodenfokussierte Studie von Emmert-Fees et al. (2024) ergab.

  Ernährungswissenschaftliche Studien erfordern die Modellierung hoch komplexer Systeme, bspw. des menschlichen Körpers und seiner Reaktionen auf Umwelteinflüsse. Es ist davon auszugehen, dass zahlreiche nichtlineare Effekte und Interaktionen vorliegen, die sich mit begrenzten Daten und Datenverarbeitungsressourcen nicht vollständig abbilden lassen.

Die obigen Ausführungen zeigen, dass die Ergebnisse statistischer Modelle aus unterschiedlichen Gründen mit Unsicherheit behaftet sein können (Jahn et al., 2022; Schmaus, 2023). Die Quantifizierung dieser Unsicherheit stellt eine erhebliche Herausforderung dar (Schmaus, 2023). Ein kompetenter Umgang mit Modellierungsstudien beinhaltet dabei die bewusste Berücksichtigung und die transparente Kommunikation von Unsicherheit in den Ergebnissen. Anstelle eines isolierten Punktschätzers für das Ergebnis sollte ein Intervall oder eine Verteilung möglicher Ergebnisse betrachtet werden, um die Unsicherheiten transparent zu kommunizieren (Jahn et al., 2022). Modellierungsstudien zu Fragestellungen der öffentlichen Gesundheit ignorieren häufig bestehende Unsicherheiten in den Ergebnissen, wie eine methodenfokussierte Studie aufzeigte (Emmert-Fees et al., 2024). Das ist insbesondere problematisch, wenn die Ergebnisse als Grundlage einer Diskussion über politische Maßnahmen bzw. für deren vermeintlich faktenbasierte Legitimation herangezogen werden.

2.2 Effekte und Einflussfaktoren

Die im vorliegenden Gutachten evaluierten Modellierungsstudien untersuchen verschiedene Auswirkungen (*Effekte*[9]) einer (potenziellen) Zuckersteuer. Im Folgenden werden zugrunde liegenden Annahmen über Effekte in Modellierungsstudien (Abschnitt 2.2.1) sowie mögliche weitere Einflussfaktoren (Abschnitt 2.2.2) aus methodischer Perspektive beleuchtet.

2.2.1 Zusammenhänge, Ursachen und Wirkung

Es ist wichtig, stets klar zwischen den Beziehungen zwischen Variablen zu unterscheiden **Kausalität** beschreibt die eindeutige Beziehung von Ursache und Wirkung, d. h. eine Änderung einer Variable A hat eine Änderung einer anderen Variable B zur Folge. **Korrelation** hingegen bedeutet, dass es zwar einen statistischen Zusammenhang zwischen den Variablen A und B gibt, aber eine Änderung von A nicht zwangsläufig auch eine Änderung von B zur Folge hat. Aus Korrelation folgt dabei nicht zwingend Kausalität. Dabei gilt, dass Korrelation nicht automatisch Kausalität impliziert. Zudem lässt sich nicht immer eindeutig feststellen, welche Variable die Änderungen in der anderen verursacht.

Der anerkannte Goldstandard zum Nachweis von Kausalität sind *randomisierte kontrollierte Studien* (englisch: *randomized-controlled trials*; Arnold et al., 2019; Jahn et al., 2022), die aber für die Untersuchung vieler anderer gesellschaftlicher Fragestellungen – einschließlich des Anwendungskontexts der Zuckersteuer – aus forschungspraktischen Gründen nicht geeignet sind (The Royal Swedish Academy of Sciences, 2021). Wenn randomisierte Experimente nicht durchführbar sind und stattdessen Beobachtungsdaten herangezogen werden, gestaltet sich die Ableitung kausaler Schlussfolgerungen deutlich schwieriger (Jahn et al., 2022). Der stichhaltige Nachweis kausaler Zusammenhänge zählt zu den anspruchsvollsten Aufgaben der empirischen Statistik[10]. In Modellierungsstudien beruhen die entwickelten Modelle auf Annahmen über zugrunde liegende (kausale) Beziehungen (Jahn et al., 2022). Entscheidend ist dabei, ob diese unterstellten Kausalitäten auf tatsächlich nachgewiesenen oder lediglich hypothetischen Zusammenhängen basieren. Die Aussagekraft der Ergebnisse von Modellierungsstudien in Bezug auf (kausale) Zusammenhänge ist also insofern eingeschränkt, als sie davon abhängt, wie gut die im Modell angenommenen kausalen Zusammenhänge zuvor belegt wurden (Mertens et al., 2022). Modellierungsstudien können zwar hilfreich sein, um (potenziell kausale) Muster zu verstehen und zu erklären, stellen aber selbst keine kausalen Beziehungen im statistischen Sinne her. Im Klartext bedeutet dies: Kausalwirkungen zwischen Einfluss- und Zielgrößen gehen als Voraussetzung in die Konstruktion des Modells ein – sie werden also ex ante unterstellt[11]. Ein Rückschluss auf die Gültigkeit der Kausalitätsannahme erfolgt ex post und indirekt, wenn die durch das Modell simulierten Ergebnisse hinreichend gut mit der Realität übereinstimmen.

[9] Ein Effekt ist die funktionale Darstellung einer tatsächlichen oder vermuteten Ursache-Wirkungs-Beziehung zwischen einem Einflussfaktor und einer abhängigen Variablen.

[10] Die Komplexität dieses Problems wird auch durch die Vergabe des Alfred-Nobel-Gedächtnispreises für Wirtschaftswissenschaften im Jahr 2021 verdeutlicht, welcher zur Hälfte an Forscher verliehen wurde, die für ihre methodischen Beiträge zur Analyse von Kausalzusammenhängen ausgezeichnet wurden. Sie beschäftigten sich mit der Frage, welche Schlussfolgerungen über Kausalität aus sogenannten *natürlichen Experimenten* gezogen werden können (The Royal Swedish Academy of Sciences, 2021).

[11] In Modellierungsstudien zur Untersuchung der Effekte einer Steuer auf zuckergesüßte Getränke werden typischerweise folgende (kausale) Zusammenhänge ex ante unterstellt: Ausgangspunkt ist die (hypothetische) Einführung einer solchen Steuer, die – abhängig von ihrer Ausgestaltung – unterschiedliche Marktreaktionen auslösen kann. Hersteller reagieren mit Preiserhöhungen oder Rezepturanpassungen, woraufhin sich das Kauf- und Konsumverhalten der Bevölkerung verändert. Ein reduzierter Verzehr zuckergesüßter Getränke kann den Gesamtzuckerkonsum verringern, was wiederum zu einer geringeren Kalorienaufnahme und potenziellen Gewichtsveränderungen führen kann. Diese Veränderungen wirken sich wiederum auf die Gesundheit, die Lebensqualität und gesundheitsökonomische Aspekte aus.

Dabei handelt es sich aber nicht um einen (wissenschaftlichen) Beweis, sondern um eine Plausibilisierung der Annahmen.

Um kausale Zusammenhänge im Gesundheitsbereich „nachzuweisen“, wird hauptsächlich auf kontrafaktische Analysen zurückgegriffen (Arnold et al., 2019). Dabei handelt es sich bspw. um Simulationsstudien, welche die Auswirkungen von Maßnahmen der Gesundheitspolitik – etwa der Einführung einer Steuer auf zuckerhaltige Lebensmittel mit der Intention, den Zuckerkonsum in der Bevölkerung sowie das Auftreten damit verbundener Erkrankungen zu reduzieren – mit der Situation ohne eine solche Maßnahme vergleichen. Bereits bei der Konstruktion der zu analysierenden Modelle werden allerdings Annahmen über zugrunde liegende (kausale) Beziehungen getroffen. So wird etwa angenommen, dass Preisänderungen bei steuerpflichtigen Produkten ursächlich zu Konsumänderungen dieser Produkte führen oder dass ein Rückgang des Konsums besteuerter zuckerhaltiger Lebensmittel eine Gewichtsreduktion nach sich zieht. Selbst wenn die auf diesen Annahmen basierenden Ergebnisse plausibel erscheinen, lässt sich ohne entsprechende Nachweise keine eindeutige Ursache-Wirkungs-Beziehung konstatieren.

2.2.2 Unberücksichtigte Einflussfaktoren und isolierte Effekte

Der Nachweis und die Interpretation von (kausalen) Beziehungen in den Daten ist auch dahingehend schwierig, da beobachtete Zusammenhänge möglicherweise durch andere, nicht im Modell berücksichtigte Variablen beeinflusst worden sein können (Jahn et al., 2022).[12]

Für die Beurteilung der tatsächlichen Wirkung geplanter Interventionen wie bspw. einer Zuckersteuer muss deren isolierter Effekt möglichst genau ermittelt werden. In der Statistik wird diesem Problem begegnet, indem neben dem vermuteten Kausalfaktor weitere potenzielle Einflussfaktoren als Kontrollvariablen in das Modell einbezogen werden (Arnold et al., 2019). Dadurch kann der Einfluss der Kontrollvariablen auf die abhängige Variable herausgerechnet werden (Doering & Bortz, 2016). Eine große Herausforderung besteht jedoch darin, alle relevanten Einflussfaktoren angemessen zu berücksichtigen. Im Beispiel der Zuckersteuer müssten potenzielle Substitutionseffekte (z. B. Verzehr anderer kalorienhaltiger Getränke oder Lebensmittel) in die Modellierung einfließen. Ebenso müssten sämtliche weitere Risikofaktoren, die die Entwicklung von Adipositas begünstigen, auf individueller Ebene modelliert werden. Adipositas gilt als multifaktorielle Erkrankung, die nicht nur durch Ernährungsgewohnheiten, sondern auch durch zahlreiche weitere Einflussfaktoren wie das soziale Umfeld oder das Maß an physischer Aktivität geprägt wird (Hummel et al., 2013; World Health Organization, 2022). Werden diese nicht berücksichtigt, besteht das Risiko, dass Modellierungsstudien zu gravierenden Fehlschlüssen hinsichtlich der potenziellen Auswirkungen einer solchen Steuer führen.

[12] Die Literatur unterscheidet (teils uneinheitlich) verschiedene solche Drittvariableneffekte (z. B. Scheinkorrelation, Interaktion, Konfundierung, Mediation, Moderation; Weiber und Mühlhaus, 2014). Diese Effekte unterscheiden sich darin, ob und wie die dritte Variable mit der abhängigen und/oder der unabhängigen Variablen in Verbindung steht. Eine genauere Differenzierung von Drittvariableneffekten ist für den Zweck des vorliegenden Gutachtens nicht erforderlich.

3 Beurteilungskriterien zur Evaluation von Modellierungsstudien

Auf Grundlage der in Kapitel 2 beschriebenen Charakteristika von Modellierungsstudien werden in diesem Kapitel die Beurteilungskriterien für die Studienevaluationen abgeleitet, die in die drei Kategorien *Datengrundlage*, *Modellspezifikation* und *Ergebniskommunikation* unterteilt sind. Daten dienen dabei als Rohmaterial, Modelle als analytisches Werkzeug zu ihrer Verarbeitung und die Ergebnisse stellen das daraus hervorgehende Produkt dar.

3.1 Datengrundlage

Die Datengrundlage wird in Modellierungsstudien häufig nur unzureichend oder gar nicht diskutiert. Da hochwertige Daten aber eine unverzichtbare Voraussetzung für evidenzbasierte Politikgestaltung sind, spielt ihre Beurteilung im vorliegenden Gutachten eine zentrale Rolle.

Verfügbarkeit und Qualität der zugrunde liegenden Eingangsdaten

Wie in Abschnitt 2.1 erläutert, hängt die Genauigkeit der Ergebnisse von Modellierungsstudien maßgeblich von der Verfügbarkeit und Qualität der zugrunde liegenden Eingangsdaten ab, die meist aus verschiedenen externen Quellen stammen. Um verlässliche Studienergebnisse zu erzielen, muss die Passung der Daten zum Untersuchungsgegenstand sichergestellt und das Risiko von Verzerrungen minimiert werden. Die Eingangsdaten der evaluierten Studien werden in Bezug auf ihre Aktualität, ihre Vollständigkeit, die Erhebungsmethoden (z. B. Befragung, Beobachtung, Experiment), die Anzahl der Erhebungszeitpunkte (z. B. quer- oder längsschnittlich[13]), damit verbundene potenzielle zufällige und nichtzufällige Fehler (siehe Abschnitt 2.1) sowie ihre Anwendbarkeit im Studienkontext bewertet[14]. Außerdem wird die Kompatibilität der verschiedenen Datenquellen evaluiert.

Bei der Untersuchung möglicher Auswirkungen einer Besteuerung spezifischer Produkte ist es bspw. entscheidend, wie sich die *Preiselastizität der Nachfrage*, also die Nachfrage nach einem Gut als Reaktion auf Preisveränderungen, entwickelt. Diese ist grundsätzlich nicht bekannt und muss daher im Rahmen von Studien ermittelt werden. Die Ergebnisse sind jedoch typischerweise mit Unsicherheit behaftet oder nicht repräsentativ. So kann eine Preiselastizität, die ausschließlich in städtischen Gebieten erhoben wurde, nicht ohne Weiteres auf ländliche Regionen übertragen werden. Darüber hinaus muss geprüft werden, ob die zugrunde liegenden Daten für den räumlichen, zeitlichen und sachlichen Kontext der jeweiligen Studie geeignet sind. So lassen sich empirisch ermittelte Preiselastizitäten nämlich nicht unmittelbar auf andere Länder übertragen, da sie von zahlreichen weiteren Faktoren abhängen, wie etwa dem Preisniveau von Substituten. Auch die Wahl der Erhebungsmethode spielt eine entscheidende Rolle. So weichen Konsumschätzungen, die auf Befragungen von Individuen basieren, von Schätzungen anhand von Absatzzahlen ab (Emmert-Fees et al., 2024). Absatzzahlen selbst variieren abhängig von den betrachteten Vertriebskanälen, etwa ob der Konsum ausschließlich basierend auf Einzelhandelsdaten (z. B. Supermärkte) oder einschließlich gastronomischer Daten (z. B. Restaurants, Kinos) erfasst wurde.

[13] In einer *Querschnittstudie* wird eine Stichprobe zu einem einzigen Zeitpunkt untersucht. Eine *Trendstudie* besteht aus mehreren zeitlich versetzten Querschnittstudien, bei denen zumindest teilweise dieselben Variablen erfasst werden. Sie dienen dazu, gesellschaftliche Veränderungen im Laufe der Zeit zu analysieren. Eine *Längsschnitt- bzw. Panelstudie* untersucht wiederholt dieselbe Stichprobe (Panel) über einen längeren Zeitraum hinweg. Sie eignet sich daher besonders zur Analyse individueller Veränderungen im Lebensverlauf (Doering & Bortz, 2016).

[14] Insbesondere wenn die Daten nicht selbst erhoben wurden, kann die Prüfung möglicher Verzerrungen eine erhebliche Herausforderung darstellen. Indikationen liefern insbesondere die Beschreibungen der jeweiligen Datenerhebung und des Umgangs mit Nonresponse.

Stichprobe

Zur Bewertung der Datengrundlage einer Studie sollte insbesondere geprüft werden, inwieweit die zugrunde liegende Stichprobe die interessierende Population repräsentiert. Stichproben in Modellierungsstudien können entweder *real* oder *synthetisch* sein:

- Reale Stichproben unterliegen – selbst bei zufälliger Auswahl und großem Umfang – natürlicherweise gewissen Schwankungen, die als (zufälliger) Stichprobenfehler[15] bezeichnet werden. Um das Ausmaß zufälliger Fehler darzustellen, werden üblicherweise Konfidenzintervalle angegeben. Diese lassen sich jedoch nur dann korrekt ermitteln, wenn sog. probabilistische[16] Stichproben vorliegen.
- Synthetische Stichproben sind künstlich erstellte Datensätze, die eine Population mithilfe von Parametern der Bevölkerungsverteilung und demografischen Merkmalen möglichst realitätsnah nachbilden, um spezifische Szenarien zu simulieren. Die Ergebnisse eines Modells, das auf solchen Stichproben basiert, hängen davon ab, wie gut die Stichprobe die Eigenschaften der betrachteten Population widerspiegelt und welche Annahmen über den Lebensverlauf der untersuchten Individuen getroffen werden (Mertens et al., 2022).

Die Zusammensetzung einer Stichprobe muss zum relevanten Untersuchungszeitpunkt Rückschlüsse auf die Gesamtheit der interessierenden Population ermöglichen. Dieser Anspruch wird als *Repräsentativität*[17] einer Stichprobe bezeichnet und ist mitentscheidend dafür, ob und wie die Studienergebnisse verallgemeinert werden können. Die Stichproben der im Gutachten evaluierten Studien werden folglich hinsichtlich ihrer Repräsentativität bewertet.

3.2 Modellspezifikation

Zur Abschätzung des Modellrisikos – also des Risikos, das sich aus der Abweichung eines Modells von der Realität ergibt – untersucht das Gutachten die Spezifikationen der jeweils verwendeten Modelle. Die Festlegung bestimmter Modellannahmen und die gleichzeitige Ablehnung alternativer Annahmen führen zu Unsicherheit im Modell, die sich auf verschiedene Aspekte der Modellspezifikation zurückführen lässt:

Modellierungsverfahren

Die Ergebnisse von Modellierungsstudien werden durch das verwendete Modellierungsverfahren und die statistischen Analysen bestimmt, deren Angemessenheit vom jeweiligen Forschungsziel abhängt. Zur Beurteilung der Aussagekraft der Ergebnisse müssen dabei die Limitationen der jeweiligen Verfahren (siehe Anhang A) berücksichtigt werden.

[15] Grundsätzlich gilt: Je größer die Stichprobe, desto kleiner ist bei unverzerrten Stichproben der Stichprobenfehler. In der Praxis bedeutet dies, dass größere, unverzerrte Stichproben genauere Schätzungen liefern (Mittag & Schüller, 2023). Kritisch zu bewerten ist insbesondere, wenn Subpopulationen, wie etwa bestimmte Altersgruppen untersucht werden sollen, deren Fallzahlen in der Stichprobe relativ klein sind (Li & O'Donoghue, 2012).

[16] Eine probabilistische Stichprobe erfordert, dass die Wahrscheinlichkeit jedes Individuums, in die Stichprobe zu gelangen, positiv und bekannt ist. Nur solche Stichproben können den Anspruch auf Repräsentativität erfüllen, wobei ggf. Angleichungen an die Gesamtpopulation ex ante (durch Quotierung) und/oder ex post (durch Gewichtung) erfolgen können.

[17] Die Interpretation des Begriffs *Repräsentativität* ist keineswegs trivial. Insbesondere reicht die Repräsentativität einer Stichprobe alleine nicht aus, um auf qualitativ hochwertige Ergebnisse zu schließen. Für diese Sicherstellung müssen auch Stichproben- und Nicht-Stichprobenfehler mitberücksichtigt werden (Münnich, 2020).

Einfluss- und Zielgrößen

Die Struktur des Modells wird durch die Festlegung der Einflussgrößen (unabhängige Variablen) und Zielgrößen (abhängige Variablen) definiert. Dies umfasst zum einen die Operationalisierung dieser Größen, also wie theoretische Konstrukte (z. B. Lebensqualität) in messbare Variablen (z. B. Lebensjahre, krankheitsbelastete Jahre) überführt werden (Doering & Bortz, 2016). Zum anderen ist von Bedeutung, wie diese Größen miteinander in Beziehung stehen. Diese Beziehungen werden durch funktionale Zusammenhänge zwischen Einfluss- und Zielgrößen beschrieben, die entweder linear oder nichtlinear[18] modelliert werden können. Auch Wechselwirkungen zwischen den Einflussfaktoren sind ggf. relevant. Bildet die Modellstruktur die realen Wirkzusammenhänge nicht adäquat ab, besteht das Risiko fehlerhafter Schlussfolgerungen.

Weitere Einflussfaktoren

Um Verzerrungen in den Ergebnissen zu minimieren, sollte ein Modell neben den zentralen Einfluss- und Zielgrößen für die Analyse auch weitere Variablen einbeziehen, die mit diesen in Beziehung stehen. Werden solche relevanten Einflussfaktoren im Modell nicht ausreichend berücksichtigt, besteht das Risiko, dass aus den Ergebnissen Zusammenhänge zwischen Einfluss- und Zielgrößen abgeleitet werden, die zu stark verallgemeinert oder fehlerhaft sind. Welche Einflussfaktoren im Anwendungskontext Zuckersteuer tatsächlich relevant sind, lässt sich derzeit nicht abschließend klären. Einige bedeutsame Faktoren, die das Konsumverhalten und die Konsumveränderung infolge einer Preisänderung betreffen, werden aber in der methodenfokussierten Studie von Emmert-Fees et al. (2024) identifiziert:

- Der Basiskonsum zuckergesüßter Getränke, also der Konsum vor einer (potenziellen) Besteuerung dieser Produkte, variiert je nach Alter und Geschlecht. Zudem reagieren Individuen mit unterschiedlichem Basiskonsum unterschiedlich auf die Einführung einer Steuer, was sich in variierenden Preiselastizitäten der Nachfrage in den verschiedenen Alters- und Geschlechtsgruppen niederschlägt. Bei der Modellierung der Auswirkungen einer Steuereinführung sind daher die Annahme eines pauschalen Basiskonsums und einer einheitlichen Preiselastizität der Nachfrage zu starke Vereinfachungen. Vielmehr sind als Kontrollvariablen mindestens *Alter* und *Geschlecht* in das Modell aufzunehmen[19].
- Die Preisänderung eines Guts (z. B. durch Besteuerung) kann sich auf den Konsum anderer Güter auswirken, was als sog. *Kreuzpreiselastizität* bezeichnet wird. Bspw. können zuckergesüßte Getränke nach einer Preiserhöhung (teilweise) durch andere, unbesteuerte kalorienhaltige Getränke (z. B. Fruchtsäfte, Milchmischgetränke oder auch alkoholische Getränke) substituiert werden[20]. Bei der Modellierung der Auswirkungen einer Steuereinführung muss daher für eine eventuelle kalorische Kompensation kontrolliert werden (Hummel et al., 2013) – ein Aspekt, der häufig nicht beachtet wird (Thiboonboon et al., 2024) –, da andernfalls das Risiko besteht, den tatsächlichen Effekt erheblich zu überschätzen.

[18] Beispiele für nichtlineare funktionale Zusammenhänge sind sigmoidale oder exponentielle Beziehungen.
[19] Eine fehlerhafte Modellierung des Effekts von Preisänderungen kann einerseits unter das Modellrisiko fallen, wenn Preiselastizitäten nicht nach Subgruppen differenziert in das Modell aufgenommen werden. Andererseits kann sie Teil des Stichprobenrisikos sein, wenn die Studien, auf denen die geschätzten Preiselastizitäten basieren, hohen zufälligen oder nichtzufälligen Fehlern behaftet sind.
[20] Wie die Preiselastizitäten der Nachfrage stehen auch die Kreuzpreiselastizitäten bestimmter Substitute von zuckergesüßten Getränken in Zusammenhang mit dem Basiskonsum dieser Getränke (Emmert-Fees et al., 2024).

3.3 Ergebniskommunikation

Die Darstellung und Kommunikation der Ergebnisse ist zentraler Bestandteil einer jeden Studie, so auch bei Modellierungsstudien. Dabei ist nicht nur die inhaltliche Präzision von Bedeutung, sondern auch die Art und Weise, wie die Ergebnisse vermittelt werden, da Studien ggf. von unterschiedlichen Zielgruppen, etwa von Wissenschaftlern, Medien, politischen Entscheidungsträger oder der breiten Öffentlichkeit rezipiert werden.

Ergebnisdarstellung

Die Ergebnisdarstellung vermittelt – insbesondere bei quantitativen Studien – die zentrale Aussage einer wissenschaftlichen Untersuchung. Um die Objektivität, Reproduzierbarkeit und Vertrauenswürdigkeit sicherzustellen, müssen die Ergebnisse einer Studie neutral, vollständig, transparent, konsistent, verständlich und nachvollziehbar dargestellt werden. Werden diese Prinzipien nicht eingehalten, besteht die Gefahr, dass die Ergebnisse vom Leser missverstanden oder fehlerhaft interpretiert bzw. wiedergegeben werden.

So kann eine unvollständige Darstellung, wie etwa die selektive Präsentation signifikanter Ergebnisse bestimmter Subgruppen, eine Verschleierung nicht vorhandener oder nicht nachweisbarer Effekte sein. Auch die Visualisierung der Ergebnisse erfordert besondere Sorgfalt, da bereits die Wahl der Darstellungsform (z. B. Kreis- oder Balkendiagramm) und der zugrunde liegenden Skalen dazu führen kann, bestimmte Aussagen zu betonen und damit die Wahrnehmung der Leserschaft in eine bestimmte Richtung zu lenken. Die Art der Darstellung von Studienergebnissen kann die Rezeption also bewusst oder unbewusst beeinflussen. Ebenso kann die Wahl der Ergebnisdarstellung in absoluten oder relativen Zahlen unterschiedliche Botschaften vermitteln. Relative Häufigkeiten lassen bei kleiner Ausgangsbasis Risiken tendenziell größer erscheinen, sodass ihre alleinige Angabe potenziell irreführend ist. Daher plädieren u. a. Wegwarth und Gigerenzer (2011) sowie Krauss et al. (2020) für die Kommunikation von Risiken und Risikoveränderungen in natürlichen Häufigkeiten („x von y“). Ähnlich schwer oder missverständlich sind aber auch Hochrechnungen in absoluten Zahlen ohne jegliche Kontextualisierung, vor allem, wenn Unsicherheiten nicht kommuniziert werden. So sind hochgerechnete Summen volkswirtschaftlicher Ersparnisse eines Staats mit 80 Mio. Einwohnern über einen Zeitraum von zwanzig Jahren bestenfalls grobe Schätzungen, die unbedingt einer Kontextualisierung bedürfen. Generell kann eine konsistente Darstellung sowohl in absoluten als auch in relativen Werten also dazu beitragen, die Interpretierbarkeit der Effekte zu verbessern.

Modellbewertung

Die Bewertung der Qualität und Zuverlässigkeit von Simulationsmodellen basiert (insbesondere bei Mikrosimulationsstudien) auf drei Typen von Analysen (Mertens et al., 2022):

- Validitätsanalysen: Es bedarf einer Validierung des Modells, um dessen Glaubwürdigkeit i. S. d. Realitätsnähe zu untermauern (Li & O'Donoghue, 2012). Die Validitätsanalyse stellt die Übereinstimmung der Modellierungsergebnisse mit anderen beobachteten Daten oder mit theoretischen Überlegungen sicher (Mertens et al., 2022). Sie untersucht damit, ob das Modell das reale Phänomen korrekt abbildet, etwa durch geeignete Annahmen, und gibt Aufschluss über die Verlässlichkeit der Ergebnisse. Fehlerquellen können an verschiedenen Stellen im Prozess auftreten, bspw. bei der Modellspezifikation, im Algorithmus, im Code oder auch bei der Interpretation der Ergebnisse (Bungartz et al., 2013).

- Sensitivitätsanalysen: Sie liefern Informationen über die Robustheit eines Modells, die für die Entscheidungsfindung auf Basis der Modellierungsergebnisse relevant sind. Sensitivitätsanalysen ermitteln die Variabilität der Zielgrößen in Abhängigkeit von der Variabilität der Einflussgrößen und zeigen damit auf, wie stark Unsicherheiten in den Eingabewerten die Ergebnisse beeinflussen (Saltelli et al., 2004). In der Praxis wird typischerweise analysiert, wie sich die Variation schwer oder nicht plausibel zu schätzender Modellparameter auf die Ergebnisse auswirkt (Mertens et al., 2022). Sensitivitätsanalysen dienen somit sowohl der Identifikation einflussreicher Parameter als auch der Abschätzung ihrer Auswirkung auf die Ergebnisse. Dies ermöglicht den Vergleich der Effekte verschiedener Faktoren auf eine ausgewählte Zielgröße (Schmaus, 2023), um die Parameter zu identifizieren, die den größten Einfluss auf das Ergebnis haben.

  Bspw. wird der Basiskonsums zuckergesüßter Getränke in Studien unterschiedlich operationalisiert. Manche nutzen den von der Industrie gemeldeten Absatz, andere den von Verbrauchern angegebenen Verzehr[21]. Sensitivitätsanalysen, die solche Annahmen variieren, können Unsicherheiten in den Simulationsergebnissen aufdecken, die aus der spezifischen Operationalisierung der Parameter resultieren (siehe z. B. Emmert-Fees et al., 2024).

- Unsicherheitsanalysen: Sie quantifizieren die (globale) Unsicherheit im Modell (Saltelli et al., 2008), die durch ungenaue oder unvollständige Eingangsdaten sowie durch zugrunde liegende Modellannahmen entstehen kann. Der Grad der Unsicherheit in den Annahmen ist ausschlaggebend für den Grad an Evidenz, den eine Modellierungsstudie liefern kann. Die Quantifizierung der Unsicherheit im Modell stellt entsprechend eine wesentliche Grundlage für die Entscheidungsfindung dar (Mertens et al., 2022).

Die drei erläuterten Analysen ergänzen sich und ermöglichen gemeinsam eine umfassende Bewertung der Qualität und Zuverlässigkeit eines Modells. Die Modellbewertung schafft Transparenz darüber, inwieweit das Modell eine verlässliche Grundlage für die Entscheidungsfindung bietet.

Einordnung der Studienergebnisse

Die transparente Einordnung von Studienergebnissen einschließlich ihrer Stärken und Schwächen ist unerlässlich, um fundierte Implikationen für Theorie und Praxis abzuleiten. Dabei sind die Limitationen einer Studie (z. B. Mängel in der Datenqualität, Unsicherheiten der Modellparameter, unberücksichtigte Einflussfaktoren oder eine nicht-repräsentative Stichprobe) deutlich zu kommunizieren. Dies ermöglicht eine realistische Einschätzung der Aussagekraft der Ergebnisse und deren korrekte Interpretation. Insbesondere bei einer an die Öffentlichkeit gerichteten Publikation birgt eine unzureichende Kommunikation der Limitationen das Risiko einer verkürzten, irreführenden oder fehlerhaften Rezeption der Ergebnisse. Den Studienautoren obliegt die besondere Verantwortung, eine fundierte Grundlage für kritische Diskussionen zu schaffen – insbesondere dann, wenn eine große Reichweite oder eine breite öffentliche Wahrnehmung der Ergebnisse absehbar ist. Dazu gehört etwa die präzise Differenzierung zwischen korrelativen und kausalen Zusammenhängen sowie eine Einordnung der Generalisierbarkeit der Befunde. Gleichzeitig muss eventuellen Rezipienten bewusst sein, dass die bloße Wiedergabe von Ergebnissen ohne die damit verbundenen Einschränkungen zu Fehlinterpretationen führen kann.

[21] Emmert-Fees et al. (2024) berichtet in seiner methodenfokussierten Studie, dass der von der Industrie gemeldete Absatz zuckergesüßter Getränke etwa 1,86 mal höher war als der von Verbrauchern selbst angegebene Verzehr.

Zusammenfassung der Beurteilungskriterien zur Evaluation von Modellierungsstudien

Datengrundlage

- Untersuchung der **Stichprobe** hinsichtlich ihrer Größe und Zusammensetzung, um zu beurteilen, inwieweit eine strukturelle Übereinstimmung mit der interessierenden Gesamtpopulation in den wesentlichen demografischen Parametern besteht (Voraussetzung für Repräsentativität) und die Stichprobe insgesamt und in den interessierenden Subgruppen ausreicht, um den zufälligen Fehler bei der Schätzung der Effekte in akzeptablen Grenzen zu halten.
- Untersuchung der **weiteren Eingangsdaten** des Modells hinsichtlich ihrer Aktualität, ihrer Vollständigkeit, der Erhebungsmethoden und -zeitpunkte, möglicher zufälliger und nichtzufälliger Fehler, ihrer Anwendbarkeit im Studienkontext sowie ihrer Kompatibilität untereinander, um zu beurteilen, wie gut ihre Passung zum Untersuchungsgegenstand ist.

Modellspezifikation

- Untersuchung des **Modellierungsverfahrens** und der statistischen Analysen hinsichtlich ihrer Eignung und jeweiligen Limitationen, um zu beurteilen, inwieweit sie zur Erreichung des Forschungsziels angemessen sind.
- Untersuchung der **Einfluss- und Zielgrößen** hinsichtlich ihrer Operationalisierung und funktionalen Zusammenhänge, um zu beurteilen, ob die zu untersuchenden realen Phänomene hinreichend präzise abgebildet werden.
- Untersuchung der **weiteren Einflussfaktoren** hinsichtlich ihrer Berücksichtigung und Relevanz im spezifischen Anwendungskontext, um zu beurteilen, ob ein Risiko von Fehlschlüssen, etwa durch die Nichtberücksichtigung von Störfaktoren bzgl. der modellierten Zusammenhänge und Effekte besteht.

Ergebniskommunikation

- Untersuchung der **Ergebnisdarstellung** hinsichtlich Auffälligkeiten in Bezug auf Neutralität, Vollständigkeit, Transparenz, Konsistenz, Verständlichkeit und Nachvollziehbarkeit, um zu beurteilen, inwieweit die Ergebnisse objektiv und unter Minimierung des Risikos einer möglichen Fehlinterpretation präsentiert werden.
- Untersuchung der Vollständigkeit der **Modellbewertung** hinsichtlich der Diskussion von Validität, Sensitivität und Unsicherheit, um zu beurteilen, inwieweit Qualität und Zuverlässigkeit des Modells überprüft werden.
- Untersuchung der **Einordnung der Studienergebnisse** hinsichtlich ihrer Generalisierbarkeit und der Offenlegung von Studienlimitationen, um zu beurteilen, inwieweit eine fundierte Grundlage für kritische Diskussionen und Ableitungen von Implikationen für Theorie und Praxis geschaffen wird.

4 Evaluation von Modellierungsstudien zu den Effekten einer Zuckersteuer

4.1 Studie von Emmert-Fees et al. (2023), Deutschland

Vollständige Quellenangabe zur Studie:

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), Artikel e1004311. https://doi.org/10.1371/journal.pmed.1004311

Executive Summary

Die Studie von Emmert-Fees et al. (2023) untersucht die potenziellen Auswirkungen einer Steuer auf zuckergesüßte Getränke in Deutschland mittels eines Mikrosimulationsansatzes. Dazu werden drei Szenarien modelliert, um das Auftreten verschiedener Krankheiten über einen Zeitraum von 20 Jahren abzuschätzen. Die Simulation ergibt in allen Szenarien eine Verringerung der Krankheitsinzidenzen sowie der damit verbundenen, geschätzten Kosten durch die Einführung einer solchen Steuer. Der Anteil der potenziell reduzierbaren Krankheitsfälle und Gesundheitskosten an den Gesamterkrankungen bzw. -kosten ist jedoch sehr klein. Die Einsparungen bei den Gesundheitskosten belaufen sich in allen Szenarien auf unter 0,04 %. Die methodische Vorgehensweise der Studie ist im Grundansatz nachvollziehbar, jedoch bestehen Kritikpunkte hinsichtlich der Vernachlässigung relevanter Faktoren. Dazu zählen Substitutionseffekte, wichtige Einflussgrößen (jenseits von Alter, Geschlecht und BMI) sowie Heterogenität in den Preiselastizitäten. Zusätzlich weist die zugrunde liegende Datenbasis wesentliche Schwächen auf, welche sich negativ auf die Verlässlichkeit der Modellergebnisse auswirken. Zu den Defiziten gehören veraltete und lückenhafte Datensätze, geringe Stichprobengrößen, sowie die mangelnde Repräsentativität der verwendeten Daten. Kritik besteht im Weiteren hinsichtlich der Darstellung der Modellergebnisse, da diese stärkere Effekte und eine höhere Sicherheit suggeriert als tatsächlich aus den Modellergebnissen ableitbar ist. Insgesamt erfüllt die Studie aus statistisch-methodischer Perspektive nicht die erforderlichen Qualitätsstandards, um als belastbare Grundlage für evidenzbasierte Entscheidungen zu dienen.

4.1.1 Zusammenfassung der Studie

Die Inhalte der Studie von Emmert-Fees et al. (2023) werden nachfolgend zusammengefasst.

Untersuchungsgegenstand

Die Studie simuliert die potenziellen Auswirkungen einer hypothetischen Besteuerung zuckergesüßter Getränke[22] in Deutschland anhand von drei Szenarien:

[22] Erfrischungsgetränke und Fruchtgetränke, denen Zucker zugesetzt wurde (Emmert-Fees et al., 2023, S. 6).

(I) Einführung einer 20 %-igen Ad-Valorem-Verbrauchsteuer[23] auf zuckergesüßte Getränke,

(II) Einführung einer 20 %-igen Ad-Valorem-Verbrauchsteuer auf zuckergesüßte Getränke sowie Fruchtsäfte[24],

(III) Einführung einer gestaffelten Herstellerabgabe[25], die gemäß Annahmen zu einer durchschnittlichen Reduktion des Zuckergehalts in zuckergesüßten Getränken von 30 % führt.

Im Fokus der Simulation stehen zum einen gesundheitliche Auswirkungen in Form der Entwicklung der Inzidenzen verschiedener Erkrankungen (Typ 2 Diabetes mellitus (kurz: *T2DM*), koronare Herzerkrankungen (kurz: *KHK*), Schlaganfälle und Adipositas) in der deutschen Bevölkerung im Alter von 30 bis 90 Jahren für den Zeitraum 2023 bis 2043. Zum anderen werden die damit verbundenen gesundheitsökonomischen Auswirkungen in Form von volkswirtschaftlichen Kosten prognostiziert.

Methodik

Die Studie verwendet einen Mikrosimulationsansatz, um relevante Parameter für 200.000 synthetische Individuen (siehe Kapitel 3) zu simulieren. Diese Population bildet charakteristische Merkmale der deutschen Bevölkerung im Alter zwischen 30 und 90 Jahren nach und wird für den Zeitraum 2013 bis 2043[26] anhand von Zensusdaten und Bevölkerungsprognosen des statistischen Bundesamtes modelliert. Die Ergebnisse der Simulation werden anschließend auf die Gesamtbevölkerung dieser Altersgruppe für die jeweiligen Simulationsjahre hochgerechnet.

In den Szenarien (I) und (II) wird die Wirkung der jeweiligen Steuer auf den Konsum von zuckergesüßten Getränken und Fruchtsäften unter der Annahme eines unveränderten Zuckergehalts dieser Getränke mithilfe der Berechnung von Preiselastizitäten modelliert. In Szenario (III) wird die Wirkung der Steuer unter der Annahme eines infolge von Rezepturänderungen um 30 % reduzierten Zuckergehalts bei gleichbleibendem Konsum zuckergesüßter Getränke modelliert.

In allen Szenarien werden der allgemeine Konsum zuckergesüßter Getränke sowie der spezifische Konsum von Zucker aus zuckergesüßten Getränken und Fruchtsäften (neben weiteren Parametern wie Alter, Geschlecht und BMI) als individuelle Risikofaktoren für die Modellierung der verschiedenen Krankheitsinzidenzen angenommen. Diese werden mithilfe einer Monte-Carlo-Simulation vorhergesagt. Hierbei wird die Existenz von Vorerkrankungen teilweise mit berücksichtigt, siehe Emmert-Fees et al. (2023, Abbildung 1) für eine systematische Darstellung der Abhängigkeiten. Daraus werden wiederum mittels Modellierung der individuellen Risiken Prävalenzen und Mortalitäten berechnet, welche schließlich die Grundlage zur Abschätzung der Auswirkungen auf die Lebensqualität (gemessen in QALYs[27]) und die durch die betrachteten Krankheiten verursachten gesellschaftlichen Kosten bilden.

[23] Die Steuer wird auf den Verkaufspreis aufgeschlagen und der finale Verkaufspreis entsprechend erhöht.
[24] Gemeint sind Fruchtsäfte, Fruchtnektare und andere Säfte (bspw. Gemüsesäfte), die von Natur aus Zucker enthalten. Auch diesen Säften kann zusätzlich Zucker zugesetzt sein (Emmert-Fees et al., 2023).
[25] Die Abgabe erhöht sich gestaffelt mit steigendem Zuckergehalt eines Getränks und soll Hersteller motivieren, die Rezepturen ihrer zuckergesüßten Getränke anzupassen.
[26] Der Simulationszeitraum besteht aus einem Kalibrierungszeitraum (2013-2023), an dem diverse Modellparameter an beobachteten Daten geschätzt werden, und einem Prognosezeitraum (2023-2043), in dem die Wirkung der Szenarien vorhergesagt wird.
[27] QALY ist die Abkürzung für *quality-adjusted life year* und kann mit *qualitätskorrigiertes Lebensjahr* übersetzt werden. Es handelt sich um eine Kennzahl zur Messung des Nutzens medizinischer Behandlungen oder Gesundheitsinterventionen unter Berücksichtigung der Lebensdauer und Lebensqualität der betroffenen Personen.

Ergebnisse

Die Simulation zeigt in allen drei Szenarien eine Reduktion des durchschnittlichen Zuckerverzehrs aus zuckergesüßten Getränken und Fruchtsäften. Die größte mittlere Reduktion des Zuckerverzehrs wird mit durchschnittlich 5,9 Gramm pro Tag, 95 % CI[28] [5,4; 6,0], in Szenario (II) simuliert, gefolgt von 2,3 g/Tag, 95 % CI [2,3; 2,4], in Szenario (III) und 1,0 g/Tag, 95 % CI [0,1; 1,7], in Szenario (I). Die weiteren Modellergebnisse liefern ein uneinheitliches Bild in Bezug auf die Wirksamkeit der einzelnen Szenarien: Szenario (II) zeigt die größten Effekte hinsichtlich der Reduktion bzw. Verzögerung von Schlaganfällen und Adipositas, während Szenario (III) größere Effekte bei der Prävention bzw. Verzögerung von T2DM und koronaren Herzerkrankungen aufweist. Szenario (II) wirkt sich stärker auf die Lebensqualität aus, wohingegen Szenario (III) größere Effekte auf die Lebenserwartung sowie auf gesundheitliche und gesellschaftliche Kosten hat.[29]

Schlussfolgerungen der Studienautoren

Die Studienautoren interpretieren die Modellergebnisse derart, dass alle untersuchten Steuerszenarien substanzielle Gesundheitsvorteile sowie erhebliche Kosteneinsparungen bewirken könnten. Im Vergleich der Szenarien erziele eine vom Zuckergehalt abhängige Abgabe auf zuckergesüßte Getränke mutmaßlich die stärksten Effekte. Die Einbeziehung von Fruchtsäften in eine Besteuerung sei womöglich mit weiteren gesundheitlichen Vorteilen verbunden.

4.1.2 Evaluation der Studie

Datengrundlage

Die Auswahl der synthetischen Stichprobe bestehend aus 200.000 Individuen im Alter von 30 bis 90 Jahren basierend auf aktuellen amtlichen Bevölkerungsdaten ist methodisch angemessen. Zwar könnte eine Simulation der vollen Grundgesamtheit die potenzielle Fehlerquelle durch Stichprobenbildung eliminieren und die Genauigkeit der Schätzwerte erhöhen, aber die gewählte Stichprobengröße erscheint ausreichend, um akzeptable Schätzwerte zu generieren. Die Nichtberücksichtigung von Kindern, Jugendlichen und jungen Erwachsenen in der Simulation wird von den Studienautoren durch die unzureichende Verfügbarkeit relevanter Daten für diese Altersgruppen begründet.

Eine belastbare Simulation der potenziellen Auswirkungen einer hypothetischen Zuckersteuer in Deutschland wird hingegen durch erhebliche Einschränkungen in der Verfügbarkeit und Qualität der notwendigen Eingangsdaten erschwert.

Für die Modellierung der Verteilungen von Konsumverhalten, BMI und Inzidenzen kardiovaskulärer Erkrankungen werden überwiegend Daten aus der KORA-Studie (Helmholtz Zentrum München, 2024) aus den Kohorten S4 (1999), F4 (2007) und FF4 (2014) einbezogen, da neuere und somit repräsentativere Daten nach Angabe der Studienautoren für Deutschland nicht verfügbar waren. Für das Konsumverhalten liegen nur Daten aus der

[28] Ein Konfidenzintervall (englisch: confidence interval; kurz: *CI*) entspricht einem zufälligen Intervall, welches bei bekannter Verteilung der Stichprobe eine vorgegebene Überdeckungswahrscheinlichkeit für den wahren, aber unbekannten Parameter besitzt. Das bedeutet, dass bei einem vorgegebenen Konfidenzniveau von 95 %, welches meist standardmäßig gewählt wird, 95 % der durch Stichproben ermittelten Konfidenzintervalle den wahren Wert überdecken würden.

[29] Die simulierten Unterschiede lassen sich vermutlich darauf zurückführen, dass eine Reduktion des Zuckerverzehrs aus zuckergesüßten Getränken tendenziell andere Bevölkerungsgruppen betrifft als eine Reduktion des Zuckerverzehrs durch verringerten Fruchtsaftkonsums, denn der Konsum von zuckergesüßten Getränken und Fruchtsäften zeigt deutliche alters- und geschlechtsspezifische Variationen (siehe Abbildungen D-J in Appendix S1 in Emmert-Fees et al., 2023).

jüngsten KORA-Kohorte vor, welche keine Personen unter 38 Jahren enthält. Daher nutzen Emmert-Fees et al. (2023) zur Modellierung des Konsumverhaltens ergänzend Daten aus der Nationalen Verzehrsstudie II (NVS II; Max Rubner-Institut, n. d.). Die Konsumdaten aus den beiden Quellen unterscheiden sich jedoch aufgrund unterschiedlicher Erhebungsmethoden stark (siehe auch Emmert-Fees et al., 2023, Appendix S1, Abbildungen D und H). Emmert-Fees et al. (2023) schätzen den Konsum schließlich auf Basis beider Datensätze, doch es bleibt unklar, welche der beiden Datenquellen den Konsum realistischer abbildet.

Weitere Modellparameter basieren auf Studien mit sehr kleinen Stichprobengrößen[30], was die Verlässlichkeit der Daten erheblich einschränkt. Einige der Datenquellen sind auch bereits veraltet[31] oder es fehlen die nötigen Variablen, welche dann anhand anderer Variablen geschätzt werden müssen[32], was weitere Unsicherheit in die Modellergebnisse bringt. Für andere Modellparameter standen für Deutschland überhaupt keine geeigneten Daten zur Verfügung[33], sodass hilfsweise Datengrundlagen insbesondere aus den USA und Großbritannien herangezogen werden. Die Übertragbarkeit auf die deutsche Bevölkerung ist jedoch fragwürdig, da länderspezifische Unterschiede in Hersteller- und Konsumentenverhalten, Gesundheitsvorsorge und weiteren relevanten Faktoren anzunehmen sind.

Die verwendeten Modellparameter, wie etwa zur Effektschätzung von Konsum und BMI auf Krankheitsinzidenzen, stammen überwiegend aus Beobachtungsstudien, meist Querschnittstudien, die im Vergleich zu Interventionsstudien eine geringere Evidenzbasis bieten. Insbesondere können Querschnittstudien lediglich korrelative und keine kausalen Zusammenhänge aufzeigen. Diese Einschränkung resultiert aus der fehlenden Verfügbarkeit von Interventionsstudien in diesem Anwendungskontext. Diese methodische Problematik in der Studie von Emmert-Fees et al. (2023) wurde auch von Nawroth und Kumar (2024) aufgegriffen und in einer Replik der Studienautoren diskutiert (Laxy & Emmert-Fees, 2025).

Für die Parametrisierung des Modells werden Daten aus insgesamt 36 unterschiedlichen Quellen herangezogen. Eine vollständige Übersicht dieser Datenquellen ist in Emmert-Fees et al. (2023), Appendix S1, Tabelle A, aufgeführt. Allein diese große Menge an unterschiedlichen Datenquellen, die teilweise mit abweichenden Definitionen und Erhebungsmethoden für die einzelnen Kenngrößen arbeiten, führt zu einem erheblichen Maß an Unsicherheit im Modell.

Modellspezifikation

Mikrosimulationsmodelle sind grundsätzlich die Methode der Wahl, um komplexe Zusammenhänge in großen Kohorten zu simulieren (siehe Anhang A.1.2). Die Validität der Modellergebnisse hängt jedoch maßgeblich davon ab, inwieweit verlässliche Daten für die einzelnen Teilzusammenhänge verfügbar sind und wie stark die Modellergebnisse von den zugrunde liegenden Annahmen und Eingabeparametern beeinflusst werden. Es ist gleichwohl zu betonen, dass ein Mikrosimulationsmodell keine kausalen Zusammenhänge nachweisen kann. Kausalität wird im Modell als Annahme vorausgesetzt und muss daher bereits im Vorfeld für jeden einzelnen

[30] Bspw. werden zur Abschätzung der Produktivitätseinbußen aufgrund von Schlaganfällen Ergebnisse basierend auf 151 Patienten aus Winter et al. (2008) herangezogen.
[31] Konsumdaten beruhend auf der NVS II stammen aus den Jahren 2005-2007 (Max Rubner-Institut, n. d.), Daten zu Krankheitskosten aus Winter et al., 2008 und Icks et al., 2013 beziehen sich auf das Jahr 1999.
[32] Bspw. sind die Risikofaktoren für KHK und Schlaganfälle im verwendeten Datensatz nicht erfasst und werden aus jenen für kardiovaskuläre Erkrankungen geschätzt.
[33] Bspw. fehlt es an Daten zur Steuerweitergaberate, Prävalenzen von KHK und Schlaganfällen sowie Korrelationen von Konsumverhalten mit BMI, KHK und T2DM für Deutschland.

Simulationsschritt hinreichend belegt sein. Die vorliegende Studie kann somit aus methodischen Gründen keinen kausalen Effekt nachweisen, sondern nur eine Abschätzung möglicher Effektstärken geben, falls ein solcher Kausalzusammenhang vorliegt.

Hinsichtlich der Modellspezifikation sind folgende Kritikpunkte anzumerken:

- Vernachlässigung der Heterogenität von Preiselastizitäten: Individuelle Preiselastizitäten können durch verschiedene Einflussfaktoren, wie etwa dem durchschnittlichen persönlichen Konsum von zuckergesüßten Getränken, beeinflusst werden (Blake et al., 2019). In der vorliegenden Studie wird jedoch angenommen, dass bei Preisanstiegen alle Konsumenten ihren Konsum in gleicher Weise einschränken. Dies führt zu einer erheblichen Pauschalisierung der Studienergebnisse.
- Verallgemeinerung der Zuckerreduktion in zuckergesüßten Getränken: In Szenario (III) wird infolge der Rezepturänderungen stark vereinfacht eine fixe Reduktion des Zuckergehalts um 30 % bei allen zuckergesüßten Getränken angenommen. In der Realität sind jedoch Abweichungen davon zu erwarten, welche die simulierten gesundheitlichen und gesundheitsbezogenen Effekte beeinflussen können, wie die Studienautoren in einer Sensitivitätsanalyse selbst darlegen[34].
- Nichtberücksichtigung zeitlicher Trends: Zeitliche Entwicklungen in Bezug auf Preiselastizitäten, Konsumverhalten, den Zuckergehalt von zuckergesüßten Getränken (mit Ausnahme der Rezepturänderung im Rahmen von Szenario (III)) sowie die Gesundheitsversorgung werden nicht betrachtet. Die Modellergebnisse bilden also die realen Phänomene und ihre Beziehungen vermutlich nicht hinreichend präzise ab.
- Begrenzte Berücksichtigung von Einflussfaktoren: Neben Alter, Geschlecht und BMI werden keine weiteren potenziellen Einflussfaktoren berücksichtigt. Dabei können körperliche Aktivität wie auch der sozioökonomische Status (z. B. Bildungsstand, Haushaltseinkommen) jedoch weitere wichtige Determinanten sein. Insbesondere werden Substitutionseffekte nur in Bezug auf Fruchtsäfte berücksichtigt. Die von Emmert-Fees et al. (2023) simulierten Effekte müssen entsprechend nicht zwangsläufig in der dargestellten Form existieren.

Ergebniskommunikation

Die Modellergebnisse werden überwiegend in Form von absoluten Summenwerten über die 20-jährige Simulationsperiode dargestellt, ohne den Bezug zur jeweiligen Ausgangsbasis herzustellen. Dies birgt das Risiko von Fehlrezeptionen und in der Folge unrealistisch hoher Erwartungen an die möglichen Effekte einer Steuer. Beispielhaft genannt sei die geschätzte Reduktion der Gesundheitskosten, die mit Beträgen zwischen 2,26 Milliarden Euro, 95 % CI [1,19; 3,60], in Szenario (I) und 3,85 Milliarden Euro, 95 % CI [2,07; 6,08], in Szenario (III) angegeben wird. Unter der vereinfachenden Annahme einer gleichmäßigen Verteilung der Einsparungen über den Simulationszeitraum von 20 Jahren entspricht dies jährlichen Einsparungen von etwa 110 Millionen Euro in Szenario (I) bis 190 Millionen Euro in Szenario (III). Zum Vergleich: Die Gesamtkosten des deutschen Gesundheitssystems beliefen sich im Jahr 2022 laut Statistischem Bundesamt (2024) auf 498 Milliarden Euro. Die von Emmert-Fees et al. (2023) berechneten Kosteneinsparungen infolge einer Besteuerung zuckergesüßter

[34] Genaue Erläuterungen zur Sensitivitätsanalyse erfolgen im Abschnitt *Ergebniskommunikation*.

Getränke entsprächen demnach einer relativen Kostenreduktion von etwa 0,023 % in Szenario (I) und 0,039 % in Szenario (III) – ein Anteil, der im Kontext der Gesamtkosten als gering einzustufen ist.

Die Schlussfolgerungen der Autoren sind größtenteils vorsichtig formuliert (bspw. durch Formulierungen wie „could lead to“ und „likely to lead to“), doch suggerieren einige Formulierungen (bspw. „would lead to“) eine Gewissheit, die nicht im Einklang mit den Limitationen des Modells steht. Insgesamt werden die Modellergebnisse und die daraus abgeleiteten Schlussfolgerungen in der Studie als deutlich verlässlicher dargestellt, als dies gerechtfertigt erscheint. So schreiben die Studienautoren bspw.: „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, S. 17). Zwar werden bestimmte Unsicherheiten in den Eingangsdaten angemessen modelliert, doch bleibt die Berücksichtigung anderer potenzieller Unsicherheiten unvollständig oder wird gänzlich unterlassen. Beispielsweise werden die Unsicherheiten durch Annahmen zur Methodik (bspw. für die Berechnung der Preiselastizitäten) nicht berücksichtigt. Außerdem wird die Unsicherheit hinsichtlich der Wirkung einer gestaffelten Herstellerabgabe auf den Zuckergehalt der zuckergesüßten Getränke in Szenario (III) nicht modelliert und die Nichtrepräsentativität verschiedener Datensätze sowie Unsicherheiten in den gesellschaftlichen Kosten durch frühzeitigen Tod nicht berücksichtigt. Außerdem werden lediglich sehr kleine Schwankungsbreiten bei den Krankheitskosten angenommen. Dies schränkt die Validität und Generalisierbarkeit der ermittelten Modellergebnisse ein. Da diese Unsicherheitsquellen nicht berücksichtigt wurden, spiegeln die in der Studie angegebenen Konfidenzintervalle, die nur die Unsicherheit aus zufälligen, aber nicht aus nicht-zufälligen Fehlern abbilden, die tatsächliche Variabilität nicht angemessen wider.

In der Studie werden verschiedene Sensitivitätsanalysen (siehe Anhang A.3.2) zu den zugrunde liegenden Modellparametern der Steuerszenarien durchgeführt[35] (siehe Emmert-Fees et al., 2023, Appendix S1, Tabelle V). Eine der Sensitivitätsanalysen (Sensitivitätsanalyse 4) zeigt bspw., dass die Modellergebnisse einer gestaffelten Herstellerabgabe eine erhebliche Sensitivität gegenüber der angenommenen Reduktion des Zuckergehalts der zuckergesüßten Getränke aufweisen: Die Analyse ist analog zu Szenario (III) konzipiert, geht jedoch von einer Zuckerreduktion in den Getränken um lediglich 10 % aus. Die simulierten Effekte fallen durchgängig geringer aus als in den drei Hauptszenarien. Ein detaillierter Vergleich der Ergebnisse der Sensitivitätsanalyse mit Szenario (III) deutet darauf hin, dass die Reduktion des Zuckergehalts annähernd linear mit den Modellergebnissen korreliert. Dies impliziert, dass bereits eine Abweichung von der angenommenen Zuckerreduktion in den Getränken um ein Fünftel (eine Reduktion von 24 % statt 30 %) zu relativen Unsicherheiten um etwa ein Fünftel in den Modellergebnissen führt. Das lässt die Bewertung von Szenario (III) als empfehlenswerte Strategie fragwürdig erscheinen, da die resultierenden Effekte vergleichbar (in Bezug auf Inzidenzen von T2DM und KHK sowie daraus abgeleitete Werte) oder sogar schlechter (in Bezug auf Inzidenzen von Adipositas und Schlaganfällen sowie daraus abgeleitete Werte) als jene von Szenario (II) ausfallen würden. Sensitivitätsanalysen dienen der Abschätzung von Unsicherheiten. Eine prominentere Darstellung der durchgeführten Analysen im Haupttext der Studie sowie eine Ausweitung der Sensitivitätsanalysen auf andere Unsicherheitsquellen im Modellierungsprozess wäre daher angebracht.

[35] Höhe der Ad-Valorem-Verbrauchsteuer, (keine) Substitution durch Fruchtsäfte, geringeres Ausmaß der Rezepturänderung seitens der Hersteller, Kombination von gestaffelter Herstellerabgabe und Ad-Valorem-Verbrauchsteuer

Eine Nachfolgestudie aus ähnlichem Autorenkreis thematisiert die Unsicherheiten der Simulation von Emmert-Fees et al. (2023): Emmert-Fees et al. (2024) analysieren in einem vergleichbaren Kontext, wie unterschiedliche Annahmen (z. B. (Nicht-)Berücksichtigung von Substitution; spezifische Ausgestaltung der Steuer) und Modellparameter (z. B. Preiselastizitäten der Nachfrage) die geschätzten Auswirkungen einer Besteuerung zuckergesüßter Getränke beeinflussen.

Die ausgeprägte Sensitivität der Simulationsergebnisse gegenüber bestimmten Annahmen und Modellparametern wirft erhebliche Zweifel an der Belastbarkeit der in Emmert-Fees et al. (2023) präsentierten Modellergebnisse und der daraus gezogenen Schlussfolgerungen auf.

Emmert-Fees et al. (2023) weisen gleichwohl transparent auf mehrere Limitationen ihrer Untersuchung hin[36], allerdings werden manche Einschränkungen nur unzureichend erwähnt oder bleiben gänzlich unberücksichtigt. Hierzu gehören insbesondere die (mögliche) Nichtrepräsentativität von Studien zu (1) Steuerweitergaberate, (2) Zuckerreduktion in zuckergesüßten Getränken durch Reformulierung von Rezepturen und (3) Einfluss von Zuckerverzehr auf den BMI, sowie die Vernachlässigung von Energiesubstitution bei verringerter Aufnahme von zuckergesüßten Getränken/Fruchtsäften durch andere kalorische flüssige und feste Lebensmittel.

[36] Erwähnte Limitationen: Nichtrepräsentativität der Basisdaten aus der KORA-Studie; keine Modellierung von zeitlichen Trends im Konsumverhalten; keine Berücksichtigung von Effekten abseits von Diabetes und kardiovaskulären Erkrankungen; keine Modellierung des Langzeitzuckerkonsums; keine Berücksichtigung der Heterogenität der Preiselastizitäten; keine Berücksichtigung der Inflationskrise der Jahre 2022/23; keine Modellierung von Personen unter 30 Jahren; keine genaue Modellierung einer gestaffelten Abgabe auf den Zuckergehalt der betroffenen zuckergesüßten Getränke; keine Berücksichtigung der Kosten für Einführung der Steuern und Reformulierung seitens der Hersteller; nicht repräsentative oder veraltete Daten für Inzidenzen und Prävalenzen; Annahmen zu Zuckergehalt; Nichtberücksichtigung weiterer relevanter Einflussfaktoren; sehr kleine Stichprobengrößen

Zusammenfassung der Limitationen der Studie von Emmert-Fees et al. (2023)

Datengrundlage

- **Stichprobe:** Nachbildung der deutschen Bevölkerung (beschränkt auf 30- bis 90-Jährige) basierend auf aktuellen amtlichen Bevölkerungsdaten grundsätzlich als repräsentativ anzunehmen, jedoch ohne Gültigkeit für Kinder und junge Erwachsene.
- **Weitere Eingangsdaten:** Daten aus insgesamt 36 Datenquellen; teilweise veraltet (Konsumdaten teils aus 2005-2007, Daten Krankheitskosten teils aus 1999); teils basierend auf sehr kleinen Stichproben (z. B. Produktivitätseinbußen durch Schlaganfälle anhand 151 Patienten); Übertragbarkeit auf Deutschland nicht immer gewährleistet (z. B. Prävalenzen von Schlaganfällen); eingeschränkte Kompatibilität von Datensätzen (Konsumdaten); fehlende Variablen (z. B. Risikofaktoren für Schlaganfälle); überwiegend basierend auf Beobachtungsstudien (z. B. zur Effektschätzung von Konsum und BMI auf Krankheitsinzidenzen); daher insgesamt eingeschränkte Validität der Modellergebnisse.

Modellspezifikation

- **Modellierungsverfahren:** Kausalität wird bei Mikrosimulationsmodellen (wie hier auf Basis von Monte-Carlo-Simulationen) ex ante bei der Modellkonstruktion unterstellt, nicht aber durch das Modell selbst nachgewiesen.
- **Einfluss- und Zielgrößen:** Nichtberücksichtigung der Heterogenität von Preiselastizitäten führt zu erheblicher Pauschalisierung der Ergebnisse; aufgrund zahlreicher weiterer Annahmen (z. B. Reduktion des Zuckergehalts um 30 % bei allen Getränken in Szenario (III), Konstanz des Konsumverhaltens, gleichbleibende Verteilung von Krankheitsinzidenzen und -kosten) bilden die Modellergebnisse die realen Phänomene und deren Beziehungen vermutlich nicht hinreichend präzise ab.
- **Weitere Einflussfaktoren:** Unzureichende Berücksichtigung von Substitution sowie Nichtberücksichtigung weiterer relevanter Faktoren (z. B. körperliche Aktivität, sozioökonomischer Status), sodass modellierte Effekte nicht zwangsläufig in der dargestellten Form existieren müssen.

Ergebniskommunikation

- **Ergebnisdarstellung:** Trotz transparenter Angabe lediglich begrenzte Aussagekraft der Konfidenzintervalle; ausschließliche Darstellung in absoluten Summenwerten ohne angemessene Kontextualisierung vermittelt den Eindruck von stärkeren Effekten, als tatsächlich vorliegen.
- **Modellbewertung:** Sensitivitätsanalysen relativieren Ergebnisse der Hauptanalyse und stellen sie damit in Frage; lediglich vereinzelte Sensitivitätsanalysen und fehlende Überprüfung der Auswirkungen zentraler Modellannahmen erschweren die Einschätzung modellbedingter Unsicherheit.
- **Ergebniseinordnung:** Unzureichende Diskussion der Limitationen; vereinfachende Einordnung der Modellergebnisse (z. B. Darstellung der Herstellerabgabe als Optimallösung ohne Berücksichtigung von Variabilität); Belastbarkeit der Modellergebnisse ist aufgrund Unsicherheiten zu hinterfragen; Schlussfolgerungen suggerieren höhere Sicherheit als aus den Modellergebnissen ableitbar.

Gesamtbeurteilung der Studie von Emmert-Fees et al. (2023)

Die methodische Vorgehensweise der Studie von Emmert-Fees et al. (2023) ist im Grundansatz nachvollziehbar. Allerdings bestehen methodische Limitationen, insbesondere in Bezug auf die Vernachlässigung relevanter Einflussfaktoren. Dazu zählen Substitutionseffekte, wichtige Einflussgrößen (jenseits von Alter, Geschlecht und BMI) sowie Heterogenität in den Preiselastizitäten. Zudem weist die zugrunde liegende Datenbasis wesentliche Schwächen auf, die die Verlässlichkeit der Modellergebnisse beeinträchtigen. Hierzu zählen veraltete und lückenhafte Datensätze, geringe Stichprobengrößen, sowie eine eingeschränkte Übertragbarkeit auf die untersuchte Population. Kritik besteht im Weiteren hinsichtlich der Darstellung der Modellergebnisse, da diese stärkere Effekte und eine höhere Sicherheit suggeriert, als tatsächlich aus den Modellergebnissen ableitbar ist. Insbesondere die Präsentation der Ergebnisse in absoluten Summenwerten ohne angemessene Kontextualisierung, sowie die unzureichende Berücksichtigung von Limitationen sind in diesem Zusammenhang problematisch. Angesichts dieser Schwächen sind die Modellierungsergebnisse und Schlussfolgerungen der Studienautoren als unzuverlässig einzustufen, sodass die Studie keine belastbare Grundlage für politische Entscheidungen darstellt.

4.2 Studie von Schwendicke und Stolpe (2017), Deutschland

Vollständige Quellenangabe zur Studie:

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

Executive Summary

Die Studie von Schwendicke und Stolpe (2017) untersucht die potenziellen kurzfristigen Auswirkungen einer hypothetischen Besteuerung zuckergesüßter Getränke auf Übergewicht und Adipositas in Deutschland mithilfe einer Monte-Carlo-Simulation. Die Studienautoren kommen zu dem Schluss, dass eine solche Steuer insbesondere bei jüngeren Menschen und Haushalten mit niedrigerem Einkommen zu einer signifikanten Verringerung von Übergewicht und Adipositas führen könnte. Allerdings weist die Studie methodische Schwächen auf, welche die Belastbarkeit und Generalisierbarkeit der Ergebnisse erheblich einschränken. So spiegeln die zugrunde liegenden Daten, wie etwa Preiselastizitäten aus US-amerikanischen Studien sowie veraltete Konsumdaten, die spezifischen sozialen und ökonomischen Rahmenbedingungen in Deutschland nicht angemessen wider. Zudem beruhen die Modellergebnisse auf stark vereinfachenden Annahmen, darunter die lineare Beziehung zwischen Kalorienaufnahme und Gewicht ohne Berücksichtigung kompensatorischer Effekte, die stabile Konsumreaktion nach Einführung der Steuer sowie die unzureichende Berücksichtigung möglicher Substitutionseffekte. Insgesamt beeinträchtigen diese methodischen Einschränkungen die Verlässlichkeit der Ergebnisse und berichteten Schlussfolgerungen erheblich. Insgesamt erfüllt die Studie aus statistisch-methodischer Perspektive nicht die erforderlichen Qualitätsstandards, um als belastbare Grundlage für evidenzbasierte Entscheidungen zu dienen.

4.2.1 Zusammenfassung der Studie

Die Inhalte der Studie von Schwendicke und Stolpe (2017) werden nachfolgend zusammengefasst.

Untersuchungsgegenstand

Die Studie simuliert die potenziellen kurzfristigen[37] Auswirkungen einer hypothetischen 20%-igen Ad-Valorem-Verbrauchsteuer auf zuckergesüßte Getränke in Deutschland. Im Fokus stehen Auswirkungen auf das Konsumverhalten und potenzielle Effekte auf die Prävalenz von Übergewicht und Adipositas.

Methodik

Betrachtet wird die deutsche Bevölkerung im Alter von 15 bis 79 Jahren. Die Steuerauswirkungen werden mittels Monte-Carlo-Simulation geschätzt. Zunächst werden auf Basis von Preiselastizitäten der Nachfrage und Kreuzpreiselastizitäten (im weiteren Verlauf dieses Abschnitts kurz: „Preiselastizitäten“) die Veränderungen im Konsumverhalten infolge der Steuer simuliert. Daraus werden Veränderungen in der Kalorienaufnahme abgeleitet und deren Effekte auf den BMI geschätzt, welcher zur Ermittlung der Prävalenz von Übergewicht und Adipositas dient.

[37] Es werden die Effekte betrachtet, die innerhalb weniger Monate nach der Einführung der Steuer auftreten.

Ergebnisse

Die Simulation ergibt, dass eine Zuckersteuer in Deutschland kurzfristig zu einer Verringerung von etwa 1,03 Mio. Fällen von Übergewicht und 497.000 Fällen von Adipositas führen könnte[38]. Die größten Effekte werden für jüngere Altersgruppen sowie einkommensschwächere Bevölkerungsgruppen simuliert. Bei älteren Personen hingegen fallen die Effekte lediglich marginal oder sogar entgegengesetzt aus.

Schlussfolgerungen der Studienautoren

Die Studienautoren interpretieren die Ergebnisse dahingehend, dass eine 20%-ige Steuer auf zuckergesüßte Getränke in Deutschland positive Effekte auf die öffentliche Gesundheit haben dürfte, vor allem bei jüngeren Menschen und einkommensschwächeren Personengruppen. Sie weisen darauf hin, dass die Ergebnisse stark von den zugrunde liegenden Annahmen und den verwendeten Daten zu Preiselastizitäten abhängen.

4.2.2 Evaluation der Studie

Datengrundlage

Die (synthetische) Stichprobe basiert auf einer Simulation der deutschen Bevölkerung im Alter von 15 bis 79 Jahren, für die amtliche Bevölkerungsdaten in Form der Fortschreibung des Zensus von 2011 zum Stand Mai 2012 (Statistisches Bundesamt, n. d. a) verwendet wurden. Die Bevölkerungsdaten waren zum Zeitpunkt der Analyse vor der Studienveröffentlichung 2017 zwar aktuell, könnten jedoch aufgrund wesentlicher demografischer Veränderungen seither – etwa durch Migration oder sozioökonomische Entwicklungen – von der heutigen deutschen Bevölkerung abweichen. Dies schränkt die Aussagekraft der Ergebnisse im Hinblick auf ihre aktuelle Gültigkeit stark ein.

Die verwendeten Daten zum Basiskonsum stammen aus der Nationalen Verzehrsstudie II (NVS II; Max Rubner-Institut, n. d.) und sind nach Geschlecht, Alter und Einkommen differenziert. Da die Daten der NVS II zwischen November 2005 und Januar 2007 erhoben wurden, waren sie Daten bei der Studienveröffentlichung bereits etwa zehn Jahre alt. Mögliche Verzerrungen der Ergebnisse durch Veränderungen in Konsum- und Markttrends zwischen Datenerhebung und Veröffentlichung sind daher anzunehmen, werden jedoch nicht thematisiert. Zudem sind weitere Veränderungen im Lebensmittelverzehr seit der Veröffentlichung 2017 bis heute möglich, was die Gültigkeit der Ergebnisse einschränkt. Die Daten zu Körperindizes stammen aus dem Mikrozensus 2013 sowie der KiGGS-Studie[39]. Während der Mikrozensus ausschließlich auf Befragungsdaten basiert, enthält die KiGGS-Studie – zumindest in ihrer Basiserhebung –objektive Messungen von Körpermaßen wie Größe und Gewicht; spätere Erhebungswellen hingegen umfassen größtenteils Befragungsdaten. Befragungsbasierte Daten können durch sozial erwünschtes Antwortverhalten (z. B. falsche Angaben vom Körpergewicht) beeinflusst werden, was zu Verzerrungen in den Ergebnissen führen kann (Schüller, 2015). Die Datensätze zu Basiskonsum und Körperindizes sind außerdem nicht vollständig kompatibel[40]. Die Inkompatibilität der Daten kann zu einer potenziell verzerrten Gewichtung einzelner Altersgruppen führen und somit Abweichungen in den Ergebnissen verursachen. Die Studienautoren weisen bereits in einer früheren Studie auf die Problematik der Inkompatibilität

[38] Die Studie liefert hierzu lediglich Punktschätzer, keine Konfidenzintervalle oder Standardabweichungen.
[39] Es handelt sich dabei um eine Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland (Robert Koch-Institut, 2024).
[40] Während die Bevölkerungsdaten aus der Fortschreibung des Zensus von 2011 (Stand Mai 2012) in Zehnjahresintervalle (z. B. 20-29 sowie 30-39 Jahre) klassifiziert sind, nutzt die NVS II andere Altersklassen (z. B. 19-24 sowie 25-34 Jahre).

hin (Schwendicke et al., 2016), diskutieren jedoch an keiner Stelle, welche Konsequenzen dies für die Validität ihrer Studienergebnisse haben könnte.

Die Daten zu den Preiselastizitäten stammen aus US-amerikanischen Studien und sind bis auf Einkommensgruppen nicht weiter (etwa nach Alter, Geschlecht oder Konsumgewohnheiten) differenziert. Auf Deutschland sind diese Daten nur bedingt übertragbar, da erhebliche Unterschiede in ökonomischen, kulturellen und sozialen Rahmenbedingungen vorliegen. Somit ist fragwürdig, ob die Studienergebnisse dem Grunde nach valide für Deutschland sind. Weiterhin bleibt unklar, ob die Preiselastizitäten auch Konsumveränderungen im gastronomischen Bereich berücksichtigen und somit möglicherweise ein unvollständiges Bild des Konsumverhaltens zeichnen.

Modellspezifikation

Mit einer Monte-Carlo-Simulation (siehe Anhang A.3.1) werden die Auswirkungen einer hypothetischen Steuer auf zuckergesüßte Getränke auf deren Konsum und auf den BMI simuliert. Das Verfahren ist grundsätzlich geeignet, um Unsicherheiten in den Eingangsparametern zu berücksichtigen. Allerdings kann ein solches Modell keine kausalen Zusammenhänge nachweisen. Kausalität wird als Annahme vorausgesetzt und ist daher bereits im Vorfeld für jeden Simulationsschritt zu belegen. Da dies nicht erfolgt, kann die vorliegende Studie lediglich Schätzungen möglicher Effektstärken liefern – unter der Annahme, dass Kausaleffekte vorliegen. Die Anzahl von 100 Simulationsdurchläufen pro Gruppe ist ungewöhnlich klein. Damit werden keine robusten Schätzungen gewährleistet. Zudem wird nicht ordentlich dokumentiert, welche Verteilungen den variierenden Faktoren in der Simulation zugrunde gelegt werden, was die Nachvollziehbarkeit der Modellierung sowie eine unabhängige Überprüfung und Replikation der Ergebnisse erschwert.

Die Modellierung der Zusammenhänge der Einfluss- und Zielgrößen weist mehrere Limitationen auf. Die BMI-Veränderung wird als Funktion in Abhängigkeit des veränderten Kalorienkonsums modelliert, basierend auf der Annahme einer linearen Beziehung zwischen Kalorienaufnahme und Körpergewicht. Diese Vereinfachung vernachlässigt jedoch, dass Übergewicht und Adipositas durch komplexe Prozessketten beeinflusst werden (Hummel et al., 2013). Weder individuelle Stoffwechselunterschiede noch die dynamischen Wechselwirkungen zwischen Ernährung, Bewegung sowie sozialen und psychischen Faktoren werden ausreichend berücksichtigt[41] Die Annahme einer linearen Beziehung zwischen Energieaufnahme und Körpergewicht vereinfacht die tatsächlichen Zusammenhänge stark und vermittelt damit ein verzerrtes Bild der zugrunde liegenden Mechanismen. Zudem betrachtet die Studie ausschließlich kurzfristige Effekte in Form einer Anpassung des Konsumverhaltens innerhalb weniger Monate nach Einführung der Steuer und geht davon aus, dass Verhaltensänderungen langfristig stabil bleiben. Mögliche langfristige Anpassungen, wie eine Gewöhnung der Konsumenten an höhere Preise oder Änderungen im Konsumverhalten, werden ignoriert. Für eine realistische Modellierung der BMI-Veränderung müssten nichtlineare Effekte und komplexere Wechselwirkungen berücksichtigt werden. Außerdem werden Preiselastizitäten stark pauschalisiert, indem sie ausschließlich nach zwei Einkommensgruppen (niedrig/mittel vs. hoch) differenziert betrachtet werden, während Unterschiede nach Alter, Geschlecht und Basiskonsum unberücksichtigt bleiben. Diese Annahme vernachlässigt die heterogenen Reaktionen auf Preisänderungen und schmälert damit die Aussagekraft der Ergebnisse.

[41] Im Gutachten wird kürzer auch von „kompensatorischen Effekten“ gesprochen.

Weitere relevante Einflussfaktoren werden in der Studie nur unzureichend berücksichtigt. Zwar werden Fruchtsäfte und Milch als Substitute für zuckergesüßte Getränke berücksichtigt, andere mögliche Substitute werden aber nicht in die Simulation einbezogen. Weitere potenziell relevante Faktoren wie regionale Unterschiede, gesundheitspolitische Rahmenbedingungen oder demografische Trends fehlen ebenfalls, was die Belastbarkeit der Ergebnisse schwächt.

Ergebniskommunikation

Negativ fällt auf, dass Unsicherheiten in den Ergebnissen nicht präzise dargestellt werden. Statt vollständiger Konfidenzintervalle geben die Studienautoren häufig nur Standardabweichungen an, und auf Populationsebene fehlen selbst diese. Dies erschwert die Beurteilung der Unsicherheiten, die jedoch besonders für abzuleitende Maßnahmen von großer Bedeutung sind. Die Ergebnisdarstellung der Punktschätzer erfolgt überwiegend in absoluten Zahlen. Relative Angaben, die eine bessere Vergleichbarkeit zwischen Subgruppen wie Alters- oder Einkommensklassen ermöglichen würden, fehlen weitgehend. Eine konsistente Darstellung sowohl in absoluten als auch in relativen Werten würde die Interpretierbarkeit der Effekte deutlich verbessern.

In der Studie werden Sensitivitätsanalysen (siehe Anhang A.3.2) durchgeführt, die sich jedoch auf vereinzelte Parameter wie Kreuzpreiselastizitäten und Steuerweitergaberaten beschränken. Unsicherheiten, die aus zentralen Modellannahmen resultieren (bspw. die lineare Beziehung zwischen Energieaufnahme und Gewichtsveränderung oder die Annahme eines langfristig stabilen Konsums), werden hingegen nicht variiert. Umfassende Sensitivitätsanalysen hinsichtlich zentraler Modellannahmen sind unerlässlich, um die Robustheit und Verlässlichkeit der Studienergebnisse fundierter zu bewerten.

Zwar sprechen die Studienautoren wesentliche Limitationen ihrer Studie offen an, allerdings fehlt eine vertiefte Diskussion dieser Einschränkungen. Die Studienautoren weisen auf die potenziell veralteten Konsumdaten hin und erwähnen die Möglichkeit verzerrter Körperindizes, die auf Befragungsdaten basieren. Auch die Übertragung von Preiselastizitäten aus US-amerikanischen Studien wird kritisch betrachtet. Die fehlende Differenzierung der Preiselastizitäten nach Alter und Basiskonsum wird zwar thematisiert, jedoch bleibt eine detaillierte Analyse möglicher Verzerrungen und deren Auswirkungen auf die Ergebnisse aus. Die Studienautoren diskutieren stark vereinfachende Annahmen wie die lineare Beziehung zwischen Kalorienaufnahme und Gewicht sowie die begrenzte Berücksichtigung von Substitutionseffekten. Langfristige Verhaltensanpassungen im Konsum werden erwähnt, aber nicht differenziert in Bezug auf ihre potenzielle Relevanz für die Ergebnisse diskutiert. Die Studienautoren erkennen die Unsicherheiten ihrer Modellannahmen an, stellen diese jedoch nicht in ausreichender Tiefe dar. Insgesamt werden die möglichen Auswirkungen all dieser Limitationen auf die Aussagekraft der Ergebnisse nur unzureichend beleuchtet.

Die Studienautoren schließen aus ihren Ergebnissen, dass die Einführung einer Steuer auf zuckergesüßte Getränke in Deutschland Übergewicht und Adipositas reduzieren dürfte. Zwar weisen sie transparent darauf hin, dass die geschätzten Effekte nur im Rahmen der getroffenen Annahmen gelten und maßgeblich von den verwendeten Preiselastizitätsdaten abhängen, dennoch erwecken ihre Schlussfolgerungen den Eindruck kausaler Wirkzusammenhänge, die das Modell allerdings nicht belegen kann.

Zusammenfassung der Limitationen der Studie von Schwendicke und Stolpe (2017)

Datengrundlage

- **Stichprobe:** Nachbildung der deutschen Bevölkerung (beschränkt auf 15- bis 79-Jährige) basierend auf amtlichen Bevölkerungsdaten aus 2012, die aufgrund wesentlicher demografischer Veränderungen seither (z. B. durch Migration) von der heutigen deutschen Bevölkerung abweichen können.
- **Weitere Eingangsdaten:** Veraltete Konsumdaten (NVS II aus 2005-2007) bilden Konsum- und Marktentwicklungen nicht adäquat ab und verzerren Ergebnisse; überwiegend Befragungsdaten (statt Messungen) zu Körperindizes, die aufgrund von Selbstauskunft verzerrt sein können; Inkompatibilität von Konsum- und Körperdaten, dadurch potenziell verzerrte Altersgruppengewichtung und Abweichungen in den Ergebnissen; stark pauschalisierte, womöglich verzerrte Preiselastizitäten aus US-Studien, die auf Deutschland wegen länderspezifischer Unterschiede nur bedingt übertragbar sind, weshalb die Validität der Studienergebnisse fragwürdig ist.

Modellspezifikation

- **Modellierungsverfahren:** Kausalität wird bei Monte-Carlo-Simulationen ex ante bei der Modellkonstruktion unterstellt, nicht aber durch das Modell selbst nachgewiesen; Auswahl und Verteilung der Parameter nicht offengelegt, daher eingeschränkte Nachvollziehbarkeit der Modellierung; geringe Zahl von 100 Simulationsdurchläufen pro Gruppe gewährleistet keine robusten Schätzungen.
- **Einfluss- und Zielgrößen:** Die vereinfachende Annahme einer linearen Beziehung zwischen Kalorienaufnahme und Körpergewicht widerspricht der multifaktoriellen Entstehung von Übergewicht und Adipositas; lediglich Untersuchung kurzfristiger Effekte ohne Berücksichtigung langfristiger Verhaltensanpassungen, sodass die Aussagekraft der Ergebnisse eingeschränkt ist; Nichtberücksichtigung der Heterogenität von Preiselastizitäten führt zu erheblicher Pauschalisierung der Ergebnisse.
- **Weitere Einflussfaktoren:** Lediglich teilweise Berücksichtigung möglicher Substitution (Milch und Fruchtsäfte) und Nichtberücksichtigung anderer Faktoren (z. B. regionale Unterschiede), sodass die simulierten Effekte die realen Zusammenhänge möglicherweise nicht korrekt beschreiben.

Ergebniskommunikation

- **Ergebnisdarstellung:** Fehlende Quantifizierung der Variabilität auf Populationsebene sowie generelles Fehlen von Konfidenzintervallen erschwert die Abschätzung der zufälligen Fehler; die Darstellung überwiegend in absoluten Zahlen erschwert die Vergleichbarkeit zwischen Subgruppen,
- **Modellbewertung:** Lediglich vereinzelte Sensitivitätsanalysen und fehlende Überprüfung der Auswirkungen zentraler Modellannahmen erschweren die Einschätzung modellbedingter Unsicherheit.
- **Ergebniseinordnung:** Die Schlussfolgerung (Steuer reduziert Übergewicht und Adipositas) behauptet einen kausalen Zusammenhang, der nicht vom Modell selbst belegt wird; eine tiefgreifende Diskussion der Auswirkungen der Limitationen auf Aussagekraft der Ergebnisse bleibt aus.

Gesamtbeurteilung der Studie von Schwendicke und Stolpe (2017)

Die Studie von Schwendicke und Stolpe (2017) weist erhebliche methodische Schwächen auf, insbesondere eine unzureichende Datengrundlage, stark vereinfachende Modellannahmen und die Nichtberücksichtigung relevanter Einflussfaktoren. Diese schränken die Aussagekraft und Zuverlässigkeit der Ergebnisse stark ein. Veraltete und nur unzureichend passende Eingangsdaten, darunter nicht kompatible Altersklassifikationen und Preiselastizitäten mit begrenzter Übertragbarkeit auf die Zielgruppe, beeinträchtigen die Validität der Modellierung. Zudem beruhen die Modellannahmen auf stark vereinfachten und realitätsfernen Prämissen, wie der linearen Beziehung zwischen Kalorienaufnahme und Gewicht, der unzureichenden Modellierung von Substitutionseffekten und der Ignoranz langfristiger Verhaltensanpassungen. Diese Annahmen vernachlässigen Wechselwirkungen und Dynamiken realen Verhaltens und stehen im Widerspruch zur Komplexität der Faktoren, die Übergewicht und Adipositas bedingen. Die Verlässlichkeit der Ergebnisse ist daher fragwürdig. Außerdem ist von der Interpretation der Wirksamkeit der Steuer – wie von den Studienautoren vorgenommen – abzusehen, da das verwendete Modell nicht in der Lage ist, Kausalzusammenhänge nachzuweisen. Insgesamt verdeutlicht die Evaluation, dass die Studie nicht den Anforderungen genügt, um als Grundlage für politische Entscheidungen zu dienen.

4.3 Studie von Rogers, Cummins et al. (2023), England

Vollständige Quellenangabe zur Studie:

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), Artikel e1004160. https://doi.org/10.1371/journal.pmed.1004160

Executive Summary

Die Studie von Rogers, Cummins et al. (2023) simuliert Zusammenhänge zwischen den Ereignissen rund um die Einführung einer Steuer auf zuckergesüßte Getränke im Vereinigten Königreich und der Entwicklung der Adipositasprävalenz bei englischen Vorschulkindern sowie Sechstklässlern mithilfe einer unterbrochenen Zeitreihenanalyse. Die Modellierungsergebnisse zeigen lediglich bei Mädchen in der 6. Klasse eine niedrigere Adipositasprävalenz gegenüber einer hypothetischen Situation ohne Steuer. Die Studienautoren schließen daraus einerseits, dass die Steuer zur Reduktion der Adipositasprävalenz älterer Grundschulkinder beitragen kann, andererseits, dass die Steuer alleine nicht ausreicht, um Adipositas bei Kindern umfassend zu reduzieren. Die Verlässlichkeit der Modellierungsergebnisse ist allerdings erheblich eingeschränkt, insbesondere aufgrund der nicht-repräsentativen Stichprobe sowie der Nichtberücksichtigung weiterer relevanter Faktoren, die eine eindeutige Zuschreibung der Effekte auf die Steuer nicht zulässt. Dies beeinträchtigt die Robustheit der ohnehin stark generalisierten Schlussfolgerungen. Insgesamt erfüllt die Studie aus statistisch-methodischer Perspektive nicht die erforderlichen Qualitätsstandards, um als belastbare Grundlage für evidenzbasierte Entscheidungen zu dienen.

4.3.1 Zusammenfassung der Studie

Die Inhalte der Studie von Rogers, Cummins et al. (2023) werden nachfolgend zusammengefasst.

Untersuchungsgegenstand

Die Studie modelliert Zusammenhänge zwischen den Ereignissen rund um die Ankündigung der britischen *Soft Drinks Industry Levy* (kurz: *SDIL*)[42] und der Entwicklung der Adipositasprävalenz bei englischen Kindern, mit besonderem Fokus auf mögliche Geschlechtsunterschiede sowie regional-sozioökonomische Disparitäten.

Methodik

Betrachtet werden Kinder, die zwischen September 2013 und November 2019 die Vorschule (Alter: vier bis fünf Jahre) oder die 6. Klasse (Alter: zehn bis elf Jahre) an staatlichen Schulen in England besuchten. Die Analyse stützt sich auf Daten zu Größe und Gewicht der Kinder, aus denen die Prävalenz von Übergewicht und Adipositas bestimmt wird. Der sozioökonomische Status wird basierend auf dem Standort der Schule ermittelt. In der Hauptanalyse wird der Effekt der Steuerankündigung (März 2016) mittels einer *unterbrochenen*

[42] Hersteller unterliegen im Rahmen einer gestaffelten Verbrauchsteuer unterschiedlich hohen Abgaben auf abgepackte, zuckergesüßte Getränke, abhängig vom Zuckergehalt: Für Getränke mit weniger als 5 g Zucker pro 100 ml Flüssigkeit fällt keine Steuer an, bei einem Zuckergehalt zwischen 5 und 8 g je 100 ml beträgt die Abgabe 0.18£ pro Liter und ab einem Zuckergehalt von 8 g je 100 ml beträgt sie 0.24£ (Sasse & Metcalfe, 2022).

Zeitreihenanalyse untersucht, indem die realen Datenwerte mit einem simulierten Trend ohne Steuerankündigung (= kontrafaktisches Szenario) verglichen werden.[43] Die Analyse wird sowohl insgesamt als auch getrennt für verschiedene sozioökonomische Gruppen durchgeführt. Die Zeitreihen werden mit einem *ARIMA*[44]-Modell analysiert. Ergänzend zur Hauptanalyse werden verschiedene Sensitivitätsanalysen (siehe Anhang A.3.2) durchgeführt, die zum einen verschiedene Definitionen der Intervention (also Unterbrechungen der Zeitreihe) betrachten, nämlich (1) die Rezepturänderung zuckergesüßter Getränke ab November 2016[45] sowie (2) die tatsächliche Einführung der Steuer im April 2018, und die zum anderen (3) die Modellierung von Adipositasprävalenz auf die Prävalenz von Übergewicht und Adipositas erweitern.

Ergebnisse

Den Modellierungsergebnissen zufolge liegt die Adipositasprävalenz bei Mädchen in der 6. Klasse am Ende des Untersuchungszeitraums um 1,6 Prozentpunkte, 95 % CI [1,1; 2,1] unter dem simulierten Trend ohne Steuer, was einer relativen Reduktion der Adipositasprävalenz[46] von 8,0 %, 95 % CI [5,4; 10,5], entspricht. In sozioökonomisch benachteiligten Gebieten zeigt das Modell einen stärkeren Effekt (2,4 Prozentpunkte im am meisten benachteiligten Gebiet, 95 % CI [1,6; 3,2]). In manchen Subgruppen, wie etwa bei Jungen in der 6. Klasse mit Schulstandort in sozioökonomisch privilegierten Regionen, ergibt die Modellierung eine um 1,6 Prozentpunkte, 95 % CI [0,7; 2,5], bzw. 10,1 %, 95 % CI [4,3; 15,9], höhere Adipositasprävalenz. Bei Vorschulkindern werden keine Effekte gefunden. Ähnliche Ergebnisse zeigen sich bei der Definition der Intervention als Zeitpunkt der Rezepturänderung zuckergesüßter Getränke (Sensitivitätsanalyse 1) oder bei der Betrachtung der kombinierten Prävalenz von Übergewicht und Adipositas (Sensitivitätsanalyse 3). Wird als die Intervention der Zeitpunkt der Steuereinführung definiert (Sensitivitätsanalyse 2), zeigen sich hingegen keine signifikanten Unterschiede in der Adipositasprävalenz.

Schlussfolgerungen der Studienautoren

Aus ihren Modellierungsergebnissen folgern die Studienautoren, dass die SDIL dazu beitragen kann, Adipositas bei älteren Grundschulkindern und Gesundheitsungerechtigkeit bei Kindern allgemein zu verringern. Darüber hinaus plädieren sie für zusätzliche Maßnahmen zur Senkung der Adipositasprävalenz bei Kindern, insbesondere bei Jungen und jüngeren Altersgruppen.

4.3.2 Evaluation der Studie

Die Studie von Rogers, Cummins et al. (2023) wird nachfolgend anhand der in Kapitel 3 vorgestellten Beurteilungskriterien evaluiert.

Datengrundlage

Die Studie basiert auf Querschnittdaten zur Größe und zum Gewicht von Vorschulkindern und Schülern der 6. Klasse, die jährlich (nur für diese Altersgruppen) im Rahmen des *National Child Measurement Programme*

[43] Der Zeitpunkt der Steuerankündigung markiert den Beginn der Intervention, ab dem die realen Daten mit einem kontrafaktischen Szenario verglichen werden. Dieser Zeitpunkt kann daher als „Unterbrechung" der Zeitreihe betrachtet werden.
[44] ARIMA ist das Akronym für „autoregressive integrated moving average".
[45] Bereits nach Ankündigung der SDIL passten viele Hersteller ihre Rezepturen an, um unter den Zuckergehaltsschwellen zu bleiben, die eine (höhere) Abgabe erfordern (Dickson et al., 2023). Die Zahl an Getränken mit veränderter Rezeptur stieg ab Ende 2016 (Sasse & Metcalfe, 2022).
[46] Im Detail handelt es sich nicht um einen Rückgang der Adipositasprävalenz nach Ankündigung der SDIL, sondern um eine Abschwächung des Anstiegs der Adipositasprävalenz im Rahmen des insgesamt zunehmenden Trends von Adipositas bei Kindern.

(kurz: *NCMP*; Office for Health Improvement and Disparities, 2024) erfasst werden. Die Messungen sind generell objektiv und umfassen nahezu alle staatlichen Schulen in England. Die Teilnahmequoten sind hoch (etwa 90 %), da Eltern aktiv widersprechen müssen, um ihre Kinder von den Messungen auszunehmen. Ein Hinweis auf eine nicht repräsentative Stichprobe ergibt sich aus der Beobachtung, dass insbesondere übergewichtige und adipöse Mädchen tendenziell häufiger aus dem Programm genommen werden. Laut dem NCMP-Leitfaden könnte diese Selektion zu Verzerrungen in Bezug auf Mädchen führen (Hancock & Copley, 2016). Da die Ergebnisse von Rogers, Cummins et al. (2023) nur für Mädchen der 6. Klasse eine signifikante Reduktion der Adipositasprävalenz im Vergleich zum kontrafaktischen Szenario zeigt, ist sorgfältig zu prüfen, ob eine Unterrepräsentation übergewichtiger und adipöser Mädchen in der Stichprobe für diesen Effekt verantwortlich ist. Da außerdem im Rahmen des NCMP jedes Jahr andere Kinder gemessen werden, bieten die Daten keine Informationen über individuelle Entwicklungen im Zeitverlauf, sondern stellen Querschnittverteilungen der Körperindizes zum jeweiligen Messzeitpunkt dar. Hinzu kommt, dass die Daten der (Vor-)Schulkinder zu verschiedenen Zeitpunkten innerhalb des Schuljahres erhoben wurden. Die publizierten Abbildungen zeigen die Datenpunkte der analysierten Zeitreihe in monatlichen Abständen, doch es fehlen genauere Informationen zur monatlichen Zusammensetzung der Stichprobe, auf deren Grundlage die Adipositas- und Übergewichtsprävalenz berechnet wurde. Daher können systematische Unterschiede im Zeitverlauf nicht ausgeschlossen werden, was die Zufälligkeit und uneingeschränkte Vergleichbarkeit der Datenpunkte infrage stellt.

Die Klassifikation von Übergewicht und Adipositas auf der Grundlage der NCMP-Daten erfolgt gemäß den Standards des National Health Services (National Health Service, 2023). Übergewicht und Adipositas werden nicht anhand eines festen BMI-Wertes kategorisiert, sondern altersstufenabhängig unter Berücksichtigung der Verteilung gemäß Cole et al. (1995) definiert, was eine differenzierte Betrachtung ermöglicht, die der natürlichen Variabilität des BMI in unterschiedlichen Altersgruppen Rechnung trägt. Der Bezug auf immer gleiche Referenzwerte unterstützt Vergleichbarkeit. Allerdings ist zu hinterfragen, ob die zugrunde liegenden Referenzdaten aus dem Jahr 1990 noch zeitgemäß sind, da Veränderungen in der Bevölkerungsstruktur und Lebensweise seither die BMI-Verteilung beeinflusst haben könnten.

Die Schätzung des sozioökonomischen Status anhand des Schulstandorts berücksichtigt keine individuellen sozialen Faktoren und ist dementsprechend lediglich als grobe, mit hoher Unsicherheit behaftete Klassifikation anzusehen.

Modellspezifikation

Unterbrochene Zeitreihenanalysen (siehe Anhang A.2.1) eignen sich grundsätzlich, um zeitliche Veränderungen (Einflussgröße: Zeit) einer Variablen (Zielgröße hier: Entwicklung der Adipositasprävalenz) vor und nach einer Intervention zu identifizieren. Die Wirkrichtung der Einfluss- auf die Zielgrößen wird bei Zeitreihenanalysen ex ante angenommen und nicht durch das Modell nachgewiesen. Da Zielgrößen in der Realität häufig von vielen Faktoren gleichzeitig beeinflusst werden, ist eine Bestimmung des isolierten Effekts durch die Intervention allerdings schwierig. So kann die Prävalenz von Adipositas bspw. nicht nur durch die Einführung der SDIL, sondern auch durch andere Faktoren wie Aufklärungskampagnen beeinflusst werden. Unterbrochene Zeitreihenanalysen können solche Einflüsse und deren Wechselwirkungen nur schwer kontrollieren, was die Isolierung des Effekts der SDIL erschwert. Derartige Einflussfaktoren werden im Modell von Rogers, Cummins et al. (2023) nicht

berücksichtigt. Das Modell berücksichtigt lediglich Alter, Geschlecht und die Monate rund um die Sommerferien[47] als Prädiktoren im Modell.

Außerdem ist bei Zeitreihenanalysen die Berechnung des kontrafaktischen Szenarios (hier: simulierter Trend ohne Steuer) sehr empfindlich gegenüber dem gewählten Zeitpunkt, der zur Modellierung des Trends herangezogen wird. Um den möglichen Trend ohne Steuer zu simulieren, werden die historisch beobachteten Daten bis zum Interventionszeitpunkt verwendet. Wird der (unbekannte) Interventionszeitpunkt aber falsch modelliert, kann dies in verzerrten Schätzungen der Effekte resultieren. Zur Abschätzung eines solchen möglichen Fehlspezifikationseffekts führen die Studienautoren Sensitivitätsanalysen mit zwei weiteren Interventionszeitpunkten durch, was eine verbesserte Einschätzung der tatsächlichen Situation ermöglicht.

Jandoc et al. (2015) empfehlen einheitliche Methoden- und Berichtsstandards für unterbrochene Zeitreihenanalysen, denen die Studie von Rogers, Cummins et al. (2023) weitgehend folgt. Teilweise sind jedoch Schwächen und Lücken hinsichtlich der methodischen Spezifikation erkennbar. (1) So wird zum einen vereinfachend ein linear fortgesetzter Trend der Adipositasprävalenz seit Ankündigung der SDIL für das kontrafaktisches Szenario angenommen. Es ist aber davon auszugehen, dass die Entwicklung der Adipositasprävalenz in der Realität kaum streng linear erfolgt, sodass daraus abgeleitete Ergebnisse wohl mit Unsicherheit behaftet sind. (2) Zum anderen fehlen wesentliche Angaben[48] zum verwendeten ARIMA-Modell und zur Überprüfung der erforderlichen Voraussetzungen. Die fehlenden Angaben implizieren, dass Rogers, Cummins et al. (2023) die Modellspezifikation entweder unvollständig dokumentieren oder dass sie ein lineares Regressionsmodell fälschlicherweise als ARIMA-Modell, welches grundsätzlich das geeignetste Verfahren zur Modellierung mit Saisonalitäten, Autokorrelation und Rauschen darstellt, bezeichnen. Dies schränkt die Nachvollziehbarkeit des zugrunde liegenden Modells ein. Ohne umfassende Prüfung der Voraussetzungen ist die Beurteilung der Validität der Ergebnisse eingeschränkt. Die Schätzung der Modellparameter im Modell erfolgt aber generell nach gängigen, auf die Datenbasis abgestimmten Methoden. (4) Ebenso machen Rogers, Cummins et al. (2023) keine expliziten Angaben dazu, ob die Anzahl der zugrunde liegenden Datenpunkte ausreichend ist, auch wenn Jandoc et al. (2015) die Bedeutung einer ausreichenden Anzahl von Datenpunkten für die Robustheit von Zeitreihenanalysen betonen. Die in den Abbildungen der Studie gezeigten Datenpunkte deuten aber darauf hin, dass die Datenbasis für die Durchführung der Analysen grundsätzlich als ausreichend angesehen werden kann.

Ergebniskommunikation

Alle Ergebnisse werden transparent zusammen mit Konfidenzintervallen angegeben. Jedoch sei darauf verwiesen, dass Konfidenzintervalle lediglich den zufälligen Stichprobenfehler widerspiegeln, jedoch nicht den nichtzufälligen Fehler, der durch die nicht-repräsentative Stichprobe (Unterrepräsentation übergewichtiger und adipöser Mädchen) entsteht. Die Aussagekraft der angegebenen Konfidenzintervalle ist daher eingeschränkt. Zudem werden die simulierten Differenzen in der Adipositasprävalenz zwischen den beobachteten und den simulierten Verläufen lediglich als absolute bzw. relative Unterschiede in Prozent bzw. Prozentpunkten dargestellt. Eine Kontextualisierung (z. B. Umrechnung in absolute Zahlen zur Angabe der Anzahl adipöser Kinder in den

[47] Für Vorschulkinder wurden die Monate September, Oktober, Juni und Februar und für Schulkinder in der 6. Klasse die Monate September und Juli berücksichtigt. Diese wurden in einer Voranalyse als relevant für die Adipositasprävalenz identifiziert.

[48] Es fehlen Werte zu autoregressive Termen, Differenzterme zur Entfernung von Trends oder Nicht-Stationarität sowie Fehlerterme als Moving-Average-Komponenten, sowie vollständige Informationen zu Verschiebungsparametern und deren Untersuchung sowie zur Behandlung der Nicht-Stationarität.

Szenarien mit und ohne Steuer; auch als Angabe in natürlichen Häufigkeiten denkbar) erfolgt allerdings nicht. Dies erschwert die Einordnung der Ergebnisse und reduziert die Transparenz der Analyse. Es fehlen Angaben zu den berechneten Trendparametern, also den Veränderungsraten des realen und des kontrafaktischen Verlaufs, was die Plausibilisierung der Ergebnisse erschwert. Weiterhin fehlen – entgegen der Empfehlung von Greenland et al. (2016) – Angaben zu p-Werten und Effektstärken, die eine fundierte Einordnung der statistischen Relevanz der Unterschiede zwischen realem und simuliertem Trend ermöglichen würden.

Die Studienautoren thematisieren mehrere Limitationen ihrer Studie, darunter die selektive Teilnahme adipöser Mädchen, ohne plausibel zu begründen, warum dies (wie von ihnen angenommen) zu einer Unterschätzung der Effekte führen sollte. Außerdem werden die grobe Schätzung des sozioökonomischen Status, der Mangel an Daten nach der Einführung der SDIL zur Einschätzung von Langzeiteffekten, sowie die unzureichende Untersuchung von saisonalen Schwankungen als Limitationen aufgeführt. Ebenfalls wird die Empfindlichkeit des simulierten Trends ohne Steuer thematisiert: Je nach gewähltem Interventionszeitpunkt kann die geschätzte Wirkung der Steuer deutlich verzerrt sein. Um dieses Risiko zu reduzieren, wurden in der Studie Sensitivitätsanalysen mit unterschiedlichen Interventionszeitpunkten durchgeführt. Die Studienautoren berichten weiterhin korrekterweise, dass es sich bei ihrem Modellergebnis lediglich um einen (korrelativen) Zusammenhang zwischen der Ankündigung der SDIL und einer Reduktion von Adipositasfällen bei 10- bis 11-jährigen Mädchen handelt.

Die Studienautoren folgern aus ihren Modellierungsergebnissen, dass die SDIL zur Verringerung von Adipositas bei älteren Grundschulkindern beitragen kann, jedoch wird diese Annahme nicht hinreichend durch das Modell gestützt. Der Effekt basiert ausschließlich auf den Ergebnissen für Mädchen der 6. Klasse, deren Daten vermutlich verzerrt sind, da übergewichtige und adipöse Mädchen in der Stichprobe unterrepräsentiert sind. Einen derart allgemeinen Effekt abzuleiten, ist außerdem stark generalisiert, zumal bei Jungen kein vergleichbarer Effekt nachgewiesen wurde. Die Studienautoren folgern daher auch, dass die Steuer allein nicht ausreicht, um Adipositas bei Kindern – insbesondere bei Jungen und jüngeren Altersgruppen – wirksam zu reduzieren, und dass zusätzliche Maßnahmen erforderlich sind. Da lediglich Kinder in der Vorschule und der 6. Klasse untersucht wurden, könnten Analysen aller Altersstufen detailliertere Erkenntnisse dazu liefern. Ein Appell für zusätzliche Maßnahmen lässt sich aus den Ergebnissen daher nur bedingt ableiten.

Darüber hinaus folgern die Studienautoren, dass die SDIL bei Kindern zur Verringerung von Gesundheitsungerechtigkeiten in Bezug auf Adipositas beitragen kann[49]. Diese Folgerung ist jedoch verkürzt, da sie auf einer groben Zuordnung des sozioökonomischen Status ohne Berücksichtigung individueller Unterschiede basiert. Zudem stützt sie sich zu großen Teilen auf mutmaßlich verzerrte Effekte in der Stichprobe der Mädchen. Die Modellierung zeigt je nach Geschlecht und sozioökonomischem Status sowohl signifikante als auch nicht-signifikante Effekte in verschiedene Richtungen (siehe Tabelle 2 in der Studie von Rogers, Cummins et al., 2023). Der postulierte mögliche Effekt zur Verbesserung der Gesundheitsgerechtigkeit bleibt daher äußerst hypothetisch und erfordert zweifellos spezifischere Analysen, um belastbare Aussagen zu ermöglichen –insbesondere im Hinblick auf die Gestaltung politischer Maßnahmen.

[49] Nach Angaben von Rogers, Cummins et al. (2023) ist die Prävalenz von Adipositas in sozioökonomisch benachteiligten Gebieten höher als in privilegierten. Eine Reduktion der Gesundheitsungerechtigkeit bedeutet in diesem Zusammenhang, dass die Adipositasprävalenz in benachteiligten Gebieten nach der Steuer weniger stark ansteigt als in privilegierten, wodurch insgesamt eine Angleichung zwischen den verschiedenen sozioökonomischen Gruppen erzielt wird.

Zusammenfassung der Limitationen der Studie von Rogers, Cummins et al. (2023)

Datengrundlage

- **Stichprobe:** Strukturell von der interessierenden Population abweichende Stichprobe der Subgruppe der Mädchen (Unterrepräsentation übergewichtiger und adipöser Mädchen), sodass Ergebnisse verzerrt sind und gerade der simulierte Effekt für Sechstklässlerinnen in Frage gestellt werden muss.
- **Weitere Eingangsdaten:** Körperindizes als Reihe von Querschnittsdaten potenziell nicht zufällig, sodass systematische Unterschiede im Zeitverlauf nicht ausgeschlossen werden können; lediglich grobe (regionale) Klassifikation des sozioökonomischen Status, sodass diesbezügliche Schlussfolgerungen mit Unsicherheit behaftet sind.

Modellspezifikation

- **Modellierungsverfahren:** Kausalität wird bei Zeitreihenanalysen ex ante bei der Modellkonstruktion unterstellt, nicht aber durch das Modell selbst nachgewiesen; generelle Empfindlichkeit von Zeitreihenanalysen bei Berechnung des simulierten Trends (Szenario ohne Steuer) gegenüber Interventionszeitpunkt mit Risiko für verzerrte Effektschätzungen; lediglich unzureichende Angaben zur Modellspezifikation (z. B. ARIMA versus lineare Regression; fehlende Prüfung der erforderlichen Voraussetzungen), somit eingeschränkte Nachvollziehbarkeit.
- **Einfluss- und Zielgrößen:** Vereinfachende Annahme eines linear fortgesetzten Trends der Adipositasprävalenz seit Ankündigung der SDIL für kontrafaktisches Szenario, die der realen Entwicklung möglicherweise nicht gerecht wird und die Modellanpassung mutmaßlich verschlechtert.
- **Weitere Einflussfaktoren:** Nichtberücksichtigung relevanter Faktoren (z. B. Aufklärungskampagnen), sodass Effekte der SDIL nicht zwangsläufig in der dargestellten Form existieren müssen.

Ergebniskommunikation

- **Ergebnisdarstellung:** Trotz transparenter Angabe lediglich begrenzte Aussagekraft der Konfidenzintervalle, die Unsicherheit durch nicht-repräsentative Stichprobe nicht abbilden können; das Fehlen von p-Werten und Effektstärken verhindert die fundierte Einordnung der statistischen Relevanz; fehlende Kontextualisierung erschwert Ergebniseinordnung; fehlende methodische Angaben zu Trendparametern erschweren Plausibilisierung der Ergebnisse.
- **Modellbewertung:** Trotz durchgeführter Sensitivitätsanalysen zu den Interventionszeitpunkten ist von Unsicherheit in den Ergebnissen aufgrund hoher Empfindlichkeit der simulierten Trends hinsichtlich dieser Zeitpunkte auszugehen.
- **Ergebniseinordnung:** Die Schlussfolgerung (SDIL kann zur Reduktion von Adipositas bzw. Gesundheitsungerechtigkeiten bei älteren Schulkindern beitragen) ist nicht hinreichend durch das Modell gestützt (Effekt ausschließlich bzw. zu großen Teilen auf simuliertes Ergebnis in der Subgruppe der Sechstklässlerinnen zurückzuführen, der mutmaßlich verzerrt ist); Appell für zusätzliche Maßnahmen zur Reduktion der Adipositasprävalenz nur bedingt aus Ergebnissen ableitbar.

Gesamtbeurteilung der Studie von Rogers, Cummins et al. (2023)

Die Verlässlichkeit der Modellierungsergebnisse von Rogers, Cummins et al. (2023) ist insbesondere aufgrund der nicht-repräsentativen Stichprobe und der Nichtberücksichtigung weiterer relevanter Faktoren erheblich eingeschränkt. Die geringere Teilnahme übergewichtiger und adipöser Mädchen bei den Messungen führt zu einer potenziellen Verzerrung der Modellierungsergebnisse, die von den Studienautoren zwar genannt, aber nicht ausreichend diskutiert wird. Aufgrund der Nichtberücksichtigung relevanter Faktoren (z. B. parallele Aufklärungskampagnen), können die simulierten Effekte außerdem nicht eindeutig der Steuer zugeschrieben werden. Diese Schwächen untergraben die Robustheit der Schlussfolgerungen, weshalb die Studie nicht als belastbare Grundlage für politische Entscheidungen herangezogen werden sollte.

4.4 Studie von Cobiac et al. (2024), England

Vollständige Quellenangabe zur Studie:

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), Artikel e1004371. https://doi.org/10.1371/journal.pmed.1004371

Executive Summary

Die Studie von Cobiac et al. (2024) modelliert den Einfluss einer im April 2018 im Vereinigten Königreich in Kraft getretenen Steuer auf zuckergesüßte Getränke auf die Zuckeraufnahme durch Getränke in englischen Haushalten sowie damit verbundene gesundheitliche Auswirkungen bei Kindern bis 17 Jahren, unter Berücksichtigung sozioökonomischer Benachteiligung. Grundlage der Modellierung bildet eine unterbrochene Zeitreihenanalyse, mit der Veränderungen im Zuckerkonsum erfasst werden. Potenzielle Effekte der Konsumänderung auf Übergewicht und gesundheitsbezogene Aspekte werden mit zwei Kohortensimulationen modelliert. Die Modellierung ergibt eine Reduktion des Zuckerkonsums in englischen Haushalten, insbesondere in sozioökonomisch benachteiligten Regionen. Darüber hinaus prognostiziert das Modell bei Kindern eine Reduktion der Prävalenz von Übergewicht, eine Verbesserung der Zahngesundheit und Lebensqualität sowie eine vernachlässigbare Veränderung der Lebenserwartung. Die Ergebnisse sind kritisch zu bewerten, da die Studie fundamentale Schwächen in der Datengrundlage aufweist: Die verwendeten Verkaufsdaten spiegeln den tatsächlichen Konsum der untersuchten Stichprobe nicht angemessen wider. Zudem beeinträchtigen vereinfachende Modellannahmen und die Vernachlässigung relevanter Einflussfaktoren die Belastbarkeit der Ergebnisse zusätzlich. Insgesamt erfüllt die Studie aus statistisch-methodischer Perspektive nicht die erforderlichen Qualitätsstandards, um als belastbare Grundlage für evidenzbasierte Entscheidungen zu dienen.

4.4.1 Zusammenfassung der Studie

Die Inhalte der Studie von Cobiac et al. (2024) werden nachfolgend zusammengefasst.

Untersuchungsgegenstand

Die Studie modelliert potenzielle Effekte der Besteuerung zuckergesüßter Getränke im Rahmen der im März 2016 angekündigten und im April 2018 eingeführten britischen SDIL auf die Zuckeraufnahme durch Getränke in englischen Haushalten sowie die daraus resultierenden mittel- und langfristigen gesundheitlichen Auswirkungen für Kinder, wobei jeweils der Grad der regionalen sozioökonomischen Benachteiligung berücksichtigt wird.

Methodik

Basierend auf Daten zu wöchentlichen Haushaltseinkäufen wird mithilfe einer unterbrochenen Zeitreihenanalyse simuliert, wie sich die durch den Kauf verschiedener – nicht ausschließlich zuckergesüßter – Getränke erworbene Zuckermenge im Zeitraum von März 2014 bis November 2019 im Vergleich zu einem kontrafaktischen Szenario

ohne Steuer verändert hat. Auf dieser Grundlage werden – unter der Annahme eines stabilen Einkaufstrends über einen Zeitraum von zehn Jahren – mittelfristige Effekte auf den BMI, Zahnkaries und die Lebensqualität (QALYs)[50] sowie langfristige Effekte auf die Lebenserwartung von englischen Kindern zwischen null und 17 Jahren mittels zweier verschiedener Kohortensimulationen modelliert. Die Analysen werden sowohl insgesamt als auch getrennt für verschiedene sozioökonomische Gruppen durchgeführt.

Ergebnisse

Die Simulationsergebnisse zeigen, dass sich nach Einführung der SDIL die durch Getränkekäufe erworbene Menge an Zucker um 15 Gramm pro Haushalt und Woche reduziert, 95 % CI [10,3; 19,7]. Den Modellprognosen zufolge führe diese Reduktion bis 2025 bei Kindern zu jährlich 3.600 weniger Kariesfällen, 95 % CI [946; 6.330], sowie 64.100 weniger Fällen von Übergewicht und Adipositas, 95 % CI [54.400; 73.400], was insgesamt einen Gewinn von 19.500 QALYs, 95 % CI [14.800; 24.600], zur Folge hätte. Die prognostizierte Veränderung der Lebenserwartung ist marginal[51] und daher vernachlässigbar.

Schlussfolgerungen der Studienautoren

Die Studienautoren schlussfolgern aus dem Rückgang der durch Getränkekäufe erworbenen Zuckermenge, dass die SDIL zu einer Reduktion des Zuckergehalts von Getränken geführt habe, was sich positiv auf gesundheitliche Aspekte bei englischen Kindern auswirken könne.

4.4.2 Evaluation der Studie

Die Studie von Cobiac et al. (2024) wird nachfolgend anhand der in Kapitel 3 vorgestellten Beurteilungskriterien evaluiert.

Datengrundlage

Für die Prognose der *mittelfristigen* gesundheitlichen und gesundheitsbezogenen Auswirkungen der SDIL auf Kinder werden Bevölkerungsdaten der gesamten Kinder- und Jugendpopulation (null bis 17 Jahre) in England aus dem Jahr 2015 sowie Bevölkerungsprojektionen zur Modellierung der Geburtsjahrgänge nach 2015 verwendet. Für die Prognose der *langfristigen* gesundheitsbezogenen Auswirkungen werden Bevölkerungsdaten des Geburtsjahrgangs 2015 in England verwendet. All diese Daten stammen vom Office for National Statistics. Da es sich um aktuelle amtliche Bevölkerungsdaten handelt, welche zudem die gesamte interessierende Population abbilden, ist eine grundlegende strukturelle Übereinstimmung mit der Zielpopulation anzunehmen.

Die Daten zu den Getränkekäufen[52] werden aus wöchentlichen Querschnittdaten zu Haushaltseinkäufen abgeleitet, welche im Mittel etwa 17.000 englische Haushalte umfassen. Datenlieferanten sind Haushalte, die ihre Daten zu wöchentlichen Einkäufen einem Marktforschungsunternehmen gegen ein geringes jährliches Entgelt zur Verfügung stellen. Die Daten wurden basierend auf Selbstauskünften erfasst. Aufgrund eines potenziell sozial erwünschten Antwortverhalten oder Angaben aus der Erinnerung sind diese generell anfällig für Verzerrungen. Um die Repräsentativität sicherzustellen, nutzen Cobiac et al. (2024) Gewichtungsfaktoren des Marktforschungsunternehmens,

[50] Quality Adjusted Life Years (QALYs) errechnen sich aus er prognostizierten verbleibenden Lebenszeit, multipliziert mit einem Nutzwertfaktor, der zwischen null für die denkbar schlechteste und eins für die bestmögliche Lebensqualität liegt.
[51] Sie bewegt sich je nach Grad der sozioökonomischen Benachteiligung im Spektrum von wenigen Tagen bis zu einem Monat.
[52] Es handelt sich explizit um Kaufdaten und nicht um Konsumdaten. Die damit einhergehenden Studienlimitationen werden im nachfolgenden Abschnitt *Modellspezifikation* erläutert.

erläutern das Gewichtungsverfahren aber nicht näher. Um aus den Daten auf Haushaltsebene im weiteren Verlauf die Pro-Kopf-Veränderung des Zuckerkaufs durch Getränke zu berechnen, wird eine weitere Quelle zur durchschnittlichen Haushaltsgröße (2,4 Personen) herangezogen. Unklar bleibt, ob die Übereinstimmung der durchschnittlichen Haushaltsgröße in der Stichprobe mit derjenigen in der Population überprüft wurde. Zudem ist es problematisch, bei einer Untersuchung von Kindern Daten einer Durchschnittsperson eines Haushalts zu verwenden, da diese das spezifische Kaufverhalten der Zielgruppe nicht angemessen widerspiegelt[53]. Außerdem wurden die Getränkekäufe ausschließlich auf Basis von Einzelhandelsdaten ermittelt, ohne gastronomische Daten zu berücksichtigen, was zu einer Verzerrung der Ergebnisse führt. Cobiac et al. (2024) nehmen unter Verweis auf die Studie von Cornelsen et al. (2019) an, dass der Anteil der Ausgaben für Käufe außerhalb des eigenen Haushalts etwa 10-12 % betrage. Andere Quellen schätzen diesen Anteil deutlich höher, etwa auf 45 % (Statista, 2022). Diese Anteile können je nach Region und sozioökonomischem Status variieren (Law et al., 2022), was Cobiac et al. (2024) auch entsprechend anmerken.

Die Zuordnung des Grades an sozioökonomischer Benachteiligung zu den Haushalten erfolgt anhand der Postleitzahlen der Haushalte auf Basis des *Index of Multiple Deprivation* (Smith et al., 2015), was grobe regionale Schätzungen ermöglicht, aber individuelle Faktoren, die nicht unmittelbar mit der geografischen Lage zusammenhängen, ignoriert. Darauf basierend werden die Haushalte in der Modellierung in fünf gleich große Gruppen kategorisiert, die nach dem Grad der Benachteiligung gestaffelt sind. Eine adäquate Abbildung und Abgrenzung sozialer Disparitäten wird dadurch allerdings nicht erreicht.

Die Modellparameter zur Gesundheit stammen aus verschiedenen Datenquellen und beruhen auf unterschiedlichen Erhebungsmethoden. Zum einen handelt es sich um Querschnittdaten aus staatlich veranlassten Befragungen zu Körperindizes und Zahngesundheit, die in Form von Selbstauskünften erhoben wurden, partiell ergänzt durch objektive Messungen (National Health Service, 2011, 2015, 2016). Der Großteil der Gesundheitsparameter stammt analog zu den Konsumdaten aus Daten zu England, lediglich die Daten zur Zahngesundheit beinhalten ebenfalls Datenpunkte aus Wales und Nordirland, wodurch keine vollständige geografische Deckungsgleichheit besteht. Wesentliche Abweichungen zu rein englischen Daten sind jedoch nicht zu erwarten, da Wales und Nordirland ebenfalls Teil des Vereinigten Königreichs sind und der SDIL gleichermaßen unterliegen. Daten zur Lebensqualität und -erwartung stammen aus Metastudien, deren Aussagekraft stark von der Methodik der einbezogenen Studien abhängt. Dabei bleibt sowohl die Anwendbarkeit im spezifischen Studienkontext als auch die Qualität der Originaldaten unklar. Die Parameter zur Schätzung der Qualitätsbereinigung von Lebenszeit und -erwartung basierend auf Zahngesundheit und Übergewicht wurden ursprünglich für Erwachsene entwickelt und validiert, es gibt jedoch nur begrenzte Evidenz für ihre Angemessenheit bei Kindern (Brown et al., 2018).

Modellspezifikation

Zur Schätzung der Veränderung der durch Getränkekäufe erworbenen Zuckermenge wurde der Zeitraum von März 2014 bis November 2019 gewählt, um Effekte durch den Brexit oder die Corona-Pandemie auszuklammern. Dies erhöht die Interpretierbarkeit der Ergebnisse. Zur Ermittlung der Ergebnisse wird eine Abfolge von drei verschiedenen Modellierungsverfahren verwendet:

[53] Die damit einhergehenden Folgen für die Validität der Studienergebnisse werden im Abschnitt *Ergebniskommunikation* erläutert.

(1) Zur Abschätzung der Veränderung der durch Getränkekäufe erworbenen Menge an Zucker in einem Haushalt wird eine unterbrochene Zeitreihenanalyse (siehe Anhang A.2.1) mit zwei Interventionspunkten (Ankündigung sowie Einführung der SDIL) durchgeführt, die als Einflussfaktoren die monatlichen Durchschnittstemperaturen, sowie die (Nach-)Weihnachtszeit[54] berücksichtigt. Wie bereits in Abschnitt 4.3.2 beschrieben, sind unterbrochene Zeitreihenanalysen generell empfindlich gegenüber der Wahl des Interventionszeitpunkts. Die Wirkrichtung der Einfluss- auf die Zielgrößen wird bei Zeitreihenanalysen ex ante angenommen und nicht durch das Modell nachgewiesen. Werden Kontrollvariablen nicht ausreichend berücksichtigt, müssen die beobachteten Effekte der SDIL nicht zwangsläufig in der vom Modell dargestellten Form existieren.

(2) Zur Abschätzung der Prävalenz von Übergewicht (und Adipositas) sowie der Zahngesundheit von Kindern (= mittelfristige Effekte) wird eine Kohortensimulation (siehe Anhang A.1.1) über einen Zeitraum von zehn Jahren (2015 bis 2025) durchgeführt, welche die Veränderung der durch Getränkekäufe erworbenen Zuckermenge als zentrale Einflussgröße annimmt. Im Anschluss wird die Kohortensimulation um die Prognose der QALYs der Kinder erweitert, wofür Adipositasprävalenz und Zahngesundheit als Einflussgrößen dienen. Die Wirkrichtung der Einfluss- auf die Zielgrößen wird bei Kohortensimulationen ex ante angenommen und nicht durch das Modell nachgewiesen. Werden Kontrollvariablen nicht ausreichend berücksichtigt, müssen die beobachteten Effekte der SDIL nicht zwangsläufig in der vom Modell dargestellten Form existieren.

(3) Zur Prognose der Lebenserwartung (= langfristige Effekte) wird eine Kohortensimulation über einen Zeitraum von 100 Jahren durchgeführt, die BMI und Zahngesundheit als Einflussfaktoren berücksichtigt.

In allen drei Verfahren wird die tatsächliche Situation der Steuereinführung mit einem kontrafaktischen Szenario ohne SDIL[55] verglichen. Es wird angenommen, dass die beobachtete Steuerwirkung auch langfristig in unveränderter Form bestehen bleibt, ohne jedoch mögliche dynamische Effekte wie Verhaltensanpassungen der Konsumenten, Preisstrategien der Produzenten oder Veränderungen in der Marktstruktur zu berücksichtigen. Die Vernachlässigung dieser Dynamik stellt eine erhebliche Vereinfachung des Modells im Vergleich zur komplexen Realität dar und kann großen Einfluss auf die Verlässlichkeit der Ergebnisse, insbesondere bei einem Modellierungszeitraum über mehrere Jahrzehnte, haben.

Stark vereinfachend setzt die Studie die Änderung des Kaufverhaltens mit einer Änderung des Konsumverhaltens gleich: (1) Es wird angenommen, dass die durch Getränkekäufe erworbene Zuckermenge vollständig konsumiert wird, was nicht zwangsläufig zutrifft. (2) Die fehlende Differenzierung nach Alter und Geschlecht lässt die Heterogenität des Konsums vollständig unberücksichtigt. Es kann daher nicht davon ausgegangen werden, dass der durchschnittliche Haushaltskonsum von Zucker durch Verzehr von Getränken repräsentativ für den Konsum von Kindern ist.

[54] Diese Kontrollvariablen dienen der Abbildung saisonaler Trends. Die Relevanz der Monate Dezember und Januar wurde im Rahmen einer Voranalyse ermittelt.

[55] Hierfür wird ausgehend von einem Trend vor der SDIL extrapoliert, wie die Werte ausfielen, wenn es keine Steuer gäbe. Die geschätzten Veränderungen aufgrund der Steuer hängen stark vom gewählten Zeitpunkt ab, der für die Ermittlung des Trends verwendet wird. Zur Kontrolle allgemeiner Trends in Haushaltskäufen, die nicht durch die SDIL beeinflusst werden, nutzen die Studienautoren Käufe von Toilettenartikeln wie Shampoo und Seife als Kontrollvariable. Diese gelten als weitgehend unbeeinflusst von saisonalen oder sozioökonomischen Faktoren. Allerdings repräsentieren sie ein anderes Kaufverhalten als der Erwerb von (zuckergesüßten) Getränken, was ihre Eignung als Kontrollgröße in diesem Kontext möglicherweise einschränkt.

Außerdem wird der durch verändertes Kaufverhalten geschätzte Zuckerkonsum in Kalorien umgerechnet und mit Energiebilanzgleichungen linear in BMI-Änderungen überführt, ohne kompensatorische Effekte wie Veränderungen der körperlichen Aktivität oder kalorische Substitution durch andere Nahrungsmittel zu berücksichtigen. Zudem wird bei der Prognose von Zahnkaries eine feste Dosis-Wirkungs-Beziehung[56] zwischen Zucker und Karies angenommen, ohne weitere Einflussfaktoren wie Zahnhygiene oder Fluorid einzubeziehen. In der Gesamtheit wird demnach kein realistisches Abbild der tatsächlichen Zusammenhänge gezeichnet, was die Verlässlichkeit der Ergebnisse einschränkt.

Ein zentraler Einflussfaktor, den die Studie berücksichtigt, ist der Grad der regionalen Benachteiligung. Insgesamt lässt die Modellierung allerdings mehrere relevante Einflussgrößen unberücksichtigt. Die Studienautoren erwähnen zwar, dass der Zuckerkonsum durch gekaufte Getränke bereits vor der Ankündigung der Steuer gesunken war, möglicherweise erklärbar durch parallel anlaufende andere Maßnahmen wie Aufklärungskampagnen. Dieser mögliche Einflussfaktor wird allerdings nicht im Modell berücksichtigt. Substitutionseffekte werden ebenfalls gänzlich vernachlässigt. Insgesamt können die simulierten Effekte daher wohl kaum wie vom Modell dargestellt eindeutig der Steuer zuzuschreiben sein.

Ergebniskommunikation

Alle Ergebnisse werden (differenziert nach Grad der Benachteiligung) mit Konfidenzintervallen angegeben, was die Unsicherheit aufgrund des möglichen Stichprobenfehlers nachvollziehbar macht. Eine weitere statistische Einordnung der Effekte, bspw. durch die Berechnung von p-Werten oder Effektstärken, erfolgt allerdings nicht. Dies schließt eine fundierte statistische Einordnung aus. Die Modellierungsergebnisse werden in absoluten Werten (z. B. Reduktion um 3.600 Kariesfälle, 95 % CI [946; 6.330]) angegeben, die aber nicht weiter kontextualisiert werden. Für die vollständige Einordnung der Ergebnisse fehlen Angaben zur jeweils zugrunde liegenden Basis (z. B. Gesamtzahl an Kariesfällen in der untersuchten Population). Ergänzende Angaben in natürlichen und relativen Häufigkeiten hätten die Einordnung erleichtern können.

Die Studienautoren weisen im Rahmen der Validierung ihrer Ergebnisse darauf hin, dass die von ihnen simulierte Reduktion der Zuckeraufnahme durch Getränke in englischen Haushalten mit 15,0 g pro Haushalt und Woche[57], 95% CI [10,3; 19,7], nahezu doppelt so hoch ausfällt wie die von Rogers, Pell et al. (2023) für Großbritannien simulierte Reduktion. Beide Studien verwenden vergleichbare Methoden, beruhen aber auf unterschiedlichen Datensätzen. Cobiac et al. (2024) führen diese Diskrepanz auf kleine Unterschiede in den Zeitreihendaten des Zuckerkonsums vor der Einführung der SDIL zurück. Dies zeigt, wie empfindlich (Langzeit-)Ergebnisse auf minimale Änderungen der Eingangsdaten reagieren. Cobiac et al. (2024) selbst heben hervor, dass in ihrer Studie selbst geringfügige Abweichungen von der anfänglichen Reduktion des Zuckerkaufs um $\pm 0,1$ g pro Monat die kontrafaktischen Schätzungen am Ende des Simulationszeitraums (November 2019) um $\pm 4,4$ g Zucker beeinflussen können. Die simulierten Ergebnisse unterliegen daher einer gewissen Unsicherheit und sind deshalb generell mit Vorsicht zu betrachten. Systematische Sensitvitätsanalysen zur besseren Abschätzung des Einflusses dieser Empfindlichkeit auf die simulierten Effekte führen Cobiac et al. (2024) nicht durch.

[56] Die Dosis-Wirkungs-Beziehung beschreibt, wie sich eine Veränderung der Menge (Dosis) eines Faktors auf die Stärke der Wirkung auswirkt (Bernabé et al., 2016).

[57] Eine Reduktion von 15 g Zucker pro Woche und Haushalt mit durchschnittlich 2,4 Personen entspricht einer Verringerung von 0,89 g Zucker pro Person und Tag, was einer Energieeinsparung von weniger als 3 kcal pro Tag entspricht (eigene Berechnung basierend auf den Werten von Cobiac et al., 2024). Diese Reduktion erscheint äußerst gering.

Darüber hinaus wird die Aussagekraft der Ergebnisse zusätzlich durch die bereits erwähnte fehlende Differenzierung des Zuckerkonsums zwischen Kindern und Erwachsenen verzerrt, insbesondere in Kombination mit der Tatsache, dass Kauf- statt tatsächlicher Konsumdaten verwendet wurden. Laut den Berechnungen von Cobiac et al. (2024), die auf Kaufdaten basieren, wurden kurz vor der Ankündigung der SDIL im März 2016 im Mittel 341,6 g Zucker pro Woche und Haushalt (entspricht 142,3 g pro Haushaltsmitglied) durch Getränke eingekauft. Im Vergleich dazu zeigt eine tatsächliche Verzehrsstudie, die *National Diet and Nutrition Survey 2016* (Public Health England, 2018), dass zehnjährige Kinder im Durchschnitt 80,1 g Zucker pro Woche[58] durch Getränke konsumieren. Dieser Wert umfasst, anders als die Datengrundlage von Cobiac et al. (2024), auch außer Haus konsumierte Getränke und liegt dennoch deutlich unter dem von Cobiac et al. (2024) berechneten Pro-Kopf-Einkauf von Zucker durch Getränke. Da die von Cobiac et al. (2024) simulierten Effekte der SDIL nicht auf tatsächlichem Verzehr basieren, könnten sie erheblich überschätzt sein und sollten kritisch hinterfragt werden.

Die Studienutoren setzen sich mit verschiedenen Annahmen, Unsicherheiten und Limitationen ihrer Studie auseinander. So erwähnen sie, dass Ko-Interventionen, wie etwa Aufklärungskampagnen eine klare Abgrenzung des Effekts der SDIL erschweren. Zudem diskutieren sie, dass das kontrafaktische Szenario auf der Annahme beruht, dass sich die Kauftrends ohne Intervention linear fortgesetzt hätten und die Effekte der SDIL langfristig konstant blieben – Annahmen, die kritisch hinterfragt werden müssen. Sie erwähnen auch, dass die Berechnung des durchschnittlichen Konsums pro Person individuelle Unterschiede innerhalb der Haushalte unberücksichtigt lässt. Sie geben außerdem an, dass kompensatorische Verhaltensänderungen, wie bspw. eine Anpassung der körperlichen Aktivität, ebenfalls unberücksichtigt blieben. Sie begründen dies mit der Annahme, dass solche Veränderungen im dreijährigen Beobachtungszeitraum nach Ankündigung der SDIL noch nicht eingetreten seien. Sie weisen außerdem darauf hin, dass die Schätzungen der qualitätsbereinigten Lebensjahre und der Lebenserwartung aufgrund erheblicher Schwankungen und Unsicherheiten in den aus Metastudien abgeleiteten Parametern mit Vorsicht interpretiert werden sollten.

Insgesamt erfolgt die Ergebniseinordnung aber stark generalisiert: (1) Während die Studienautoren im Ergebnisteil zutreffend feststellen, dass die Veränderungen in der Lebenserwartung nur geringfügig ausfallen (ohne jedoch genauer zu quantifizieren, wie klein diese Effekte tatsächlich sind), wird in der Zusammenfassung die Aussage getroffen, dass die SDIL langfristige Verbesserungen der Lebenserwartung bewirken werde. Diese Schlussfolgerung ist jedoch nicht nachvollziehbar, da sich die Effekte je nach Grad der sozioökonomischen Benachteiligung lediglich im Bereich von wenigen Tagen bis maximal einem Monat bewegen und somit als gering einzustufen sind. Die Aussage der Studienautoren stellt daher eine deutliche Überbewertung der simulierten Effekte dar, insbesondere da die Unsicherheit der Ergebnisse, die es gerade bei Langzeiteffekten zu beachten gilt, dabei vollständig unberücksichtigt bleibt. (2) Die Studienautoren weisen zwar zutreffend darauf hin, dass die gesundheitlichen Ergebnisse, wie reduzierte BMI-Werte oder weniger Zahnkariesfälle, hauptsächlich korrelativ mit der Einführung der SDIL verbunden sind, da andere relevante Einflussfaktoren im Modell nicht berücksichtigt wurden. Dennoch ziehen sie eine kausale Schlussfolgerung, indem sie behaupten, die SDIL führe zu einer Reduktion von Übergewicht, Adipositas und Zahnkaries sowie zu einer Verbesserung der Lebensqualität.

[58] Nach Angaben von Public Health England (2018) konsumieren sie insgesamt 52 g Zucker pro Tag, d. h. 364 g pro Woche. Davon ist ein Anteil von 22 % auf Getränke zurückführbar. Damit entfallen 80,1 g Zucker pro Woche auf Getränke.

Zusammenfassung der Limitationen der Studie von Cobiac et al. (2024)

Datengrundlage

- **Stichprobe:** Aufgrund der Verwendung amtlicher Bevölkerungsdaten der gesamten interessierenden Population bestehen keine Limitationen bezüglich der Repräsentativität der Stichprobe.
- **Weitere Eingangsdaten:** Getränkekaufdaten einer Durchschnittsperson (zumal basierend auf Selbstauskünften und ohne Berücksichtigung der Gastronomie) spiegeln spezifisches Kaufverhalten der Stichprobe nicht angemessen wider und verzerren Ergebnisse; mögliche Kompatibilität der Datenquellen (Haushaltseinkäufe und Haushaltsgröße), dadurch potenzielle Abweichungen in den Ergebnissen; Gesundheitsdaten teilweise basierend auf Selbstauskünften sowie Metastudien mit unklarer Qualität der Originaldaten und nicht nachgewiesener Eignung für den Studienkontext, sodass Ergebnisse potenziell unsicherheitsbehaftet sind; lediglich grobe (regionale) Klassifikation des sozioökonomischen Status, sodass diesbezügliche Schlussfolgerungen unsicherheitsbehaftet sind.

Modellspezifikation

- **Modellierungsverfahren:** Kausalität wird bei Zeitreihenanalysen und Kohortensimulationen ex ante bei der Modellkonstruktion unterstellt, nicht aber durch das Modell selbst nachgewiesen; generelle Empfindlichkeit von Zeitreihenanalysen bei Berechnung des simulierten Trends (Szenario ohne Steuer) gegenüber Interventionszeitpunkt mit Risiko für verzerrte Effektschätzungen.
- **Einfluss- und Zielgrößen:** Zahlreiche vereinfachende Annahmen (langfristige konstante Steuerwirkung, Gleichsetzung von Konsum- mit Kaufverhalten, Vernachlässigung der Heterogenität des Konsumverhaltens, lineare Beziehung zwischen Kalorienaufnahme und Gewicht vernachlässigt Kompensationseffekte, feste Dosis-Wirkungs-Beziehung zwischen Zucker und Karies vernachlässigt individuelle Zahnhygiene) bilden reale Phänomene und ihre Beziehungen nicht präzise ab.
- **Weitere Einflussfaktoren:** Nichtberücksichtigung relevanter Faktoren (z. B. Substitution, Aufklärungskampagnen), sodass die simulierten Effekte die realen Zusammenhänge möglicherweise nicht korrekt beschreiben.

Ergebniskommunikation

- **Ergebnisdarstellung:** Das Fehlen von p-Werten und Effektstärken verhindert die fundierte Einordnung der statistischen Relevanz; die fehlende Kontextualisierung erschwert die Ergebniseinordnung.
- **Modellbewertung:** Diskrepanzen in den Ergebnissen im Vergleich zu anderen Studien (Zuckerverzehr und -veränderung) deuten auf Verzerrungen in den Ergebnissen und eine Überschätzung der Effekte hin; die Empfindlichkeit der Ergebnisse auf minimale Änderungen der Ausgangsdaten führt zu Unsicherheiten, die nicht mithilfe von Sensitivitätsanalysen abgeschätzt werden.
- **Ergebniseinordnung:** Vernachlässigung von Unsicherheit und Überbewertung von Effekten (Lebenserwartung); die kausale Schlussfolgerung (SDIL führe zu Reduktionen von Übergewicht, Adipositas und Zahnkaries sowie verbesserter Lebensqualität) wird nicht durch das Modell gestützt.

Gesamtbeurteilung der Studie von Cobiac et al. (2024)

Die Verlässlichkeit der Studienergebnisse von Cobiac et al. (2024) ist insbesondere aufgrund der unzureichenden Datengrundlage, zahlreicher vereinfachender Modellannahmen sowie der Nichtberücksichtigung relevanter Einflussfaktoren kritisch zu hinterfragen. Besonders problematisch ist, dass der simulierte Zuckerkonsum von Jugendlichen auf Verkaufsdaten auf Haushaltsebene basiert und die tatsächliche Aufnahme nicht realistisch widerspiegelt. Auf dieser fragilen Grundlage werden gesundheitsbezogene Effekte modelliert, was belastbare Schlussfolgerungen ausschließt. Die stark vereinfachenden Annahmen zur Beziehung zwischen der Steuer und den Effekten, etwa die lineare Beziehung zwischen Kalorienaufnahme und Gewicht ohne Berücksichtigung kalorischer Kompensation oder anderer Einflussfaktoren auf das Körpergewicht, werden der Realität nicht gerecht. Darüber hinaus ist davon auszugehen, dass die simulierten Effekte nicht in der dargestellten Form existieren, sondern teilweise auf andere Einflussfaktoren wie Aufklärungskampagnen zurückzuführen sind. Validierungen legen zudem nahe, dass die von Cobiac et al. (2024) simulierten Effekte überschätzt wurden. Angesichts dieser Schwächen sind die Modellierungsergebnisse und Schlussfolgerungen der Studienautoren als unzuverlässig einzustufen, sodass die Studie keine belastbare Grundlage für politische Entscheidungen darstellt.

4.5 Studie von Gračner et al. (2022), Mexiko

Vollständige Quellenangabe zur Studie:

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

Die Studie von Gračner et al. (2022) untersucht den Zusammenhang zwischen Preisänderungen zuckergesüßter Getränke infolge der in Mexiko eingeführten Steuer und gewichtsbezogenen Parametern bei mexikanischen, in Städten wohnhaften Jugendlichen mithilfe multivariater Regressionsmodelle. Die Modellierungsergebnisse zeigen, dass eine 10%-ige Preiserhöhung von zuckergesüßten Getränken bei jugendlichen Mädchen mit einem relativen Rückgang ihrer Prävalenz von Übergewicht und Adipositas um 3 % einhergeht. Die beobachteten Gewichtsverluste sind gering und treten vorwiegend bei Mädchen mit höherem Ausgangsgewicht auf, die in Städten leben, in denen die Preissteigerungen überdurchschnittlich hoch waren. Für Jungen werden keine vergleichbaren Effekte nachgewiesen. Die Studienautoren schließen daraus, dass große Preiserhöhungen mit beobachtbaren Änderungen gewichtsbezogener Parameter einhergehen könnten. Diese Schlussfolgerung scheint angesichts der berichteten Ergebnisse sowie zahlreicher Studienlimitationen jedoch stark verallgemeinert. Besonders die Auswahl der Stichprobe, die ausschließlich Jugendliche umfasst, die in Städten leben und bei einem bestimmten Versicherungsunternehmen krankenversichert sind, schränkt die Übertragbarkeit der Ergebnisse auf die gesamte (jugendliche) Bevölkerung Mexikos erheblich ein. Insgesamt erfüllt die Studie aus statistisch-methodischer Perspektive nicht die erforderlichen Qualitätsstandards, um als belastbare Grundlage für evidenzbasierte Entscheidungen zu dienen.

4.5.1 Zusammenfassung der Studie

Die Inhalte der Studie von Gračner et al. (2022) werden nachfolgend zusammengefasst.

Untersuchungsgegenstand

Die Studie untersucht den Zusammenhang zwischen Preisänderungen[59] zuckergesüßter Getränke infolge der im Januar 2014 in Mexiko eingeführten spezifischen Verbrauchsteuer von 1 Mex$ pro Liter und verschiedenen gewichtsbezogenen Parametern, einschließlich des BMI und des Risikos sowie der Prävalenz von Übergewicht und Adipositas[60] bei Jugendlichen in mexikanischen Städten.

Methodik

Untersucht werden 12.654 Jugendliche aus 39 mexikanischen Städten, die zwischen 1999 und 2002 geboren wurden. Basierend auf Longitudinaldaten zu ihrem Körpergewicht und ihrer Körpergröße wird für den Zeitraum von 2012 bis 2017 mithilfe von multivariaten Regressionsmodellen die Prävalenz für Übergewicht und Adipositas

[59] Die Preisänderungen fielen je nach Stadt unterschiedlich aus.
[60] Übergewicht und Adipositas werden in dieser Studie nicht weiter differenziert.

in der Untersuchungsgruppe simuliert. Dabei berücksichtigt die Analyse auch das Ausmaß der Preisveränderungen der besteuerten Getränke, wobei die Städte in drei Kategorien eingeteilt werden (Preisanstieg < 5 %, 5-10 % und > 10 %). Zur Überprüfung der Robustheit der Ergebnisse werden zudem verschiedene Sensitivitätsanalysen durchgeführt.

Ergebnisse

Die Modellierungsergebnisse zeigen, dass ein Preisanstieg zuckergesüßter Getränke um 10 %[61] bei Mädchen innerhalb von zwei Jahren nach der Preisänderung mit einem absoluten Rückgang der Prävalenz von Übergewicht und Adipositas um 1,3 Prozentpunkte, 95 % CI $[-2,1;\ -0,38]$, $p = 0,006$, bzw. einem relativen Rückgang um 3,0 %[62] einhergeht. Bei Mädchen mit höherem Ausgangsgewicht wird eine Reduktion ihres BMI simuliert, die einer Gewichtsreduktion von etwa 0,35 kg[63] entspricht. In Städten, in denen die Preissteigerungen infolge der Steuer überdurchschnittlich, also mehr als 10 %, waren[64], werden größere Effekte bei den Ergebnissen der Mädchen simuliert. Für Jungen zeigen sich keine derartigen Zusammenhänge.

Schlussfolgerungen der Studienautoren

Die Studienautoren bewerten die simulierten Gewichtsverluste, die hauptsächlich bei Mädchen mit höherem Ausgangsgewicht und wohnhaft in Städten mit Preissteigerungen von mehr als 10 % auftraten, als gering. Sie folgern, dass große Preiserhöhungen mit beobachtbaren Änderungen gewichtsbezogener Parameter einhergehen könnten.

4.5.2 Evaluation der Studie

Die Studie von Gračner et al. (2022) wird nachfolgend anhand der in Kapitel 3 vorgestellten Beurteilungskriterien evaluiert.

Datengrundlage

Die Stichprobe der zwischen 1999 und 2002 geborenen Jugendlichen war im Beobachtungszeitraum (2012-2017) zehn bis 18 Jahre alt. Um die Kompatibilität mit den Preisdaten zu gewährleisten, werden ausschließlich Gesundheitsdaten von Jugendlichen aus 39 mexikanischen Städten[65] einbezogen, nicht aber aus ländlichen Gebieten. Zwischen städtischen und ländlichen Regionen bestehen typischerweise systematische Unterschiede, etwa im sozioökonomischen Status oder im Zugang zu Gesundheitseinrichtungen. Die Daten zu den Körpermaßen der Jugendlichen stammen vom öffentlichen Gesundheitsinstitut *Instituto Mexicano del Seguro Social* (kurz: *IMSS*) und umfassen Jugendliche, die über das IMSS krankenversichert waren und jährlich[66] ambulante Gesundheitsbehandlungen innerhalb der gleichen Stadt in Anspruch nahmen. Im Rahmen dieser Behandlungen wurden auch ihre Körpermaße erfasst. Für jeden Jugendlichen standen entsprechend sechs (jährlich gemittelte) Datenpunkte zu Körpergewicht und -größe für die Analyse zur Verfügung. Mithilfe solcher Longitudinaldaten ist es möglich, auch individuelle Veränderungen im Zeitverlauf zu beobachten. Die Repräsentativität dieser Daten ist allerdings eingeschränkt: Das IMSS versichert den Studienautoren zufolge Arbeitnehmer in formellen

[61] Dies entspricht der durchschnittlichen Preiserhöhung bei zuckergesüßten Getränken in Mexiko.
[62] keine Angabe eines zugehörigen Konfidenzintervalls
[63] keine Angabe eines zugehörigen Konfidenzintervalls
[64] Die durchschnittliche Preissteigerung in diesen Städten wird mit 16.6 % angegeben.
[65] Im Detail: städtische Regionen im Umkreis von 30 Kilometer vom Stadtzentrum.
[66] mindestens einmal und maximal zehnmal pro Kalenderjahr; im Durchschnitt 3,59 Behandlungen jährlich

Beschäftigungsverhältnissen sowie deren Familien[67]. Dadurch könnten Jugendliche aus Haushalten mit stabileren Einkommens- und Bildungsniveaus überrepräsentiert sein. Zusätzlich könnten systematische Unterschiede zwischen Jugendlichen, die jährlich in der Klinik untersucht wurden (wie in der Stichprobe erfasst), und solchen, die dieses Kriterium nicht erfüllen (nicht in der Stichprobe), bestehen. Laut den Studienautoren ist die untersuchte Stichprobe zudem jünger[68], etwas schwerer[69] und häufiger weiblich[70] als der durchschnittliche mexikanische Jugendliche. Die Studienergebnisse sind daher nicht uneingeschränkt auf die gesamte jugendliche Bevölkerung Mexikos übertragbar, insbesondere nicht auf Jugendliche, die in ländlichen Gebieten leben.

Die berechneten inflationsbereinigten jährlichen Durchschnittspreise für zuckergesüßte Getränke basieren auf Preisdaten des nationalen Instituts für Statistik und Geographie aus dem Zeitraum 2011-2016. Diese Daten umfassen monatliche Preisangaben für in Geschäften verkaufte Produkte in mexikanischen Städten, jedoch keine Informationen zu Verkäufen in Gastronomiebetrieben. In die Analyse wurden ausschließlich *kohlensäurehaltige* zuckergesüßte Getränke einbezogen, da die Daten für kohlensäurefreie zuckergesüßte Getränke nicht vollständig für alle untersuchten Städte vorhanden waren. Den Studienautoren zufolge werden von Jugendlichen allerdings vorrangig kohlensäurehaltige zuckergesüßte Getränke verzehrt. Außerdem hätten sich die Preise nach der Steuererhebung vor allem bei kohlensäurehaltigen zuckergesüßten Getränken signifikant verändert. Aufgrund der Nicht-Berücksichtigung der Preise in gastronomischen Vertriebsstätten sowie von kohlensäurefreien Getränken können Abweichungen zwischen den verwendeten und den tatsächlichen Preisänderungen zuckergesüßter Getränke nicht ausgeschlossen werden. Die Städte werden basierend auf den beobachteten Preisänderungen in drei Gruppen eingeteilt: geringe (< 5 %; neun Städte[71]; 1.499 Jugendliche), mittlere (5–10 %; sieben Städte; 6.986 Jugendliche) und hohe (> 10 %; 23 Städte; 4.169 Jugendliche) Preisänderungen.

Die Verknüpfung der Gesundheitsdaten mit den Preisdaten erfolgt über den Standort der von den Jugendlichen aufgesuchten Gesundheitseinrichtungen. Dieser repräsentiert den Studienautoren zufolge in der Regel auch den Wohnort der Jugendlichen, da Patienten den Gesundheitseinrichtungen anhand ihrer Wohnadresse zugewiesen werden. Die Zuweisung scheint auch dahingehend zuverlässig, da nur Jugendliche mit Arztbesuch in derselben Stadt während des sechsjährigen Untersuchungszeitraums berücksichtigt werden. Einzelfälle, etwa regelmäßige Kontrolltermine in Spezialkliniken in anderen Städten, können jedoch nicht ausgeschlossen werden.

Erfasst werden zudem sozioökonomische Indikatoren der Städte, darunter der Anteil der Haushalte in extremer Armut sowie der Anteil der Bevölkerung in informellen Beschäftigungsverhältnissen. Eine vollständige Übersicht aller verwendeten Datenquellen findet sich im Anhang der Studie.

Modellspezifikation

Die Studie verwendet multivariate Regressionsmodelle (siehe Anhang A.2.2), um die Effekte einer Besteuerung zuckergesüßter Getränke auf BMI-Perzentile[72] sowie auf das Risiko und die Prävalenz von Übergewicht und Adipo-

[67] Das IMSS versichert insgesamt mehr als 50 % der mexikanischen Bevölkerung.
[68] Das Durchschnittsalter der untersuchten Jugendlichen wird mit 11,38 Jahren angegeben und bezieht sich mutmaßlich auf den Startpunkt der Untersuchung im Jahr 2012.
[69] Laut Studienautoren ist das Risiko sowie die Prävalenz von Übergewicht und Adipositas in der IMSS-Stichprobe höher, was vermutlich darauf zurückzuführen ist, dass schwerere Personen mit größerer Wahrscheinlichkeit eine Klinik aufsuchen.
[70] Laut Studienautoren spiegelt dies wider, dass jugendliche Mädchen tendenziell häufiger Gesundheitsdienste in Anspruch nehmen als Jungen.
[71] darunter drei Städte mit Preisreduktionen nach Einführung der Steuer
[72] basierend auf standardisierten BMI z-scores

sitas zu untersuchen. Multivariate Regressionsmodelle ermöglichen die gleichzeitige Analyse mehrerer Zielgrößen. Es ist jedoch zu beachten, dass dieses Verfahren keine Kausalaussagen erlaubt, sofern die zugrundegelegten Zusammenhänge nicht bereits zuvor als kausal nachgewiesen wurden.

Das Modell nimmt ohne empirische Belege einen linearen Zusammenhang zwischen Preisänderungen und Körpermaßen an, was die Realität stark vereinfacht. Als Prädiktoren werden logarithmierte Getränkepreise mit einer Verzögerung von ein und zwei Jahren herangezogen, um zeitlich verzögerte Reaktionen zu erfassen. Zur Kontrolle allgemeiner Trends in den Ergebnissen und landesweiter Einflüsse werden feste Effekte für die beobachteten Jahre einbezogen. Zudem integriert das Modell Kontrollvariablen auf individueller, klinischer und städtischer Ebene, darunter Alter, Häufigkeit der Arztbesuche und diagnostizierte Gesundheitszustände. Trotz der Berücksichtigung zahlreicher Variablen bleiben potenziell relevante Einflussfaktoren, welche die Beziehung von Preisänderungen und Körpermaßen beeinflussen könnten, unberücksichtigt, wie etwa umfassende Substitutionseffekte (z. B. Preisentwicklung anderer kalorienhaltiger Getränke), Umweltbedingungen (z. B. Zugang zu Grünflächen), körperliche Aktivität oder Initiativen (z. B. Aufklärungskampagnen, Maßnahmen für gesünderes Essen oder mehr Bewegung in Schulen). Außerdem wird die Heterogenität innerhalb der untersuchten Stichprobe vernachlässigt, da keine individuellen Unterschiede, etwa in Bezug auf den sozioökonomischen Status oder den Basiskonsum zuckergesüßter Getränke, berücksichtigt werden.

Ergebniskommunikation

Die Ergebnisdarstellung erfolgt generell vollständig und beinhaltet im Fließtext gängige statistische Kennwerte wie Konfidenzintervalle und p-Werte. In den Ergebnistabellen werden allerdings keine p-Werte berichtet, was die Ergebniseinordnung erschwert. Veränderungen in den untersuchten gewichtsbezogenen Parametern wie die Prävalenz von Übergewicht und Adipositas werden sowohl absolut (bezogen auf die Gesamtstichprobe) als auch relativ (bezogen auf die jeweilige Gruppe) berichtet. Wo erforderlich, erfolgt eine Kontextualisierung der Werte (z. B. Umrechnung der BMI-Reduktion in Gewichtsverlust), was die Interpretation der Ergebnisse erleichtert.

Die Methodik der Studie ist detailliert beschrieben und fehlende Aspekte werden umfassend im online verfügbaren Anhang der Studie erläutert. Die Validität der Ergebnisse wird durch den Vergleich mit vorherigen Studien untermauert, in die sich die berichteten Befunde konsistent einfügen. Um die Robustheit der Hauptergebnisse zu überprüfen, werden mehrere Sensitivitätsanalysen (siehe Anhang A.3.2) durchgeführt. Diese umfassen Anpassungen für regional spezifische lineare Zeittrends, die Berücksichtigung verzögerter Reaktionen auf mögliche Substitute, alternative gewichtsbezogene Zielgrößen sowie Anpassungen in der Stichprobe. Die Ergebnisse erweisen sich zwar als robust gegenüber diesen variierten Annahmen, allerdings kann damit mögliche modellbedingte Unsicherheit in den Ergebnissen, die auf andere zentrale methodische Annahmen (z. B. unberücksichtigte relevante Einflussfaktoren, Annahme eines linearen Zusammenhangs zwischen Preisänderungen und Körpermaßen) zurückzuführen sind, nicht ausgeschlossen werden.

Die Studienautoren geben an, Berichtsrichtlinien für Beobachtungsstudien zu folgen. So benennen sie auch transparent mehrere Limitationen ihrer Studie. Dazu zählen die eingeschränkte Generalisierbarkeit der Ergebnisse auf die gesamte jugendliche Bevölkerung Mexikos aufgrund der fehlenden Repräsentativität der Stichprobe. Diese Limitation ist jedoch noch weiter auszudehnen: Die Übertragbarkeit dieser Ergebnisse auf andere Länder ist

ebenfalls fragwürdig, da die Prävalenz von Übergewicht und Adipositas von Jugendlichen in Mexiko vor der Steuer bei 46,1 % lag, während sie bspw. in Deutschland im gleichen Zeitraum bei Kindern und Jugendlichen nur bei 15,4 % lag (Bundesministerium für Gesundheit, 2024). Weiterhin berichten die Studienautoren von der Einschränkung durch Nichtberücksichtigung von Heterogenität der Stichprobe mangels verfügbarer Daten, der Nutzung lediglich angenäherter Preise sowie einer lediglich unvollständigen Berücksichtigung von Faktoren, welche die Höhe der Preisänderungen beeinflussen. Die Studienautoren weisen offen darauf hin, dass trotz umfangreicher Sensitivitätsanalysen keine Kausalaussagen aus den Ergebnissen abgeleitet werden können. Sie sprechen konsequent richtigerweise von Zusammenhängen und keinen Ursache-Wirkungs-Beziehungen. Insgesamt vermittelt die Erzählstruktur aber denoch stellenweise den Eindruck einer größeren Sicherheit in den Ergebnissen als tatsächlich gegeben ist. Dies spiegelt sich bspw. in der vergleichsweise verallgemeinerten Schlussfolgerung wider, dass große Preiserhöhungen mit beobachtbaren Änderungen gewichtsbezogener Parameter einhergehen könnten. Die simulierten Ergebnisse, die dieser Aussage zugrunde liegen, stützen sich jedoch ausschließlich auf spezifische Teilgruppen der Stichprobe (z. B. Mädchen oder übergewichtige Mädchen) und auf bestimmte Bedingungen (z. B. bei einer verzögerten Berücksichtigung der Preise um zwei Jahre, jedoch nicht nach einem Jahr).

Zusammenfassung der Limitationen der Studie von Gračner et al. (2022)

Datengrundlage

- **Stichprobe:** Aufgrund nicht-repräsentativer Stichprobe mexikanischer Jugendlicher (ausschließlich in städtischen Gebieten wohnhaft und bei einem bestimmten Unternehmen krankenversichert, daher mutmaßlich sozioökonomisch bevorteilt; darüber hinaus jünger, schwerer und weiblicher) sind Studienergebnisse verzerrt und damit nur eingeschränkt auf die gesamte jugendliche Bevölkerung Mexikos übertragbar.
- **Weitere Eingangsdaten:** Aufgrund mangelnder Datenverfügbarkeit bei Preisdaten werden lediglich kohlensäurehaltige zuckergesüßte Getränke in nicht-gastronomischen Vertriebsstätten berücksichtigt, dadurch entstehen potenzielle Abweichungen von der tatsächlichen, gesamten Preisänderung.

Modellspezifikation

- **Modellierungsverfahren:** Kausalität und lineare Zusammenhänge werden bei multivariaten Regressionsanalysen ex ante bei der Modellkonstruktion unterstellt, nicht aber durch das Modell selbst nachgewiesen.
- **Einfluss- und Zielgrößen:** Durch Annahme eines linearen Zusammenhangs zwischen Preisänderungen und Körpermaßen ohne stützende Belege erfolgt eine stark vereinfachte Abbildung der realen Phänomene und ihrer Beziehungen.
- **Weitere Einflussfaktoren:** Aufgrund unzureichender Berücksichtigung von Substitutionseffekten, Vernachlässigung der Heterogenität der Stichprobe (z. B. sozioökonomischer Status, Basiskonsum) sowie Nichtberücksichtigung von Umweltbedingungen, körperlicher Aktivität oder Initiativen müssen simulierte Effekte der SDIL nicht zwangsläufig in der dargestellten Form existieren.

Ergebniskommunikation

- **Ergebnisdarstellung:** Die teilweise lediglich unvollständige Darstellung von p-Werten erschwert die Einordnung mancher Ergebnisse.
- **Modellbewertung:** Trotz Validierung der Ergebnisse anhand vorheriger Studien und Untermauerung durch verschiedene Sensitivitätsanalysen ist von fortbestehender Unsicherheit in den Ergebnissen aufgrund Nichtberücksichtigung diverser potenziell relevanter Einflussfaktoren auszugehen.
- **Ergebniseinordnung:** Schlussfolgerung der Studienautoren (Einhergehen großer Preiserhöhungen mit beobachtbaren Änderungen gewichtsbezogener Parameter) erweckt Eindruck großer Sicherheit, obwohl Beleg in der Modellierungsstudie lediglich für bestimmte Subgruppen (Mädchen) und unter bestimmten Bedingungen (nach zwei Jahren) erfolgte.

Gesamtbeurteilung der Studie von Gračner et al. (2022)

Die Studie von Gračner et al. (2022) beruht auf Daten einer nicht-repräsentativen Stichprobe mexikanischer Jugendlicher, wodurch die Ergebnisse nur begrenzt auf die gesamte jugendliche Bevölkerung Mexikos übertragbar sind. Eine Generalisierung der Ergebnisse auf Deutschland ist aufgrund grundlegender Unterschiede u. a. in den Körpermaßen Jugendlicher nicht möglich. Darüber hinaus bildet das in der Studie verwendete Modell die komplexe Realität nur in vereinfachter Form ab, da es wesentliche Einflussfaktoren wie die Heterogenität der Stichprobe oder umfassende Substitutionseffekte außer Acht lässt. Trotz Validierung der Ergebnisse durch frühere Studien und zusätzliche Sensitivitätsanalysen bleibt daher eine gewisse Unsicherheit bestehen. Insgesamt muss die Verlässlichkeit der Ergebnisse angesichts der genannten Einschränkungen kritisch hinterfragt werden, weshalb die Studie nicht als belastbare Grundlage für politische Entscheidungen herangezogen werden sollte.

4.6 Studie von Basto-Abreu et al. (2019), Mexiko

Vollständige Quellenangabe zur Studie:

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

Die Studie von Basto-Abreu et al. (2019) modelliert gesundheitsbezogene und gesundheitsökonomische Auswirkungen einer 2014 in Mexiko eingeführten Steuer auf zuckergesüßte Getränke sowie die potenziellen Effekte einer hypothetischen Verdopplung des Steuersatzes mithilfe einer Kohortensimulation. Die Studienautoren schließen aus den Simulationsergebnissen, dass die Steuereinführung in beiden Szenarien zu einer Reduktion der Fälle von Adipositas führe, wodurch Krankheiten verhindert, die Lebensqualität verbessert und Gesundheitskosten eingespart werden könnten. Die Simulationsergebnisse basieren jedoch auf einer fragwürdigen Datengrundlage, die das tatsächliche Kauf- und Konsumverhalten der mexikanischen Bevölkerung nicht realistisch abbildet. Außerdem werden stark vereinfachende Annahmen u. a. zur Steuerwirkung zugrundegelegt. Die Belastbarkeit der Studienergebnisse ist daher grundsätzlich kritisch zu hinterfragen. Zudem ist das verwendete Modell nicht in der Lage, Evidenz dafür zu erbringen, dass die Steuer tatsächlich ursächlich für diese Effekte ist. Insgesamt erfüllt die Studie aus statistisch-methodischer Perspektive nicht die erforderlichen Qualitätsstandards, um als belastbare Grundlage für evidenzbasierte Entscheidungen zu dienen.

4.6.1 Zusammenfassung der Studie

Die Inhalte der Studie von Basto-Abreu et al. (2019) werden nachfolgend zusammengefasst.

Untersuchungsgegenstand

Die Studie simuliert die Effekte der im Jahr 2014 in Mexiko eingeführten spezifischen Verbrauchsteuer auf zuckergesüßte Getränke bei Kindern und Erwachsenen über einen Zeitraum von zehn Jahren. Dafür werden mögliche Auswirkungen der Steuer auf gesundheitsbezogene und gesundheitsökonomische Aspekte in zwei Szenarien simuliert: (I) das tatsächliche Szenario der eingeführten spezifischen Verbrauchsteuer in Höhe von 1 Mex$ pro Liter zuckergesüßte Getränke und (II) ein hypothetisches Szenario einer spezifischen Verbrauchsteuer in Höhe von 2 Mex$ pro Liter.

Methodik

Die Schätzung der Steuerauswirkungen stützt sich auf Daten zu demografischen sowie zu ernährungs- und gesundheitsbezogenen Merkmalen der mexikanischen Bevölkerung, Bevölkerungsprognosen, Daten zu mit Adipositas verbundenen Krankheiten und deren Mortalität sowie auf Daten zu Gesundheitsausgaben. Mithilfe eines Kohortensimulationsmodells werden Veränderungen in der Verteilung des BMI der mexikanischen Bevölkerung des Jahres 2014 (dem Zeitpunkt der Steuereinführung) im Alter von 2 bis 100 Jahren über einen Zeitraum

von zehn Jahren prognostiziert – sowohl unter der Annahme eines gleichbleibenden Konsums zuckergesüßter Getränke als auch unter der Annahme von Konsumveränderungen infolge einer Steuer in Höhe von 1 bzw. 2 Mex$ pro Liter. Die angenommenen Konsumveränderungen werden mithilfe einer Change-in-Change-Analyse (siehe Anhang A.3.4) in Veränderungen des BMI umgerechnet. Basierend auf den prognostizierten BMI-Verteilungen werden Fälle und Raten von Adipositas, vermiedene Krankheitsfälle (z. B. Diabetes), Lebensqualitätsindizes sowie Einsparungen bei den Gesundheitskosten abgeschätzt. Sensitivitätsanalysen untersuchen die Auswirkungen eines alternativen Diskontierungssatzes von 5 % anstelle der in der Hauptanalyse verwendeten 3 % auf die Gesundheitskosten.

Ergebnisse

Die Studie prognostiziert aufgrund der Steuer in Höhe von 1 Mex$ pro Liter zuckergesüßte Getränke folgende Veränderungen innerhalb von zehn Jahren: (1) eine Reduktion der Prävalenz von Adipositas um 239.900 Fälle, 95 % CI [173.000; 306.000], davon 94.300 bei Kindern, 95 % CI [51.600; 137.900], (2) eine Reduktion der Inzidenz verschiedener Krankheiten, darunter eine Reduktion von etwa 61.340 Fällen von Diabetes, 95 % CI [31.940, 95.170], (3) eine Verbesserung der Lebensqualität, wie etwa 918 gewonnene Lebensjahre, 95 % CI [493; 1.420], und (4) Gesundheitskosteneinsparungen in Höhe von etwa 92 Mio. US$[73], 95 % CI [47; 148], was einer Einsparung von knapp 4 US$ pro für die Steuereinführung ausgegebenem US$ entspräche, 95 % CI [2; 6]. Für das hypothetische Szenario einer Verdopplung der Steuer auf 2 Mex$ pro Liter zuckergesüßte Getränke wird in etwa eine Verdopplung dieser Effekte prognostiziert.

Schlussfolgerungen der Studienautoren

Die Studienautoren schließen aus ihren Prognosen, dass die Steuer auf zuckergesüßte Getränke in Mexiko eine wirksame Maßnahme zur Reduktion von Adipositasfällen und den damit verbundenen Krankheiten bei Kindern und Erwachsenen darstelle, die zu einer Verbesserung der Lebensqualität und Einsparungen bei den Gesundheitskosten führe. Außerdem interpretieren sie, dass eine Erhöhung der Steuer in Mexiko die positiven Effekte weiter verstärken könne. Sie mutmaßen, dass Länder mit ähnlichen Bedingungen von der Einführung einer vergleichbaren Steuer profitieren könnten.

4.6.2 Evaluation der Studie

Die Studie von Basto-Abreu et al. (2019) wird nachfolgend anhand der in Kapitel 3 vorgestellten Beurteilungskriterien evaluiert.

Datengrundlage

Die Analyse stützt sich auf eine synthetische Kohorte, die auf Basis nationaler Statisiken entwickelt wurde und die mexikanische Bevölkerung des Jahres 2014 im Alter von zwei bis 100 Jahren abbildet. Die Entwicklung dieser Kohorte im Hinblick auf gesundheitsbezogene Faktoren wird auf Basis amtlicher Bevölkerungprojektionen über einen Zeitraum von zehn Jahren simuliert, wobei die Prognose beim Tod oder beim Erreichen des Alters von 100 Jahren endet. Dafür wird auf Daten aus verschiedenen externen Studien zurückgegriffen, die nachfolgend evaluiert werden.

[73] Kosten wurden nicht in Mex$, sondern in US$ angegeben.

Die Daten zum Basiskonsum von zuckergesüßten Getränken stammen aus einer nationalen Gesundheits- und Ernährungsstudie (ENSANUT; Instituto Nacional de Salud Pública, 2012), in der die Teilnehmenden mithilfe eines semi-quantitativen Fragebogens die Menge der in den letzten 24 Stunden konsumierten Getränke angeben sollten. Die Daten berücksichtigen zwar alters- und geschlechtsspezifische Unterschiede, aber ihre generelle Aussagekraft ist fragwürdig, da die Erhebung lediglich auf eine kurze Zeitspanne beschränkt ist. Zudem können Selbstauskünfte durch fehlerhafte Angaben beeinträchtigt sein, bspw. durch Erinnerungslücken, geschätzte Angaben oder sozial erwünschtes Antwortverhalten.

Die Daten zur Veränderung des Kaufverhaltens in Bezug auf zuckergesüßte Getränke infolge der Einführung der Steuer stammen aus der Studie von Colchero, Rivera-Dommarco et al. (2017). Diese vergleicht die tatsächlichen Einkaufsdaten der zwei Jahre nach Einführung der Steuer (2014 und 2015) mit einem kontrafaktischen Szenario ohne Steuer, das auf Trendprognosen basierend auf den Realdaten von 2012 bis 2013 beruht. Die Daten spiegeln das gesamte Einkaufsvolumen eines Haushalts wider, berücksichtigen jedoch keine individuellen Unterschiede. Entsprechend wird die Veränderung des Kaufverhaltens in Bezug auf zuckergesüßte Getränke (Preiselastizität der Nachfrage) als pauschaler (prozentualer) Wert über alle Bevölkerungsgruppen hinweg abgeleitet, was die Zuverlässigkeit der Ergebnisse einschränkt, da die Heterogenität in Bezug auf Alter und Geschlecht gänzlich vernachlässigt wird. Zudem wurden die Getränkekäufe ausschließlich auf Basis von Einzelhandelsdaten ermittelt, ohne gastronomische Daten zu berücksichtigen, was zu einer Verzerrung der Ergebnisse führt. Außerdem ist fraglich, inwieweit solche ausschließlich auf städtischen Gebieten basierenden Daten repräsentativ für die Gesamtbevölkerung sind, da Hinweise vorliegen, dass der Rückgang des Konsums zuckergesüßter Getränke in ländlichen Gebieten nach der Steuereinführung weniger stark ausgefallen ist (Colchero, Molina & Guerrero-López, 2017).

Die Umrechnung von Konsumänderungen in Gewichts- bzw. BMI-Veränderungen erfolgt durch Energiebilanzgleichungen mittels einer Change-in-Change-Analyse. Die verwendeten Umrechnungsfaktoren berücksichtigen den Studienautoren zufolge andere Einflussfaktoren wie körperliche Aktivität oder kalorische Kompensation. Die Umrechungsfaktoren für Erwachsene beruhen auf einer Untersuchung zum Gewichtsverlust bei mexikanischen Lehrerinnen. Da eine einzelne Berufs- und Geschlechtsgruppe nicht repräsentativ für die Gesamtbevölkerung ist, sind bei der Übertragung der Ergebnisse auf die gesamte erwachsene Bevölkerung Verzerrungen zu erwarten. Die Umrechnungsfaktoren für Kinder und Jugendliche basieren auf einer Studie mit Kindern in den Niederlanden, deren Übertragbarkeit auf Mexiko fragwürdig ist. Zudem werden in der niederländischen Studie nur Kinder im Alter von vier bis elf Jahren untersucht (de Ruyter et al., 2012), Basto-Abreu et al. (2019) wenden die Ergebnisse aber auf Kinder und (pubertierende) Jugendliche im Alter von zwei bis 19 Jahren an.

Sowohl die von Basto-Abreu et al. (2019) verwendeten Kaufdaten als auch die Umrechnungsfaktoren von Konsumänderungen in Gewichtsveränderungen wurden, wie auch von Thiboonboon et al. (2024) kritisiert[74], nicht-experimentell erhoben, obgleich Daten aus randomisierten kontrollierten Studien – ihre Anwendbarkeit vorausgesetzt – bekanntermaßen zuverlässiger sind und damit zu weniger Unsicherheit in den Schätzwerten führen. Die Grenzen nicht-experimenteller Studien liegen vor allem in ihrer begrenzten kausalen Aussagekraft,

[74] Die Studie von Basto-Abreu et al. (2019) stellt eine der 14 Studien dar, anhand derer im systematischen Review von Thiboonboon et al. (2024) methodologische Herausforderungen in der Untersuchung der ökonomischen Auswirkungen von Steuern auf zuckergesüßte Getränke identifiziert werden.

da Effekte erst ex post konstatiert werden (Doering & Bortz, 2016)

Die Studie prognostiziert die Auswirkungen von BMI-Veränderungen auf gesundheitsbezogene Aspekte anhand relativer Risiken aus epidemiologischen Studien. Fehlen Daten für bestimmte Altersgruppen, erfolgt eine Extrapolation, die zu Verzerrungen führen kann, da Annahmen zur Krankheitsrate und Mortalität möglicherweise nicht die tatsächlichen Trends widerspiegeln. Außerdem wird angenommen, dass Krankheitsraten und Mortalitäten über den gesamten Prognosezeitraum hinweg konstant bleiben, wodurch mögliche medizinische Fortschritte sowie Epidemien oder Pandemien unberücksichtigt bleiben.

Die Behandlungskosten werden auf Basis nationaler Daten aus Mexiko, insbesondere von Gesundheitsministerien, berechnet. Fehlen solche Daten, werden US-amerikanische Behandlungskosten herangezogen und an die Unterschiede in den medizinischen Dienstleistungskosten zwischen den Ländern angepasst. Trotz Kalibrierung spiegeln die US-Daten möglicherweise nicht vollständig die Besonderheiten des mexikanischen Gesundheitssystems wider.

Modellspezifikation

Die Studie wendet als Modellierungsverfahren eine Kohortensimulation (siehe Anhang A.1.1) an. Deren Hauptbeschränkung besteht (wie bei den meisten nicht-experimentellen Studien) darin, dass keine kausalen Zusammenhänge nachgewiesen werden können. Die in der Studie berechneten gesundheitsbezogenen und/oder gesundheitsökonomischen Effekte lassen sich daher nicht eindeutig auf die Einführung der Steuer zurückführen (selbst wenn dies plausibel erscheinen mag), da Effekte durch zeitgleiche weitere Veränderungen oder Entwicklungen vorliegen könnten.

Die Studie operationalisiert Konsumveränderungen von zuckergesüßten Getränken anhand von Änderungen im Kaufverhalten dieser Getränke, aber das Kaufverhalten kann nicht uneingeschränkt mit dem tatsächlichen Konsum gleichgesetzt werden, da die erworbenen Getränke nicht zwangsläufig vollständig konsumiert werden müssen. Die Verwendung der Daten zur Veränderung des Kaufverhaltens[75] impliziert außerdem die Annahme, dass die Steuerwirkung zwei Jahre nach ihrer Einführung vollständig eintritt und über den gesamten Prognosezeitraum von zehn Jahren konstant bleibt, ohne dass hierfür ausreichende empirische Belege vorliegen. So vernachlässigt diese Annahme potenzielle kurzfristige Schwankungen, die lediglich initiale Reaktionen auf die Steuer abbilden und die langfristige Wirksamkeit (z. B. Anpassungsverhalten der Verbraucher) nicht widerspiegeln. Zudem werden in den Berechnungen Faktoren wie saisonale Schwankungen oder zeitlich begrenzte Aufklärungskampagnen, die den Konsum in diesen zwei Jahren beeinflusst haben könnten, nicht berücksichtigt.

Die Prognose der Entwicklung der Kohorte im Hinblick auf BMI und gesundheitsbezogene Faktoren wird mithilfe probabilistischer Sensitivitätsanalysen (siehe Anhang A.3.2) durchgeführt[76]. Die Berechnungen basierten auf festgelegten Verteilungen für die Schlüsselparameter des Modells. Details zur Auswahl und zur Verteilung der Parameter sind nicht dokumentiert. Unklar bleibt auch, wie Variationen in den Parametern die Ergebnisse beeinflussen. Eine ausführlichere Darstellung der Parametervariation und der Iterationsergebnisse würde zur Transparenz der Modellierung und zur Beurteilung der Unsicherheiten beitragen.

[75] Colchero, Rivera-Dommarco et al. (2017) simulieren einen Rückgang des Kaufs zuckergesüßter Getränke um 5,5 % im Jahr 2014 und 9,7 % im Jahr 2015. Basto-Abreu et al. (2019) folgern auf dieser Basis einen durchschnittlichen (konstanten) Rückgang von 7,6 % über den gesamten Untersuchungszeitraum von zehn Jahren hinweg.

[76] Es werden 10.000 Iterationen durchgeführt, um Änderungen im BMI zu prognostizieren, und 100.000 Iterationen, um gesundheitsbezogene Änderungen abzuschätzen.

Die Modellspezifikation berücksichtigt nicht alle relevanten Einflussfaktoren. So ignorieren Basto-Abreu et al. (2019) bspw. Substitutionseffekte: Neben dem Rückgang des Konsums steuerpflichtiger Getränke wurde in Mexiko auch ein Anstieg des Konsums nicht steuerpflichtiger Getränke beobachtet, die ebenfalls Zucker enthalten können[77] (Aguilar et al., 2021; Colchero, Rivera-Dommarco et al., 2017). Dies deutet darauf hin, dass steuerpflichtige Getränke teilweise durch andere kalorienhaltige Getränke substituiert wurden, sodass die Gesamtmenge der Kalorienaufnahme durch Getränke möglicherweise weniger stark zurückging, als von den Studienautoren simuliert. Der Effekt der Steuer könnte daher überschätzt werden. Darüber hinaus werden Wechselwirkungen mit parallelen Reformen, wie der Besteuerung kalorienreicher fester Lebensmittel in Mexiko, nicht berücksichtigt. Eine Studie, welche die Auswirkungen der Einführung beider Steuern (auf Getränke und Lebensmittel) in Mexiko untersuchte, zeigte, dass es eine erhebliche Substitution zwischen besteuerten und unbesteuerten Kategorien gab und dass insgesamt kein statistisch signifikanter Rückgang der eingekauften Gesamtmenge an Kalorien festgestellt werden konnte (Aguilar et al., 2021), was die von Basto-Abreu et al. (2019) getätigten Annahmen ebenso wie die daraus resultierenden Ergebnisse grundsätzlich in Frage stellt. Auch die Inflation, welche langfristige Kosten und die wirtschaftliche Effizienz erheblich beeinflusst, wird offenbar vernachlässigt. Ebenso wird der sozioökonomische Status nicht einbezogen. Dieser hätte es ermöglicht, die Auswirkungen der Steuer auf verschiedene Bevölkerungsgruppen zu analysieren und Unterschiede in der Preissensitivität, dem Zugang zur Gesundheitsversorgung und dem Gesundheitsbewusstsein zu berücksichtigen.

Ergebniskommunikation

Die geschätzten Rückgänge der Adipositasprävalenz sowie der gesundheitsbezogenen Effekte werden durchwegs in absoluten Fallzahlen angegeben. Im Fall der Adipositasprävalenz werden die Effekte zusätzlich in Prozentpunkten ausgedrückt, was eine bessere Einordnung ermöglicht. Für verhinderte Krankheitsfälle oder verbesserte Lebensjahre fehlt jedoch eine vergleichbare Darstellung. Angaben in natürlichen oder relativen Häufigkeiten hätten die Verständlichkeit erhöht. Ebenso fehlt eine Kontextualisierung der eingesparten Gesundheitskosten, deren Höhe für den zehnjährigen Untersuchungszeitraum angegeben wird, jedoch ohne ergänzende Informationen schwer einzuordnen ist. Alle Berechnungen werden mit Konfidenz- oder Unsicherheitsintervallen[78] ergänzt, was die Variabilität transparent darlegt.

Die Validierung der Schätzwerte zum Rückgang der Adipositasprävalenz bei Erwachsenen ergibt Diskrepanzen im Vergleich zu anderen Studien. Die Studienautoren führen diese Abweichungen darauf zurück, dass andere Ansätze keine kalorische Kompensation berücksichtigten. Die getätigten Annahmen von Basto-Abreu et al. (2019) sind daher als realitätsnäher einzuschätzen als andere Ansätze. Der geschätzte Rückgang der Diabetesprävalenz deckt sich aber mit Ergebnissen aus anderen Studien. Unterschiede zeigen sich bei den geschätzten Kosteneinsparungen: Basto-Abreu et al. (2019) geben diesbezüglich an, dass ihr Modell konservative Ergebnisse produziere und die Abweichungen zu anderen Studienergebnissen sich vor allem durch unterschiedliche Annahmen zu den Basiskosten erklären ließen. Da die Unterschiede erheblich sind, bleibt fraglich, ob und in welcher Höhe Kosten tatsächlich eingespart werden, zumal die Datengrundlage fragwürdig ist.

Zur Prognose der potenziellen Einsparungen bei den Gesundheitskosten wird eine Sensitivitätsanalyse durchge-

[77] z. B. 100 % Fruchtsäfte

[78] Das Unsicherheitsintervall ist ein Bereich möglicher Werte einer Größe, der alle Arten von Unsicherheiten berücksichtigt, einschließlich zufälliger und systematischer Fehler sowie subjektiver Einschätzungen.

führt, um den Einfluss des Diskontierungssatzes[79] auf die Ergebnisse abzuschätzen. Die Durchführung solcher Sensitivitätsanalysen ist grundsätzlich sinnvoll, die Modellierungen in Basto-Abreu et al. (2019) basieren allerdings auf potenziell verzerrten Verkaufsdaten zu zuckergesüßten Getränken. Daher ist anzuzweifeln, ob eine solche Sensitivitätsanalyse im letzten Modellierungsschritt –und die Modellierungsergebnisse ganz generell – aussagekräftige Ergebnisse liefern.

Die Studie benennt transparent einige methodische Einschränkungen, wie die fragwürdige Datenqualität bezüglich des Basiskonsums, die den Studienautoren zufolge tendenziell zu konservativen Schätzungen der Steuerwirkung führe. Zudem werden Probleme in der Datenverfügbarkeit zur Konsumveränderung sowie die vereinfachende Annahme einer pauschalen statt heterogenen BMI-Veränderung hervorgehoben. Die Rolle möglicher weiterer Einflussfaktoren (z. B. Substitution, parallele politische Maßnahmen), die einen Teil des Effekts erklären oder die Ergebnisse relativieren könnten, wird jedoch nicht ausreichend diskutiert.

Die Studienautoren schließen aus den Simulationsergebnissen, dass die Steuereinführung zu einer Reduktion der Fälle von Adipositas führe, wodurch Krankheiten verhindert, die Lebensqualität verbessert und Gesundheitskosten eingespart werden könnten. Diese Schlussfolgerung wirkt aber aus zwei Gründen stark generalisiert: (1) Die erheblichen Studienlimitationen, insbesondere in Bezug auf die Datengrundlage und die vereinfachenden Modellannahmen, mindern die Verlässlichkeit der Aussage und sollten zwingend transparent kommuniziert werden. (2) Obwohl die Zusammenhänge plausibel erscheinen, ist das verwendete Modell nicht in der Lage, kausal nachzuweisen, dass die beobachteten Effekte tatsächlich eindeutig auf die Steuer zurückzuführen sind.

Darüber hinaus argumentieren die Studienautoren, dass Länder mit ähnlichen Ausgangsbedingungen ebenfalls von der Einführung einer Steuer profitieren können. Da jedoch länderspezifische Unterschiede in den Ausgangsbedingungen vielfältig und komplex sein können (z. B. demografische, soziale und wirtschaftliche Rahmenbedingungen, Konsumverhalten, Preiselastizitäten und Gesundheitssystem), müssen sämtliche Voraussetzungen vor einer Übertragung der Ergebnisse sorgfältig und kritisch geprüft werden. Eine Anwendung auf Deutschland erscheint kaum plausibel, allein aufgrund der abweichenden Bedingungen hinsichtlich des BMI bei Kindern, wie in Abschnitt 4.5.2 erläutert wurde.

[79] Dabei wird ein alternativer Diskontierungssatz von 5 % anstelle der in der Hauptanalyse 3 % herangezogen.

Zusammenfassung der Limitationen der Studie von Basto-Abreu et al. (2019)

Datengrundlage

- **Stichprobe:** Aufgrund der Verwendung nationaler Bevölkerungsdaten und amtlicher Bevölkerungsprojektionen bestehen keine Limitationen bezüglich der Repräsentativität der Stichprobe.
- **Weitere Eingangsdaten:** Daten spiegeln Konsum- und Kaufverhalten nicht angemessen wider (Basiskonsum basierend auf Selbstauskünften und kurzem Befragungszeitraum; städtische Getränkekaufdaten ohne Berücksichtigung der Gastronomie sowie alters- und geschlechtsspezifischer Unterschiede) und verzerrt Ergebnisse; nicht-experimentell erhobene Umrechnungsfaktoren von Konsum- in Gewichtsänderungen basierend auf spezifischen Populationen (Erwachsene: Lehrerinnen; Kinder: niederländische Daten), sodass Übertragbarkeit auf Gesamtpopulationen aufgrund physiologischer bzw. ethnischer Unterschiede fragwürdig ist; Ausgleich fehlender Daten (gesundheitsbezogene Aspekte: Extrapolation; Behandlungskosten: US-Daten) spiegeln mexikanisches Gesundheitssystem nicht adäquat wider und verzerrt Ergebnisse potenziell.

Modellspezifikation

- **Modellierungsverfahren:** Kausalität wird bei Kohortensimulationen ex ante bei der Modellkonstruktion unterstellt, nicht aber durch das Modell selbst nachgewiesen; Verteilungsparameter für probabilistische Senstitivitätsanalysen nicht offengelegt, was Nachvollziehbarkeit einschränkt.
- **Einfluss- und Zielgrößen:** Vereinfachende Annahmen (konstante Steuerwirkung nach zwei Jahren über 10 Jahre hinweg vernachlässigt Schwankungen und Anpassungen, Gleichsetzung von Konsum- mit Kaufverhalten) bilden realen Phänomene und ihre Beziehungen nicht hinreichend präzise ab.
- **Weitere Einflussfaktoren:** Vernachlässigung von Substitutionseffekten und Wechselwirkungen mit parallelen Reformen (obwohl Evidenz auf gleichbleibende eingekaufte Gesamtmenge an Kalorien hindeutet), sodass Verlässlichkeit der Ergebnisse fraglich ist; Nichtberücksichtigung weiterer relevanter Faktoren (z. B. saisonale Schwankungen, Aufklärungskampagnen, Inflation, sozioökonomischer Status), sodass simulierte Effekte nicht zwangsläufig in der dargestellten Form existieren müssen.

Ergebniskommunikation

- **Ergebnisdarstellung:** Fehlende Kontextualisierung der Ergebnisse erschwert deren Einordnung.
- **Modellbewertung:** Diskrepanzen in den Ergebnissen im Vergleich zu anderen Studien (Adipositasprävalenz, Kosteneinsparungen) aufgrund unterschiedlicher Annahmen, wobei tatsächliche Werte fraglich bleiben; aufgrund verzerrter Konsum- und Kaufdaten ist Aussagekraft der Sensitivitätsanalysen im letzten Modellierungsschritt sowie der Ergebnisse generell begrenzt.
- **Ergebniseinordnung:** Inadäquate Diskussion unberücksichtigter Einflussfaktoren; Schlussfolgerungen (Steuer führe zu Reduktion von Adipositas, Krankheiten und gesundheitsbezogenen Kosten; Übertragbarkeit auf Länder mit ähnlichen Bedingungen) ist aufgrund Nichtberücksichtigung von Studienlimitationen zu stark generalisiert.

Gesamtbeurteilung der Studie von Basto-Abreu et al. (2019)

Die Verlässlichkeit der Studienergebnisse von Basto-Abreu et al. (2019) ist insbesondere aufgrund der unzureichenden Datengrundlage, vereinfachender Modellannahmen und der Nichtberücksichtigung relevanter Einflussfaktoren kritisch zu hinterfragen. Die Datengrundlage weist erhebliche Mängel auf: Die Daten zum Basiskonsum erscheinen aufgrund der Erhebungsmethode und der Fragestellung anfällig für Fehler. Die Kaufdaten zuckergesüßter Getränke schließen die Gastronomie aus und vernachlässigen alters- und geschlechtsspezifische Unterschiede – sie sind also verzerrt sowie pauschalisiert und spiegeln daher das reale Kaufverhalten nicht adäquat wider. Zudem basiert die Umrechnung von Konsum- in Gewichtsänderungen auf spezifischen Populationen (Erwachsene: Lehrerinnen; Kinder: niederländische Daten), die aufgrund physiologischer und ethnischer Unterschiede nicht auf die mexikanische Gesamtbevölkerung übertragbar sind. Auf dieser fragilen Grundlage werden gesundheitsbezogene Effekte simuliert, was belastbare Schlussfolgerungen ausschließt. Die stark vereinfachenden Annahmen (z. B. vollständige Steuerwirkung nach zwei Jahren, konstante Steuerwirkung über 10 Jahre hinweg, Gleichsetzung von Konsum- und Kaufverhalten) sowie die Vernachlässigung von Substitutionseffekten und parallelen politischen Maßnahmen untergraben die Aussagekraft der Ergebnisse sowie Schlussfolgerungen weiter. Insgesamt erweist sich die Studie damit keineswegs als zuverlässige Grundlage für politische Entscheidungen.

5 Zusammenfassung und Diskussion

Im vorliegenden Gutachten wurden sechs verschiedene Modellierungsstudien zu den Effekten einer Steuer auf zuckergesüßte Getränke aus statistisch-methodischer Sicht bewertet. Dieses Kapitel fasst die zentralen Limitationen der Einzelevaluationen zusammen, diskutiert sie und leitet daraus übergeordnete Erkenntnisse ab.

5.1 Zusammenfassung der Studienevaluation

Die zentralen Limitationen der sechs evaluierten Studien in Bezug auf Datengrundlage, Modellspezifikation und Ergebniskommunikation werden im Folgenden zusammengefasst. Aus Gründen der besseren Lesbarkeit werden dabei nur die Namen der Studienautoren, jedoch nicht das Erscheinungsjahr der jeweiligen Publikationen aufgeführt. Insbesondere wird dabei grundsätzlich – sofern nicht explizit anders gekennzeichnet – mit der Autorenangabe „Emmert-Fees et al." auf die Studie von Emmert-Fees et al. (2023), die als eine der sechs Studien im Gutachten evaluiert wurde, Bezug genommen und nicht auf Emmert-Fees et al. (2024).

Beurteilung der Datengrundlage

- **Stichprobe:** Die beiden evaluierten Studien, die auf realen Stichproben basieren, weichen strukturell von der interessierenden Gesamtpopulation ab: In der Untersuchung von Rogers, Cummins et al. sind übergewichtige und adipöse Mädchen unterrepräsentiert, wodurch der simulierte Effekt für Sechstklässlerinnen fragwürdig ist; Gračner et al. wiederum untersuchen ausschließlich Jugendliche, die in städtischen Gebieten leben und bei einem bestimmten Unternehmen krankenversichert sind. Diese Gruppe unterscheidet sich in Alter, Gewicht und Geschlechterverteilung vom Durchschnitt der mexikanischen Jugend und ist außerdem mutmaßlich sozioökonomisch begünstigt. Folglich sind die Ergebnisse nur bedingt auf die gesamte jugendliche Bevölkerung Mexikos übertragbar. **Da die Stichproben dieser beiden Studien die interessierende Population nicht angemessen repräsentieren, sind ihre Ergebnisse als nicht verlässlich zu bewerten.** Bei den evaluierten Studien, die auf synthetischen Stichproben beruhen, welche aus aktuellen amtlichen Bevölkerungsstatistiken abgeleitet wurden (Emmert-Fees et al.; Cobiac et al.; Basto-Abreu et al., Schwendicke und Stolpe), ist hingegen eine grundsätzliche strukturelle Übereinstimmung mit der interessierenden Gesamtpopulation anzunehmen. Allerdings ist einschränkend zu berücksichtigen, dass sich die Analyse von Schwendicke und Stolpe aus dem Jahr 2017 auf Bevölkerungsdaten aus dem Jahr 2012 stützt, die aufgrund wesentlicher demografischer Veränderungen seither – etwa durch Migration – von der heutigen deutschen Bevölkerung abweichen können.

  Abgesehen von der Studie von Basto-Abreu et al., die eine Kohorte im Alter von zwei bis 100 Jahren abbildet und damit alle für den Untersuchungskontext wichtigen Altersgruppen umfasst, beschränken sich die übrigen Untersuchungen auf spezifische Altersbereiche (Emmert-Fees et al.: 30 bis 90 Jahre; Schwendicke und Stolpe: 15 bis 79 Jahre; Rogers, Cummins et al.: Vorschulkinder und Sechstklässler; Cobiac et al.: null bis 17 Jahre; Gračner et al.: zehn bis 18 Jahre). Die Schlussfolgerungen besitzen entsprechend allenfalls innerhalb der untersuchten Altersgruppen Gültigkeit und sollten nicht darüber hinaus verallgemeinert werden. Dennoch leiten etwa Rogers, Cummins et al. – obwohl ihre Stichprobe lediglich spezifische Altersstufen von Kindern abdeckt – Effekte für Kinder oder ältere Kinder im Allgemeinen ab.

- **Weitere Eingangsdaten:** Häufig sind die verwendeten Eingangsdaten nicht oder nur teilweise auf den

Anwendungskontext der Studie **übertragbar**. Dies liegt unter anderem daran, dass **Daten aus anderen Ländern** genutzt werden, die aufgrund länderspezifischer Unterschiede nicht zur untersuchten Stichprobe passen (Emmert-Fees et al.: Prävalenzen von koronaren Herzerkrankungen und Schlaganfällen aus Großbritannien; Schwendicke und Stolpe: Preiselastizitäten aus den USA; Basto-Abreu et al.: Umrechnungsfaktoren von Konsum- in Gewichtsänderungen bei Kindern aus den Niederlanden). Teilweise basieren die Daten auf **spezifischen Subpopulationen** und sind daher nicht auf die Gesamtbevölkerung verallgemeinerbar. So verwenden Basto-Abreu et al. pauschal Daten von Lehrerinnen, um Konsum- in Gewichtsänderungen bei Erwachsenen allgemein umzurechnen. Zudem sind Daten oft nicht in ausreichender **Detailtiefe** verfügbar, um die Heterogenität der Bevölkerung differenziert abzubilden. Bspw. nehmen Basto-Abreu et al. pauschalisierte Konsumveränderungen über alle Alters- und Geschlechtsgruppen hinweg an; Rogers, Cummins et al. sowie Cobiac et al. können lediglich grobe regionale Klassifikationen des soziökonomischen Status vornehmen. Darüber hinaus bilden Daten die **Realität** nicht immer adäquat ab. So beschränken sich Daten zum Einkauf zuckergesüßter Getränke in mehreren der evaluierten Studien ausschließlich auf den stationären Einzelhandel, lassen dabei aber gastronomische Vertriebsstätten außer Acht (z. B. bei Cobiac et al., Gračner et al. und Basto-Abreu et al.), was resultierende Effekte verzerrt.

Außerdem sind die Daten teilweise **veraltet**. So stammen die Konsumdaten bei Schwendicke und Stolpe sowie teilweise auch bei Emmert-Fees et al. aus der NVS II aus den Jahren 2005-2007. Die Daten zu Krankheitskosten bei Emmert-Fees et al. reichen teilweise sogar bis 1999 zurück. Zudem basieren die Daten teils auf **kleinen Stichproben**, was ihre Zuverlässigkeit einschränkt. So basiert etwa die Abschätzung der Produktivitätseinbußen aufgrund von Schlaganfällen bei Emmert-Fees et al. auf lediglich 151 Patienten. Zudem basieren die Daten aufgrund begrenzter Verfügbarkeit größtenteils auf **nicht-experimentellen** Erhebungsverfahren, die eine geringere Evidenzbasis als experimentelle Studien aufweisen. Dies betrifft beispielsweise die Umrechnungsfaktoren von Konsum in Gewichtsänderungen bei Emmert-Fees et al. und Basto-Abreu et al. Werden Daten **aus Metastudien** genutzt (z. B. Gesundheitsdaten bei Cobiac et al.), bleibt die Qualität der zugrunde liegenden Primärdaten oft unklar, insbesondere in Bezug auf mögliche Verzerrungen und ihre Passung zum Studienkontext. Daten basierend auf **Selbstauskünften** (z. B. Schwendicke und Stolpe: Körperindizes; Cobiac et al.: Getränkeeinkäufe und Gesundheitsdaten; Basto-Abreu et al.: Basiskonsum) sind fehleranfällig, was Unsicherheiten in den Ergebnissen zur Folge haben kann. Darüber hinaus sind die Daten teilweise **unvollständig**. Bspw. fehlen bei Emmert-Fees et al. Risikofaktoren für koronare Herzkrankheiten und Schlaganfälle. Verfahren zum Ausgleich fehlender Daten (Emmert-Fees et al.: Schätzung; Basto-Abreu et al.: Extrapolation sowie Ergänzung mit Daten aus anderen Studien) gewährleisten nicht immer die Validität. Schließlich sind einige Datensätze **inkompatibel** (z. B. Konsumdaten bei Emmert-Fees et al.; Konsum- und Körperdaten bei Schwendicke und Stolpe), was zu Abweichungen in den Ergebnissen, etwa bei spezifischen Altersgruppen, führen kann.

Die Datensätze, die zur Ableitung von Modellparametern und Annahmen zur Modellstruktur herangezogen werden, weisen insgesamt in allen evaluierten Studien Schwächen auf – teils erheblich: Bspw. werden von Emmert-Fees et al. für die Parametrisierung des Modells Daten aus insgesamt 36 unterschiedlichen Quellen herangezogen, die in Summe zu einem erheblichen Maß an Unsicherheit im Modell führen. Dies legt nahe, dass die Parameter in allen evaluierten Studien (in unterschiedlichem

Ausmaß) verzerrt sein könnten. **Eine Modellierung auf einer derart fragilen Grundlage lässt keine belastbaren Schlussfolgerungen zu.** Die mangelhafte Verfügbarkeit und Qualität der Daten hat entsprechend schwerwiegende Auswirkungen auf die Aussagekraft der Ergebnisse. Wie Emmert-Fees et al. (2024) in einer methodischen Nachfolgestudie zu der im vorliegenden Gutachten evaluierten Studie von Emmert-Fees et al. (2023) eindrücklich zeigten, können Ergebnisse zur Modellierung der Effekte einer Steuer auf zuckergesüßte Getränke stark variieren, je nachdem, welche Daten zugrunde gelegt werden.

Beurteilung der Modellspezifikation

- **Modellierungsverfahren:** Die eingesetzten Modellierungsverfahren (Mikrosimulation bei Emmert-Fees et al.; Monte-Carlo-Simulation bei Schwendicke und Stolpe; Kohortensimulation bei Cobiac et al. und Basto-Abreu et al.; Zeitreihenanalyse bei Rogers, Cummins et al. und Cobiac et al.; multivariate Regressionsanalyse bei Gračner et al.) sind grundsätzlich geeignet, um reale Situationen mit hypothetischen Szenarien zu vergleichen. Allerdings wird Kausalität bei allen Verfahren ex ante bei der Modellkonstruktion unterstellt, jedoch nicht durch das Modell selbst nachgewiesen. **Wurde ein kausaler Zusammenhang nicht bereits im Vorfeld nachgewiesen – ein Aspekt, der in keiner der Studien explizit thematisiert wird –, bietet das Modellierungsverfahren keine Grundlage, um die Steuer *ursächlich* für die simulierten Effekte verantwortlich zu machen.** Analog dazu wird bei Regressionsmodellen die Annahme der Linearität vorausgesetzt, nicht aber durch das Modell gezeigt.

  Bei Zeitreihenanalysen besteht darüber hinaus das (nicht gänzlich auszuschließende) Risiko verzerrter Effektschätzungen, da die Berechnung des simulierten Trends empfindlich gegenüber dem Interventionszeitpunkt ist. Zudem sind 100 Simulationsdurchläufe pro Gruppe bei einer Monte-Carlo-Simulation, wie sie bei Schwendicke und Stolpe erfolgte, für belastbare Ergebnisse nicht ausreichend. Die Nachvollziehbarkeit der Modellierung wird in manchen Studien zudem durch unzureichende Angaben in der Modellspezifikation zusätzlich erschwert. Beispiele sind die unklare Auswahl und Verteilung der Parameter bei Schwendicke und Stolpe und Basto-Abreu et al. sowie fehlende Informationen zu ARIMA-Eigenschaften und nicht überprüfte Voraussetzungen bei Rogers, Cummins et al.

- **Einfluss- und Zielgrößen:** Die Einflussgrößen werden häufig nur unzureichend operationalisiert. Bspw. wird Konsum häufig pauschal mit Kauf gleichgesetzt (Cobiac et al.; Basto-Abreu et al.). Zudem bleibt die Heterogenität des Konsumverhaltens hinsichtlich alters- und geschlechtsspezifischer Unterschiede häufig unberücksichtigt (Emmert-Fees et al.; Schwendicke und Stolpe; Cobiac et al.; Basto-Abreu et al.). **Beide Aspekte führen zu einer starken Pauschalisierung der Ergebnisse.**

  Die Zusammenhänge zwischen Einfluss- und Zielgrößen werden häufig vereinfacht – meist linear – und ohne hinreichende empirische Belege modelliert. Ein Beispiel ist die Annahme eines direkten Zusammenhangs zwischen Konsum- und Gewichtsveränderungen (Schwendicke und Stolpe; Cobiac et al.), der physiologische Unterschiede, körperliche Aktivität und kalorische Kompensation weitgehend ignoriert und damit der multifaktoriellen Entstehung von Übergewicht und Adipositas widerspricht. Alle Studien, die gesundheitsbezogene Effekte modellieren (Emmert-Fees et al.; Schwendicke und Stolpe; Cobiac et al.; Basto-Abreu et al.), treffen Annahmen zur zeitlichen Wirksamkeit der Steuer, lassen dabei jedoch sowohl mögliche Schwankungen als auch langfristige Adaptionseffekte im Konsumverhalten unberücksichtigt. Auch die

beiden weiteren Studien treffen stark vereinfachende Annahmen: Gračner et al. unterstellen ohne empirische Belege einen linearen Zusammenhang zwischen Preisänderungen und Körpermaßen; Rogers, Cummins et al. gehen von einem linear fortgesetzten Trend der Adipositasprävalenz aus. **Die vereinfachenden Annahmen, die von allen evaluierten Studien getroffen wurden, lassen darauf schließen, dass die Modellergebnisse die realen Phänomene und deren Zusammenhänge nicht ausreichend präzise abbilden.**

- **Weitere Einflussfaktoren:** Die Berücksichtigung sämtlicher relevanter Einflussfaktoren stellt aufgrund der Komplexität der zu modellierenden Realsituation einer Steuereinführung eine erhebliche Herausforderung dar. **Die evaluierten Studien adressieren diese Problematik jedoch nur unzureichend.** Substitutionseffekte und der sozioökonomische Status werden in allen Studien entweder nicht ausreichend oder gar nicht berücksichtigt. Ebenso werden parallele Maßnahmen, wie Aufklärungskampagnen oder politische Reformen, durchweg vernachlässigt. Ein Beispiel hierfür ist die Studie von Basto-Abreu et al., welche die Effekte einer Steuer auf zuckergesüßte Getränke auf Körpergewicht, gesundheitliche und gesundheitsökonomische Aspekte in Mexiko untersucht. Die ermittelten Effekte basieren auf der Annahme eines Rückgangs des Konsums zuckergesüßter Getränke infolge der Steuer. Die Verlässlichkeit dieser Ergebnisse ist allerdings fraglich, da andere Untersuchungen in Mexiko zeigen, dass unter Berücksichtigung weiterer Lebensmittel keine signifikanten Änderungen des Gesamtkalorien-Einkaufs festzustellen sind. **Die Vernachlässigung wichtiger Einflussfaktoren birgt das Risiko, dass die simulierten Effekte die realen Zusammenhänge nicht korrekt abbilden und daraus stark verallgemeinerte oder fehlerhafte Schlussfolgerungen hinsichtlich der Effekte der Steuer gezogen werden.**

Beurteilung der Ergebniskommunikation

- **Ergebnisdarstellung:** Eine Studie (Schwendicke und Stolpe) verzichtet vollständig auf die Angabe von Konfidenzintervallen. In den anderen Studien **erfassen die Konfidenzintervalle die Variabilität der simulierten Ergebnisse nur unzureichend**, da sie ausschließlich zufällige Fehler abbilden und nicht-zufällige Fehler nicht widerspiegeln. Das Fehlen von p-Werten, Effektstärken (Rogers, Cummins et al.; Cobiac et al.; Gračner et al.) und Angaben zu Trendparametern (Rogers, Cummins et al.) erschwert zudem eine fundierte Bewertung der statistischen Relevanz der Ergebnisse. In den evaluierten Studien fehlt es außerdem teilweise an einer angemessenen Kontextualisierung, insbesondere bei der Einordnung der Effektstärken wie der Reduktion von Krankheitsprävalenzen oder Kosteneinsparungen. Ein Beispiel ist die von Emmert-Fees et al. geschätzte Gesundheitskostenreduktion, die in den drei Szenarien mit etwa drei Milliarden Euro angegeben wird. Dies entspricht jedoch einer relativen Kostenreduktion von lediglich unter 0,04 % – eine Information, die von den Studienautoren nicht kommuniziert wird und im Kontext der Gesamtkosten als gering einzustufen ist. **Durch die alleinige Darstellung der absoluten Werte wird dem Leser ein stärkerer Effekt suggeriert.** Ähnlich tritt diese Problematik auch in den Studien von Schwendicke und Stolpe, Rogers, Cummins et al., Cobiac et al. sowie Basto-Abreu et al. auf. Angaben in natürlichen oder relativen Häufigkeiten hätten die Verständlichkeit deutlich verbessert.

- **Modellbewertung:** Systematische Analysen der (globalen) Unsicherheiten in den Modellierungsergebnissen fehlen weitgehend und beschränken sich meist auf Sensitivitätsanalysen. Diese betrachten in allen Studien

lediglich Variationen einzelner Parameter, ohne zentrale Modellannahmen wie wesentliche Einflussfaktoren oder die Annahme linearer Beziehungen zu variieren. Sie sind daher nur bedingt aussagekräftig und aufgrund der mangelhaften Datengrundlage in allen Studien ohnehin fragwürdig. Cobiac et al. weisen zwar auf eine Empfindlichkeit der Ergebnisse gegenüber minimalen Änderungen in den Eingangsdaten hin, simulieren deren Auswirkungen jedoch nicht. Bei Emmert-Fees et al. fallen in einer Sensitivitätsanalyse (gestaffelte Herstellerabgabe, Zuckerreduktion in zuckergesüßten Getränken von lediglich 10 %) die simulierten Effekte geringer aus als in den drei Hauptszenarien, wodurch die Ergebnisse der Hauptanalyse relativiert werden. Bei Cobiac et al. zeigen Validitätsanalysen zudem Diskrepanzen beim Zuckerkonsum und der Reduktion des Zuckerkonsums, die auf Überschätzung der Effekte hindeuten. Ähnliche Abweichungen zeigen sich bei der Adipositasprävalenz und den Kosteneinsparungen in Gračner et al., wobei die tatsächlichen Werte fraglich bleiben. **Unsicherheiten und mögliche Verzerrungen der berichteten Ergebnisse werden in allen Studien insgesamt nur unzureichend adressiert.**

- **Ergebniseinordnung: Die Einordnung der Modellergebnisse erfolgt häufig ohne umfassende Berücksichtigung der Studienlimitationen und Unsicherheiten in den Ergebnissen, weshalb die Verlässlichkeit der Ergebnisse in allen Studien kritisch hinterfragt werden muss.** Zwar werden Studienlimitationen meist breit thematisiert, jedoch oft nur oberflächlich behandelt und selten in Bezug auf ihre potenziellen Auswirkungen auf die Ergebnisse reflektiert. **Dies führt teilweise zu übergeneralisierten Schlussfolgerungen.** So stellt bspw. Emmert-Fees et al. die gestaffelte Herstellerabgabe als optimale Steuerform dar, lässt dabei jedoch die in einer Sensitivitätsanalyse aufgezeigten Unsicherheiten unberücksichtigt. Rogers, Cummins et al. sowie Gračner et al. weisen Effekte lediglich für bestimmte Subgruppen, etwa Mädchen, aus, verallgemeinern diese jedoch auch auf Jungen. Cobiac et al. werten eine geringfügige Erhöhung der Lebenserwartung – im Bereich von wenigen Tagen bis maximal einem Monat – als signifikante Verbesserung und überinterpretieren damit die tatsächliche Relevanz des Effekts. Besonders problematisch sind fehlerhafte kausale Schlussfolgerungen: Schwendicke und Stolpe, Cobiac et al. sowie Basto-Abreu et al. folgern, dass eine Steuer Adipositas reduzieren würde, obwohl dies nicht durch die verwendeten Modelle belegt wird. Solche irreführenden Schlussfolgerungen sind besonders kritisch, da sie häufig von Medien aufgegriffen und zur Argumentation für regulatorische Maßnahmen herangezogen werden.

5.2 Zentrale Erkenntnisse

Auf Grundlage der bestehenden Forschung zu Modellierungsstudien sowie der in diesem Gutachten erfolgten Evaluationen der Einzelstudien lassen sich folgende zentrale Erkenntnisse ableiten:

1. Modellierungsstudien zur Untersuchung (hypothetischer) regulatorischer Maßnahmen, bspw. die Einführung einer Zuckersteuer in Deutschland, sind kein eigenständiger Beleg kausaler Zusammenhänge; vielmehr wird das Vorliegen von Kausalwirkungen zwischen Einfluss- und Zielgrößen bei den verwendeten Modellierungsverfahren ex ante angenommen.

2. Nicht nur das Vorliegen kausaler Wirkmechanismen, sondern auch deren Quantifizierung basiert auf Annah-

men, bspw. der Übertragbarkeit früherer empirischer Untersuchungen. Solche Annahmen sind zwangsläufig mit Unsicherheiten der Modellierungsergebnisse verbunden. Jede Entscheidung für eine bestimmte Annahme ist gleichzeitig eine Entscheidung gegen alternative Möglichkeiten. Modellierungsergebnisse werden daher maßgeblich von den zugrunde liegenden Annahmen beeinflusst.

3. Die Aussagekraft von Modellierungsstudien ist stark von der Verfügbarkeit und Qualität der benötigten empirischen Eingangsdaten abhängig. Modellierungsstudien beziehen diese in der Regel aus externen Quellen und setzen, häufig ohne explizite Prüfung, deren Repräsentativität und Genauigkeit voraus. Beides ist oftmals nicht gegeben.

4. Ungenaue, verzerrte und/oder nicht repräsentative Daten sowie falsche oder zu stark vereinfachende Annahmen über die zugrunde liegenden Wirkmechanismen führen zu Modellergebnissen, die zwar innerhalb des Modells konsistent sein mögen, aber die Realität nicht korrekt abbilden.

5. Selbst in Fällen, in denen die Annahmen dem Grunde nach zutreffen und die Eingangsdaten repräsentativ sind, bleiben die Ergebnisse von Modellierungsstudien immer mit Unsicherheiten behaftet, die erheblich sein können und bei Fragestellungen der öffentlichen Gesundheit häufig ignoriert werden.

5.3 Diskussion

„Politische Entscheidungen sollten auf einen informierten Diskurs zu Zielen und Maßnahmen gestützt sein. Wissenschaft kann dazu einen Beitrag leisten, indem sie Evidenz über Ursachen und Wirkungen zur Verfügung stellt, bspw. durch methodisch angemessene, systematische Evaluierungen. [...] Um eine informierte Debatte zu gewährleisten, müssen Datengrundlage, Auswertungsmethode und Ergebnisse von Evaluierungen zugänglich und nachvollziehbar sein.“ (Leopoldina, 2021).

Die sorgfältige Abwägung, für welchen Zweck die Ergebnisse von Modellierungsstudien verwendet werden können („fit for purpose“), stellt ein zentrales Element ihrer Beurteilung dar. Um eine unsachgemäße Anwendung zu vermeiden, müssen die Ergebnisse ihrem Zweck entsprechend transparent dargestellt und kommuniziert werden. Dabei ist klar zu differenzieren zwischen Ergebnissen, die der Exploration möglicher Forschungsfragen bzw. dem Erkenntnisgewinn innerhalb einer Disziplin dienen, und solchen, die für politische Entscheidungen, etwa im Rahmen von Gesetzgebungsverfahren, herangezogen werden sollen (Münnich, im Druck). Erstere können geringere Anforderungen an die Qualität der zugrunde liegenden Daten anlegen, sollten diese jedoch entsprechend dokumentieren. Wenn Studienautoren aber den Anspruch erheben, Datengrundlagen für die Ableitung politischer Maßnahmen bereitzustellen, müssen sie sich an strengsten Kriterien hinsichtlich der Datenqualität, der Eignung der Methodik und der Einhaltung wissenschaftsethischer Grundsätze messen lassen.

Angesichts der potenziell weitreichenden Auswirkungen auf Gesellschaft und Wirtschaft ist es entscheidend, dass sich alle beteiligten Akteure – insbesondere politische Entscheidungsträger, finanzielle Förderer und involvierte Forschende – bewusst sind, dass Studien, die als Grundlage für evidenzbasierte Politik dienen, auf allen Ebenen besonders hohen Qualitätsstandards genügen müssen. Dies gilt einerseits für die **Datengrundlagen**, die ein bestmögliches **Abbild realer Phänomene** darstellen müssen; andererseits für die verwendeten **Verfahren zur Analyse** dieser Daten, die geeignet sein müssen, um die **Beziehungen zwischen diesen Phänomenen**

adäquat zu modellieren. Modellierungsstudien sind nach derzeitigem wissenschaftlichem Stand grundsätzlich die bestmöglichen *Verfahren*, um Auswirkungen regulatorischer Maßnahmen zu untersuchen. Sie können theoretisch geeignet sein, eine Evidenzbasis für Regulierung zu schaffen, sofern sie auf belastbaren und sorgfältig validierten Annahmen sowie präzisen Datengrundlagen beruhen. Die Kausalität der modellierten Wirkmechanismen wird allerdings durch Modellierungsstudien selbst nicht nachgewiesen, sondern muss durch entsprechende Studien bereits ex ante belegt sein. Ebenso müssen die Modellparameter mit hinreichender Sicherheit aus vorhandenen Daten schätzbar sein. Die Erstellung einer für die Begründung politischer Maßnahmen geeigneten Modellierungsstudie erfordert also sowohl hochwertige und für den Anwendungskontext relevante Daten als auch realitätsgetreue Annahmen zur Modellierung der Beziehungen zwischen diesen Daten sowie geeignete statistische Verfahren. Nur auf dieser Grundlage können Ergebnisse erzielt werden, die in ihrer Gesamtheit überzeugen. Ungenaue, verzerrte und/oder nicht repräsentative Daten sowie falsche bzw. zu stark vereinfachende Annahmen über die zugrunde liegenden Wirkmechanismen führen zu Modellergebnissen, die zwar innerhalb des Modells konsistent sein mögen, aber die Realität nicht korrekt abbilden.

> Das Gutachten kommt zu dem Schluss, dass keine der evaluierten Modellierungsstudien zu den Effekten einer Zuckersteuer aus statistisch-methodischer Perspektive den hohen Qualitätsstandards genügt, die für evidenzbasierte politische Entscheidungen erforderlich sind.

Ein Kernproblem liegt darin, dass Datengrundlagen von hinreichender Qualität, die eine belastbare Grundlage für die aufgestellten Modelle schaffen würden, nicht vollumfänglich vorhanden sind – damit ist eine unverzichtbare Voraussetzung für evidenzbasierte Politikgestaltung nicht erfüllt. Das Fehlen solcher Daten liegt einerseits an der Komplexität der Wirkmechanismen, die schwer messbaren Einflussfaktoren unterliegen, andererseits an unbeobachteter (oder gar unbeobachtbarer) Heterogenität der Subpopulationen. In Ermangelung ausreichend detaillierter Datensätze müssen zahlreiche Annahmen getroffen werden. Zudem ist die hinreichend präzise Simulation der Bevölkerung auf Mikroebene hoch anspruchsvoll. Mögliche Selektionsfehler in der Datenerhebung, Messfehler durch Selbstauskünfte oder Unsicherheiten durch kleine Stichproben können erhebliche Auswirkungen auf die Modellergebnisse haben. Studienergebnisse aus anderen Ländern sind schwer auf Deutschland zu übertragen, da immer strukturelle Unterschiede zwischen der in der Studie untersuchten Stichprobe eines anderen Landes und der deutschen Gesamtbevölkerung bestehen.

Einige der im Gutachten aufgezeigten Limitationen sind kaum oder gar nicht zu überwinden. Daher soll die vorliegende Analyse nicht die wissenschaftliche Leistung der Autoren der evaluierten Studien schmälern. Die Kritik richtet sich vielmehr an die Annahme, dass solche Studien bedenkenlos als verlässliche Grundlage für regulatorische Entscheidungen herangezogen werden können. Dieses Gutachten versteht sich daher als Appell, die identifizierten Limitationen in der öffentlichen Debatte und in politischen Entscheidungsprozessen angemessen zu berücksichtigen. Dazu gehört insbesondere, Unsicherheiten in Modellierungsstudien klar zu quantifizieren und transparent zu kommunizieren[80].

[80] Zu einem ähnlichen Schluss kommen auch Emmert-Fees et al. (2024, S. 2) in ihrer methodenfokussierten Studie: „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.“ Die Erkenntnisse dieser Studie stellen eine unverzichtbare Ergänzung zur Modellierungsstudie von Emmert-Fees et al. (2023) dar und sollten daher in der Debatte angemessen berücksichtigt werden.

Bibliografie

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

Appelhans, Y. (2024). So sinnvoll wäre eine Zuckersteuer. Verfügbar 7. Oktober 2024 unter 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* (17. Aufl.). Springer. https://doi.org/10.1007/978-3-658-40465-9

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. Verfügbar 12. Dezember 2024 unter 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), Artikel 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., Freeman, J. V., & Preece, M. A. (1995). Body mass index reference curves for the UK, 1990. *Archives of Disease in Childhood*, *73*(1), 25–29. https://doi.org/10.1136/adc.73.1.25

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*, Artikel 112361. https://doi.org/10.1016/j.socscimed.2019.112361

de Ruyter, J. C., Olthof, 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* (5. Aufl.). Springer. https://doi.org/10.1007/978-3-642-41089-5

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), Artikel 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), Artikel 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), Artikel 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

Europäische 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. Verfügbar 11. Dezember 2024 unter 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. https://doi.org/10.1007/978-3-658-37985-8_36

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* (2. Aufl.). Sage. https://doi.org/10.4135/9781412983662

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. Verfügbar 16. August 2024 unter https://www.zdf.de/nachrichten/ratgeber/zuckersteuer-fruehstuecksrichtlinie-massnahmen-100.html

Helmholtz Zentrum München. (2024). KORA: Kooperative Gesundheitsforschung in der Region Augsburg. Verfügbar 5. Dezember 2024 unter 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), Artikel 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* (6. Aufl.). 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*, Artikel 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*, Artikel 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? Verfügbar 16. Mai 2024 unter 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* (3. Aufl.). Chapman; Hall/CRC. https://doi.org/10.1201/9780429298899

Magoley, N. (2024). Zuckersteuer auf Softdrinks: Was sie bringen würde. Verfügbar 7. Oktober 2024 unter 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), Artikel 800–4. https://doi.org/10.3238/arztebl.m2024.0150

Max Rubner-Institut. (n. d.). Die Nationale Verzehrsstudie II. Verfügbar 5. Dezember 2024 unter 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* (7., vollständig überarbeitete und aktualisierte Auflage). Springer. https://doi.org/10.1007/978-3-662-68224-1

Münnich, R. (n. d.). Evidence-based policies and data quality – What is missing? [im Druck]. 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. Verfügbar 1. Oktober 2024 unter 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. Verfügbar 1. Oktober 2024 unter 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. Verfügbar 25. September 2024 unter 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. Verfügbar 12. September 2024 unter 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. Verfügbar 7. Oktober 2024 unter https://www.monitor-versorgungsforschung.de/abstract/zuckersteuer-plausibilitaet-und-wertigkeit-der-studien/

Office for Health Improvement and Disparities. (2024). National Child Measurement Programme. Verfügbar 12. September 2024 unter 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, Supplement), 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. Verfügbar 5. Dezember 2024 unter 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), Artikel 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), Artikel 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. https://doi.org/10.1007/978-3-658-34063-6

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. Verfügbar 12. August 2024 unter 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*, Artikel 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. Verfügbar 17. September 2024 unter 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). Verfügbar 5. Dezember 2024 unter https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/Glossar/fortschreibungdes-bevoelkerungsstandes.html

Statistisches Bundesamt. (n. d. b). Kohorte. Verfügbar 29. Oktober 2024 unter https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Geburten/Glossar/kohorte.html

Statistisches Bundesamt. (2024). Gesundheitsausgaben in Deutschland. Verfügbar 13. August 2024 unter 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. Verfügbar 23. Oktober 2024 unter 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*, Artikel 105076. https://doi.org/10.1016/j.healthpol.2024.105076

United Nations. (1954). *Handbook of statistical organization* (1. Aufl.). Verfügbar 16. Mai 2024 unter https://unstats.un.org/unsd/publication/SeriesF/SeriesF_6E.pdf

Verified Metrics. (n. d.). Cohort modeling. Verfügbar 9. August 2024 unter 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), Artikel 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

Abkürzungsverzeichnis

A

ARIMA ..... Autoregressive Integrated Moving Average

B

BMI ..... Body Mass Index

C

CI ..... Confidence Interval

I

IMSS ..... Instituto Mexicano del Seguro Social

K

kcal ..... Kilokalorien

KHK ..... Koronare Herzkrankheiten

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

Glossar

A

Ad-Valorem-Verbrauchssteuer

Eine Ad-Valorem-Verbrauchsteuer ist eine Wertsteuer und wird als Prozentsatz des Wertes einer Ware erhoben.

B

BMI (body mass index)

Der BMI ist eine weit verbreitete Kennzahl zur Beurteilung des Körpergewichts in Relation zur Körpergröße. Er wird berechnet, indem das Körpergewicht in Kilogramm durch das Quadrat der Körpergröße in Metern dividiert wird. Der BMI dient als Grundlage zur Einordnung, ob eine Person als untergewichtig, normalgewichtig, übergewichtig oder adipös gilt.

G

Gestaffelte Steuer

Eine gestaffelte Steuer ist eine Form der Besteuerung, bei der innerhalb einer definierten Produktkategorie differenzierte Steuersätze angewendet werden, die an spezifische Produktmerkmale gekoppelt sind. Ein Beispiel hierfür ist die Variation der Steuersätze in Abhängigkeit vom Zuckergehalt bei zuckergesüßten Getränken.

Grundgesamtheit (Population)

Die Grundgesamtheit bezeichnet die Menge aller Entitäten (z. B. Personen), über die eine Aussage getroffen werden soll (z. B. die gesamte in Deutschland lebende Bevölkerung). Vollständige Daten über die Grundgesamtheit stehen in der Praxis selten zur Verfügung. In der Regel erhebt man Daten über eine (deutlich kleinere) *Stichprobe*, um Rückschlüsse auf die Eigenschaften der Grundgesamtheit zu tätigen.

K

Kausalität

Kausalität beschreibt die Beziehung von Ursache und Wirkung, d. h. eine Änderung einer Variable A hat die Änderung einer anderen Variable B zur Folge. Kausalität impliziert eine *Korrelation* dieser Variablen. Umgekehrt ist eine Korrelation zweier Variablen jedoch nicht hinreichend, um auf Kausalität zu schließen. Es ist bspw. möglich, dass zwei Variablen keine direkte wechselseitige Beeinflussung aufweisen, sondern beide von einer dritten Variable C kausal abhängen, wodurch eine beobachtete (Schein-)Korrelation entsteht.

kcal (Kilokalorie)

Kilokalorie (kcal) ist eine (veraltete) Einheit für die Energiemenge, die jedoch für Lebensmittel weiterhin Anwendung findet. Umgangssprachlich wird die kcal auch oft schlicht als Kalorie (ohne Kilo) bezeichnet.

Konfidenzintervall

Ein Konfidenzintervall (englisch: confidence interval; kurz: *CI*) entspricht einem zufälligen Intervall, welches bei bekannter Verteilung der Stichprobe eine vorgegebene Überdeckungswahrscheinlichkeit für den wahren, aber unbekannten Schätzwert besitzt. Das bedeutet, dass bei einem vorgegebenen Konfidenzniveau von z. B. 95 %, welches meist standardmäßig gewählt wird, 95 % der durch Stichproben ermittelten Konfidenzintervalle den wahren Wert überdecken würden.

Korrelation

Zwei Variablen A und B werden als korreliert bezeichnet, wenn ein statistischer Zusammenhang zwischen ihnen besteht. Wenn hohe Werte von A mit hohen Werten von B einhergehen, spricht man von positiver Korrelation. Umgekehrt liegt eine nega-

tive Korrelation vor, wenn hohe Werte von A tendenziell mit niedrigen Werten von B einhergehen. Ein Beispiel für positiv korrelierte Variablen sind Körpergröße und -gewicht beim Menschen: Größere Menschen sind tendenziell schwerer als kleine. Es ist jedoch essenziell zu beachten, dass Korrelation keine *Kausalität* impliziert. So besteht bspw. eine positive Korrelation zwischen den Variablen „Absatz von Speiseeis“ und „Auftreten von Sonnenbrand“, ohne dass zwischen ihnen eine kausale Beziehung vorliegt.

Kreuzpreiselastizität

Die Kreuzpreiselastizität (der Nachfrage) quantifiziert die prozentuale Änderung der nachgefragten Menge eines Gutes als Reaktion auf eine prozentuale Preisänderung eines anderen Gutes. Eine positive Kreuzpreiselastizität liegt vor, wenn eine Preissteigerung von Gut X zu einem Anstieg der Nachfrage nach Gut Y führt, was darauf hindeutet, dass Verbraucher das eine Gut mit dem anderen substituieren (z. B. Substitution zuckergesüßter Getränke durch Fruchtsäfte). Umgekehrt liegt eine negative Kreuzpreiselastizität vor, wenn eine Preissteigerung von Gut X mit einer sinkenden Nachfrage nach Gut Y einhergeht, was darauf hinweist, dass Verbraucher X mit Y komplementieren – wenn das eine Gut teurer wird (z. B. Grillgut), kaufen die Konsumenten weniger davon, was automatisch auch dazu führt, dass sie weniger vom anderen Gut (z. B. Grillkohle) nachfragen.

L

Lineare Regression

Die lineare Regression ist ein statistisches Verfahren zur Modellierung und Vorhersage der Werte einer Variablen Y anhand einer oder mehrerer anderer Variablen $X_1, X_2, ...$ Dabei wird ein linearer Zusammenhang zwischen Y und den X_i angenommen, der durch die Gleichung

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

beschrieben wird. Die Koeffizienten $b_0, b_1, b_2, ...,$ werden so geschätzt, dass die Summe der quadrierten Fehlerterme ε, die die Abweichungen zwischen den beobachteten und den vorhergesagten Werten darstellen, über alle Datenpunkte minimiert wird. Die Variable Y wird als *abhängige Variable* oder *Zielgröße* bezeichnet, während $X_1, X_2, ...$ als *unabhängige Variablen* oder *Prädiktoren* bezeichnet werden. Enthält das Modell nur einen Prädiktor, spricht man von einer *einfachen* linearen Regression. Werden hingegen mehrere Prädiktoren einbezogen, handelt es sich um eine *multiple* lineare Regression. Wenn darüber hinaus mehrere Zielgrößen $Y_1, Y_2, ...$ gemeinsam modelliert werden - dadurch wird obige Gleichung zu einem Gleichungssystem - spricht man von einer *multivariaten* linearen Regression.

M

Metastudie / Metaanalyse)

Eine Metastudie oder Metaanalyse ist eine systematische Zusammenfassung und Auswertung von Ergebnissen aus Primärstudien (Einzeluntersuchungen). Ziel ist es, frühere Forschungsarbeiten quantitativ zu integrieren und die gewonnenen Erkenntnisse auf Grundlage statistischer Analysen konsolidiert darzustellen.

P

Preiselastizität

Die Preiselastizität (der Nachfrage) quantifiziert die prozentuale Änderung der nachgefragten Menge eines Gutes als Reaktion auf eine Preisänderung dieses Gutes.

Q

QALY (quality-adjusted life year)

Das QALY ist ein Maß zur Bewertung eines Lebensjahres unter Berücksichtigung von Gesundheit und Wohlbefinden. Dabei wird einem Lebensjahr in voller Gesundheit oder mit maximalem Wohlbefinden ein Wert von 1 zugewiesen, während ein Zustand ohne Gesundheit oder Wohlbefinden, wie der Tod, den Wert 0 zugewiesen bekommt. Für Gesundheitszustände mit Einschränkungen liegt der Wert zwischen 0 und 1, wobei die genaue Bewertung von der Schwere der Beeinträchtigung abhängt. Dieses Konzept ermöglicht es, Lebensjahre mit höherer Lebensqualität, bspw. im Kontext der Evaluation medizinischer Interventionen, stärker zu gewichten als solche mit geringerer Lebensqualität. Eine einheitliche Definition zur Zuordnung von QALY-Werten zu spezifischen Gesundheitszuständen existiert jedoch nicht. Stattdessen basieren die Bewertungen meist auf Befragungen, die das subjektive Wohlbefinden der Betroffenen erfassen. Diese Daten werden in der Regel in Modelle integriert, um jedem Gesundheitszustand einen entsprechenden QALY-Wert zuzuweisen.

S

SDIL (Soft Drinks Industry Levy)

Die SDIL ist eine im März 2016 angekündigte und im April 2018 im vereinigten Königreich eingeführte Abgabe auf *zuckergesüßte Getränke*. Hersteller unterliegen im Rahmen einer gestaffelten Verbrauchsteuer unterschiedlich hohen Abgaben auf abgepackte, zuckergesüßte Getränke, abhängig vom Zuckergehalt: Für Getränke mit weniger als 5 g Zucker pro 100 ml Flüssigkeit fällt keine Steuer an, bei einem Zuckergehalt zwischen 5 und 8 g je 100 ml beträgt die Abgabe 0.18£ pro Liter und ab einem Zuckergehalt von 8 g je 100 ml beträgt sie 0.24£.

Signifikanzniveau

Das Signifikanzniveau bezeichnet die im Voraus festgelegte maximal zulässige *Irrtumswahrscheinlichkeit* eines statistischen Tests, also die Wahrscheinlichkeit, mit der der Test fälschlicherweise ein signifikantes Ergebnis anzeigt.

Spezifische Verbrauchsteuer

Eine spezifische Verbrauchsteuer wird als fester Geldbetrag in Abhängigkeit von einem bestimmten physikalischen Merkmal eines Produkts, wie etwa dessen Menge oder Zuckergehalt, erhoben.

Statistische Signifikanz

Ein Ereignis, wie bspw. eine Reduktion des Konsums zuckergesüßter Getränke nach der Einführung einer Steuer, wird als statistisch signifikant bezeichnet, wenn die Wahrscheinlichkeit, dass das Ereignis lediglich zufällig auftritt, gering ist. In der Praxis wird die Signifikanz häufig durch den p-Wert quantifiziert, der angibt, wie groß die Irrtumswahrscheinlichkeit ist, die beobachteten Daten zu erhalten, obwohl der Effekt in der Realität nicht vorhanden ist. Liegt der p-Wert unterhalb des festgelegten *Signifikanzniveaus*, gilt das Ergebnis als statistisch signifikant.

Steuerweitergaberate

Die Steuerweitergaberate gibt an, welcher Anteil einer Steuer auf den Endpreis für Verbraucher übertragen wird.

Stichprobe

Eine Stichprobe ist eine Teilmenge der *Grundgesamtheit*, die nach festgelegten Kriterien ausgewählt wurde. Das Ziel quantitativer wissenschaftlicher Studien ist es in der Regel, Erkenntnisse über die gesamte Population zu gewinnen, aus der die Stichprobe entnommen wurde.

Substitution
Substitution (eines Produkts) bezeichnet den Austausch eines Produkts durch ein anderes Produkt.

U

Unsicherheitsintervall
Das Unsicherheitsintervall ist ein Bereich möglicher Werte einer Größe, der alle Arten von Unsicherheiten berücksichtigt, einschließlich zufälliger und systematischer Fehler sowie subjektiver Einschätzungen. Es unterscheidet sich vom Konfidenzintervall dadurch, dass es nicht nur die statistische Unsicherheit im Zusammenhang mit der Zufälligkeit der Stichprobe widerspiegelt, sondern auch systematische und subjektive Faktoren einbezieht.

Z

Zuckergesüßte Getränke
Zuckergesüßte Getränke (engl. *sugar-sweetened beverages*; kurz: *SSBs*) sind Getränke, denen Zucker (z. B. Saccharose, Fructose, Glukose, Maissirup, Honig, Malzsirup, etc.) zugesetzt wurde.

A Modellierungsverfahren

In Modellierungsstudien kommen – abhängig von der spezifischen Forschungsfrage und dem Zweck der Untersuchung – eine Vielzahl unterschiedlicher methodischer Verfahren zum Einsatz (Jahn et al., 2022). Modellierungsverfahren, die häufig als Entscheidungsgrundlage im Gesundheitswesen dienen, werden bspw. von Jahn et al. (2022) mitsamt ihrer Gemeinsamkeiten und Unterschiede detailliert vorgestellt.

Zu den am häufigsten im Gesundheitsbereich verwendeten Modellierungsverfahren zählen sogenannte *State-Transition-Modelle* (Mertens et al., 2022; Siebert et al., 2012). Sie modellieren verschiedene Zustände (*states*, z. B. Gesundheitszustände) und Übergänge (*transitions*) zwischen diesen Zuständen (Jahn et al., 2022). Die Modellierung erfolgt dabei entweder auf Basis von Kohorten oder auf Ebene einzelner Individuen (Siebert et al., 2012). Eine ausführliche Darstellung der State-Transition-Modelle findet sich in Abschnitt A.1. Zur Analyse der Auswirkungen einer Steuer auf zuckergesüßte Getränke haben bisherige Studien auch verschiedene andere analytische Verfahren eingesetzt, um Zusammenhänge zwischen Variablen zu untersuchen. Diese analytischen Verfahren werden expliziter in Abschnitt A.2 beleuchtet. In Modellierungsstudien werden zudem häufig weitere unterstützende Verfahren eingesetzt, etwa zur Validierung oder Bewertung der Ergebnisse. Eine Erläuterung solcher Begleitverfahren findet sich in Abschnitt A.3.

A.1 State-Transition-Modelle

Die State-Transition-Modellierung ist ein intuitiver, flexibler und transparenter Ansatz der computergestützten entscheidungsanalytischen Modellierung und umfasst verschiedene Arten von Simulationen (Siebert et al., 2012). Eine Simulation stellt dabei ein bekanntes oder unbekanntes Szenario nach, um dieses besser zu verstehen. So können bspw. die Auswirkungen des Tsunami 2004 in Südostasien (bekanntes Szenario) oder des Klimawandels in der Zukunft (unbekanntes Szenario) untersucht werden (Bungartz et al., 2013).

State-Transition-Modelle analysieren Populationen entweder auf Basis von Kohorten – also Personengruppen, die ein bestimmtes Ereignis zur gleichen Zeit erfahren haben (z. B. gleiches Geburtsjahr; Statistisches Bundesamt, n. d. b) – oder auf Ebene einzelner Individuen, was als *Mikrosimulation* bezeichnet wird (Siebert et al., 2012). Werden in Mikrosimulationsmodellen zusätzlich Interaktionen zwischen Individuen einbezogen, spricht man von *agentenbasierten Modellen* (Arnold et al., 2019). Kohorten- und Mikrosimulationen berücksichtigen keine solchen Interaktionen zwischen Individuen oder Gruppen (Siebert et al., 2012), sondern fokussieren sich stattdessen mehr auf die Entwicklung der Population – was auch im Kontext der Untersuchung der Auswirkungen von Steuern von primärer Relevanz ist. Kohorten- und Mikrosimulationen werden nachfolgend detaillierter vorgestellt.

A.1.1 Kohortensimulation

Bei einer Kohortensimulation werden virtuelle Kohorten von Individuen erstellt, welche die Eigenschaften und Verhaltensweisen einer realen Population nachbilden (Iskandar, 2018). Eine Kohorte bezeichnet dabei eine geschlossene Gruppe von Individuen, die (während eines bestimmten Zeitraums) ein gemeinsames Merkmal teilen, bspw. das Geburtsjahr oder eine spezifische Krankheit (Glenn, 2005). Diese Kohorten werden repräsentativ für die zu untersuchende Gesamtbevölkerung ausgewählt bzw. zusammengestellt, sodass entsprechend Rückschlüsse auf die Gesamtbevölkerung möglich sind (Ethgen & Standaert, 2012). In der Simulation werden diese Kohorten

über einen definierten Zeitraum hinweg analysiert, indem verschiedene Zustände (z. B. Gesundheitszustände wie Krankheitsbeginn, Krankheitsverlauf und Tod) nachgestellt werden (Ethgen & Standaert, 2012; Iskandar, 2018). Um die Wahrscheinlichkeiten der Übergänge zwischen den verschiedenen Zuständen zu schätzen, wird üblicherweise auf sogenannte *Markovketten* zurückgegriffen (Iskandar, 2018; Siebert et al., 2012).

Das Ergebnis einer Kohortensimulation kann wertvolle Einblicke in die langfristigen Effekte (z. B. Kosten) verschiedener Interventionen geben und damit Forschende und politische Entscheidungsträger unterstützen, die potenziellen Auswirkungen verschiedener Strategien (z. B. Auswirkungen auf die Gesundheit und das Gesundheitssystem) zu bewerten oder die Verteilung von Ressourcen zu steuern (Iskandar, 2018). Kohortenmodelle können jedoch nur Einblicke in Trends auf Ebene der Kohorten geben, nicht aber individuelles Verhalten vorhersagen (Verified Metrics, n. d.). Außerdem berücksichtigen die mithilfe von Markovketten formulierten Übergangswahrscheinlichkeiten keine Ereignisse vor dem aktuellen Zustand (Siebert et al., 2012). Die statistische Aussagekraft der Ergebnisse von Kohortensimulationen kann außerdem von verschiedenen Faktoren beeinflusst werden, insbesondere auch von der Prävalenz des interessierenden Risikofaktors oder der interessierenden Zielgröße (Brown & Jiang, 2010).

A.1.2 Mikrosimulation

Mikrosimulationsmodelle erfassen das Verhalten einzelner Einheiten (z. B. Personen, Haushalte, Unternehmen). Sie bieten im Vergleich zu Makrosimulationsmodellen, die sich auf ganze Populationen konzentrieren, detailliertere Analysemöglichkeiten. Sie kommen zum Einsatz, wenn die Heterogenität innerhalb einer Population oder die Untersuchung individueller Verläufe von Bedeutung ist (Spielauer, 2011; Weymeirsch et al., 2024). Zudem ermöglichen sie, vergangene Ereignisse in Zukunftsprojektionen einzubeziehen (Spielauer, 2011).

Die Detailliertheit, mit der Mikrosimulationsmodelle bewerten können, wie heutige Entscheidungen und Handlungen die Zukunft gestalten, macht sie zu einem attraktiven Instrument für die Politikgestaltung, denn sie ermöglichen die Modellierung politischer Maßnahmen auf beliebig detaillierter Ebene. Allerdings erfordern Mikrosimulationsmodelle aufgrund ihrer Komplexität in der Regel hohe Investitionen in Personal und Hardware (Spielauer, 2011). Die Genauigkeit der Ergebnisse hängt zudem stark von den zugrundeliegenden Annahmen und Regeln im Modell ab, wobei selbst kleinste Änderungen an den Eingabeparametern erhebliche Auswirkungen auf die Ergebnisse haben können (Emmert-Fees et al., 2024). Zu den Limitationen von Mikrosimulationsmodellen gehört außerdem die Verfügbarkeit von Daten (Emmert-Fees et al., 2024; Weymeirsch et al., 2024). Um genaue und zuverlässige Ergebnisse zu erzielen, benötigen Mikrosimulationsmodelle umfangreiche und detaillierte Daten auf individueller Ebene. Der Zugang zu sensiblen Mikrodaten ist aber häufig stark eingeschränkt, und es müssen strenge Datenschutzrichtlinien eingehalten werden (Weymeirsch et al., 2024).

A.2 Analytische Verfahren

Zur Analyse der Auswirkungen einer Steuer auf zuckergesüßte Getränke haben bisherige Studien auch verschiedene andere analytische Verfahren eingesetzt, um Zusammenhänge zwischen Variablen zu untersuchen. Sie werden im Folgenden erläutert.

A.2.1 Zeitreihenanalyse

Eine Zeitreihenanalyse ist eine Analyse, die untersucht, wie sich eine Reihe von Datenpunkten über die Zeit hinweg entwickelt. Dabei sollten in der Analyse Faktoren wie Saisonalität und Trend der Zeitreihe berücksichtigt werden (Fahrmeir et al., 2023). Die unterbrochene Zeitreihenanalyse ist eine spezielle Form der Zeitreihenanalyse. Sie untersucht mithilfe stückweiser Regressionsmodelle „unterbrochene“ Zeitreihen, wobei die Unterbrechung eine gezielte Intervention (z. B. Steuereinführung) darstellt. Um Veränderungen zu analysieren, die auf diese Intervention zurückzuführen sind, werden die Datenpunkte vor und nach der Intervention betrachtet.

Bei der Untersuchung und Bewertung der Wirksamkeit von Maßnahmen auf Bevölkerungsebene, bei denen ein randomisiertes Studiendesign nicht durchführbar ist, gilt die unterbrochene Zeitreihenanalyse als eine der besten Alternativen (Kontopantelis et al., 2015). Das Verfahren ermöglicht es, mittel- und langfristige Trends zu berücksichtigen, die unabhängig von der Maßnahme auftreten könnten und somit Ergebnisse eines Vorher-Nachher-Vergleichs verfälschen würden (Penfold & Zhang, 2013). Eine wesentliche Einschränkung der (unterbrochenen) Zeitreihenanalyse liegt darin, dass es schwierig ist, die isolierten Effekte einzelner Komponenten zu analysieren, wenn diese zeitlich nah beieinanderliegen. Die Aussagekraft der Analyse ist außerdem nur gegeben, wenn die untersuchte Intervention der einzige Faktor ist, der sich zum definierten Zeitpunkt verändert hat (Mathes et al., 2024). Potenziell konkurrierende Einflüsse oder parallele Interventionen müssen daher sorgfältig berücksichtigt und kritisch diskutiert werden. Werden Daten auf Bevölkerungsebene verwendet, sind Schlussfolgerungen auf individueller Ebene nicht möglich; hierfür bedarf es Zeitreihendaten, die Messungen von denselben Individuen umfassen (Penfold & Zhang, 2013).

A.2.2 Multivariate Regression

Bei der multivariaten Regression werden im Gegensatz zur univariaten Regression nicht nur eine, sondern mehrere abhängige Variablen gleichzeitig analysiert. Ein Beispiel hierfür ist eine Untersuchung, bei welcher der Einfluss von Ernährungsgewohnheiten wie dem Konsum von Fleisch, Gemüse, Getreideprodukten, Obst und Schokolade auf verschiedene gesundheitliche Parameter wie BMI, Blutdruck und Cholesterinspiegel betrachtet wird (Backhaus et al., 2023).

Ein Vorteil dieser Regressionsmodelle liegt in der simultanen Schätzung der Modellparameter zur Klärung der Zusammenhänge zwischen abhängigen und unabhängigen Variablen. Eine Limitation besteht (wie in univariaten Regressionsmodellen) in der korrekten Auswahl von Prädiktorvariablen (Johnson & Wichern, 2007). Beispielsweise besteht die Gefahr von Multikollinearität, d. h. eine starke Korrelation von zwei oder mehr Prädiktoren, was zu Schwierigkeiten bei der Schätzung ihrer individuellen Effekte führt.

A.3 Unterstützende Begleitverfahren

A.3.1 Monte-Carlo-Simulation

Bei Monte-Carlo-Simulationen handelt es sich um „Szenarien, bei denen dasselbe Verfahren sehr oft in zufälligen, d. h. einer bestimmten Verteilung gehorchenden Konstellationen durchgeführt wird, um am Ende durch Mittelung das gewünschte Resultat zu erhalten“ (Bungartz et al., 2013, S. 28). Im Kontext von Simulationsstudien

ermöglicht diese Methodik die Berücksichtigung von Unsicherheiten in den Eingangsparametern, welche die Simulationsergebnisse beeinflussen können. Zu diesem Zweck werden zunächst Wahrscheinlichkeitsdichtefunktionen für die Eingangsparameter geschätzt, aus denen dann wiederholt Stichproben gezogen werden. Das Modell wird anschließend für jede Stichprobe ausgewertet, wodurch Stichprobenverteilungen der Ergebnisgrößen entstehen. Auf Grundlage dieser Verteilungen werden relevante Ergebnisstatistiken berechnet (Harrison, 2010). Monte-Carlo-Simulationen werden meist eingesetzt, um komplexe Zusammenhänge zu analysieren. Insbesondere dann, wenn Unsicherheiten berücksichtigt werden sollen. Da diese Methode auf numerischer Approximation basiert, erfordert sie jedoch umfangreiche Datenmengen und einen hohen Rechenaufwand um gute Ergebnisse zu gewährleisten. Die Qualität der Resultate hängt entscheidend von der Validität des zugrunde liegenden Modells und der Definition der Modellparameter ab.

A.3.2 Sensitivitätsanalyse

Die Sensitivitätsanalyse ist eine Methode, die in verschiedenen Bereichen, einschließlich der Wirtschafts-, Sozial- und Naturwissenschaften, verwendet wird, um zu untersuchen, wie empfindlich ein Modell auf Veränderungen seiner Eingangsparameter reagiert. Bei dieser Analyse werden die Auswirkungen von Variationen in den Eingangsgrößen auf die Ausgabewerte eines Modells untersucht, um zu bestimmen, welche Parameter die Ergebnisse am stärksten beeinflussen. Ziel der Sensitivitätsanalyse ist es, die Unsicherheiten in den Modellvorhersagen zu quantifizieren und zu verstehen, wie robust das Modell gegenüber Schwankungen in den Eingangswerten ist. Diese Analyse hilft dabei, kritische Parameter zu identifizieren, die genauere oder verlässlichere Daten erfordern, und unterstützt die Entscheidungsträger bei der Bewertung der Zuverlässigkeit und Stabilität von Modellergebnissen (Romeike & Stallinger, 2021).

A.3.3 Difference-in-Differences

Das Difference-in-Differences-Verfahren ist eines der am häufigsten verwendeten Verfahren in Studien, welche die Wirksamkeit gewisser Maßnahmen evaluieren und findet demnach breite Anwendung in der Wirtschaft, der öffentlichen Politik, der Gesundheitsforschung und zahlreichen anderen Disziplinen. Es kombiniert Vorher-Nachher-Vergleiche mit Interventions- und Kontrollgruppenvergleichen, wodurch es sowohl intuitiv verständlich als auch vielseitig einsetzbar ist. Die Methode betrachtet zwei Zeitpunkte – vor und nach der Intervention – und vergleicht die Differenz der Ergebnisse der Interventionsgruppe (mit Intervention) mit der Differenz der Kontrollgruppe (ohne Intervention). Der Effekt der Intervention ergibt sich aus der Differenz dieser beiden Differenzen. Allerdings kann die Anwendung durch Probleme wie die Verletzung der Paralleltrendanahme, heterogene Effekte innerhalb der Gruppen, zeitabhängige Variation der Interventionseffekte, konfundierende Faktoren oder Auswahlverzerrungen aufgrund nicht zufälliger Gruppenzuweisung beeinträchtigt werden. Zusätzliche Adjustierungen sind daher notwendig, um Verzerrungen zu minimieren (Fredriksson & Magalhães de Oliveira, 2019).

A.3.4 Change-in-Change

Das Change-in-Change-Verfahren ist ein statistisches Verfahren zur Bewertung von Interventionen wie politischen Maßnahmen. Im Gegensatz zur einfacheren Difference-in-Differences-Methode, die davon ausgeht, dass die Trends zwischen der Kontroll- und der Interventionsgruppe ohne Eingriff gleich verlaufen wären, ermöglicht das

Change-in-Change-Verfahren die Berücksichtigung komplexerer Veränderungen in der Verteilung der Variablen. Dadurch können genauere Ergebnisse erzielt werden, insbesondere in Fällen, in denen die Effekte in verschiedenen Bevölkerungsgruppen unterschiedlich stark sind. Das Verfahren hat jedoch auch Nachteile: Es ist komplex und erfordert umfangreiche Berechnungen und große Datenmengen, was die Anwendung bei begrenzten Ressourcen erschwert. Modellierungsfehler oder fehlerhafte Annahmen können zu verzerrten Ergebnissen führen, deren Interpretation insbesondere bei heterogenen Effekten schwierig ist (Athey & Imbens, 2006).