Aus der Mund-, Zahn-, und Kieferklinik Heidelberg
(Geschäftsführender Direktor: Prof. Dr. Peter Rammelsberg)
Poliklinik für Zahnerhaltungskunde (Ärztlicher Direktor: Prof. Dr. med. Dr. med. dent. Hans Jörg Staehle))

EMPIRICAL METHODS FOR CAUSAL INFERENCE

Inauguraldissertation zur Erlangung des Doctor scientiarum humanarum
(Dr. sc. hum.)
an der Medizinischen Fakultät Heidelberg der Ruprecht-Karls -Universität

Vorgelegt von
Frank Gabel
aus Bad Mergentheim

2020

Doktorvater:	Herr Prof. Dr. rer. pol. Dr. med. dent. Stefan Listl
Dekan:	Herr Prof. Dr. med. Hans-Georg Kräusslich

Contents

List of Figures

List of Tables

Chapter 1

Introduction

During the writing of this PhD thesis, I incidentally came across the famous painting from French artist Paul Gauguin "Where Do We Come From? What Are We? Where Are We Going?". It artistically conveys the human obsession of knowing the causes of things - why each thing comes into and goes out of existence, and why it exists in the first place – in short: answers to the philosophical question "why?". In fact, our ability to perform predictive causal reasoning and to answer questions causally has made *homo sapiens* the most advanced species in history.

One of the pioneers of causal patterns of thinking, David Hume, once stated: "[...] all reasonings concerning matter of fact seem to be founded on the relation of Cause and Effect.". In light of today's ubiquitous statistical models designed to predict various outcomes such as tomorrow's weather, the likelihood of malicious health conditions, future earthquakes, genetic predispositions from gene expression data, and so on, this statement seems quite far-fetched. However, often times, predictive models are based on the extrapolation of observed past associations onto the future while completely lacking clear causal evidence.

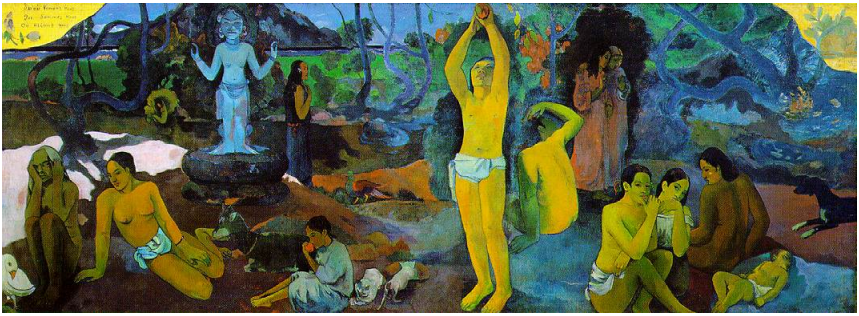

Figure 1.1: Where Do We Come From? What Are We? Where Are We Going? is an 1897 painting by French artist Paul Gauguin. In the upper left corner, the original French inscription can be seen: D'où Venons Nous / Que Sommes Nous / Où Allons Nous.

As has been repeated mantra-like, statistical correlation of variables A and B, i.e. them occurring together does not at all necessarily imply causation - in fact, causation is only one of several explanations for an observed correlation: A could actually cause B (direct causation), B could cause A (reverse causation), a third variable X could cause both A and B (consequences of a common cause), researchers could be conditioning on a collider Z, which is caused by both A and B or the association might be caused by random noise without there actually being any dependency. Therefore, simply assuming causation from correlation is a logical fallacy ("Cum hoc ergo propter hoc" - "with this, therefore because of this") and not a legitimate form of scientific argumentation. However, sometimes people commit the opposite fallacy – refusing even well-founded arguments that are based upon correlation entirely, as correlation could never imply causation. This would dismiss a large swath of important scientific evidence. To inform on causal relationships between variables of interest, well-conducted randomised controlled trials are deemed the gold standard. In randomised controlled trials, a coin flip decides about the assignment to treatment. This way and under some assumptions, treated and untreated individuals could be exchanged without

expecting a change in scientific conclusion.

In the absence of randomised controlled trials, researchers often have to resort to observational data. The challenges that are then faced in pursuing correct causal conclusions involve developing an understanding of natural and induced variation in explanatory variables from both a theoretical and empirical perspective and determining why certain variables take particular values - in other words, to reason about the data-generating process. This is necessary as, as will be seen, naive comparisons between variables across groups are likely to yield biased results. Even though no statistical technique can make the argument to move from correlation to causation persuasive, it is, under certain conditions, possible to obtain valid causal estimates of treatment effects even if randomised experiments are not feasible. In this thesis, some of these methods will be applied to scenarios in the field of health economics.

1.1 About Inference in Science

According to the Merriam-Webster dictionary, science is defined as "the state of knowing: knowledge as distinguished from ignorance or misunderstanding", aiming at trying to build and organize knowledge systematically in the form of testable explanations about certain aspects of the universe.
Accumulating this knowledge works differently for various disciplines of science. While it is widely prevalent to build on previous knowledge acquired by existing work and thereby expanding understanding, the means of expanding this knowledge span from theoretical calculations over observational studies to randomised experiments where some disciplines possess the luxury of performing the latter while some don't. For example, it's not possible to conduct several supernovae to determine whether a particular gamma-ray outburst was caused by it - contrary, different patients can be administered different medications and deduct causal statements by simply screening their particular bodily responses quite easily. But even in health

and social sciences, being able to conduct experiments is not the norm but rather the exception. Often, experiments are unfeasible due to ethical, legal, and practical impediments or due to unbearable cost.
In such cases, observational data are typically analysed ex-post. The scientific benefit is obvious as scenarios of interest are exposed to analyses that are usually out of question.

Common to all branches of science is the desire to examine a given hypothesis, a proposed explanation for a phenomenon of interest. In empirical sciences, more particularly, researchers test these hypotheses against experience by observation or experiment. Typically, data from a sampling process are available and the scientific progress consists of "inductive" inference, i.e. inferring universal statements from "singular" statements. The question of whether these inductive inferences are justified, or under what conditions, is known as the "problem of induction" (Popper, 1959).

Statistical testing of hypotheses overwhelmingly often includes the derivation of a test statistic from empirical data whose singularity given a null hypothesis (which is assumed to be true) is tested. While this approach is statistically valid and forms the basis of the majority of literature in empirical sciences, critique of it actually fills volumes and is best subsumed by 1) with large enough samples, every null hypothesis can be falsified 2) the elusive interpretation of the p-value as a "heuristic piece of inductive evidence" as opposed to items conveying probabilistic dependencies such as confidence intervals and 3) the strong tendency of journals to require statistical significance as a criterion for publication (Benjamin et al., 2018; Carver, 1978; Chow, 1997; *The Significance Test Controversy: A Reader* 2006)[1]. In recent times, there has been increasing consent that many areas of empirical science are in a 'replication' crisis of producing too many false positive non-replicable results (Loken and Gelman, 2017), thereby wasting research funding, eroding credibility and slowing down scientific progress. As a consequence, some journals have gone so far as to either ban the use of p-values altogether

[1] This has lead to the formation of the term "publication bias".

(often in favour of confidence intervals). Also, the American Statistical Association (ASA) recently took the unexpected step of releasing a statement on the "Context, Process, and Purpose" of p-values in hopes of providing some clarity about their implications and meaning (Wasserstein and Lazar, 2016). Despite these efforts towards the avoidance of misuse, the above remarks point out that knowledge is best acquired in ways enabling causal inference "by design" and not only by argumentatively well-grounded associations.

The above problem is aggravated by the problem that observational studies also often fail to address common endogeneity pitfalls such as omitted variables, omitted selection biases, simultaneous causality, common-method variance and measurements, rendering the establishment of valid cause-and-effect relationships impossible. Methods specifically designed to allow for causal conclusions such as instrumental variable estimation, regression discontinuity modelling and differences-in-differences methods bypass some of these problems. In this work, the strengths, weaknesses and limitations of these methods are demonstrated using demonstrative case studies from my own research. To assess the extent to which these methods facilitate causal conclusions, different causal frameworks are consulted which allow researchers to use a priori domain knowledge about the causal structure of interest, defining explicit research hypotheses to make valid causal inferences. What follows is theoretical groundwork of causal inference which are laid before turning to said empirical methods.

1.1.1 The causal hierarchy

In groundbreaking work, starting with the book "Causality: Models, Reasoning and Inference" and recently consolidated with "The Book of Why", Judea Pearl postulates a three-layer hierarchy concerning causal questions whereby each level requires more detailed information than the layer below and answering questions at level i $(i = 1, 2, 3)$ is only possible if information from level j $(j > i)$ is available (Pearl, 2009; Pearl and Mackenzie, 2018).

.

Association: $P(y|x)$ Associations embody purely statistical relationships

and can be characterized by naked observational data - for example, collecting weather data and finding that rainy weather is associated with fewer people buying ice-cream (and vice versa) - such associations can be inferred directly from data using tools from probability theory, namely conditional expectations: $P(\text{icecream}|\text{rain})$. Questions at this layer do not need any causal information whatsoever and are therefore placed at the bottom of the hierarchy. Much research in statistics and artificial intelligence is devoted to finding answers to these sorts of questions when the knowledge of the joint distribution is constrained by missing or limited information (Shpitser and Pearl, 2008). In tasks where prediction is the goal (practically concerning many applications of artificial intelligence), this layer is adequate as inference is neither desired nor conductible.

Intervention: $P(y|do(x), z)$ Interventions typically answer "What if"-questions - this layer not only contains what is seen, but makes it possible to change what is seen. Observational data alone cannot answer such questions, as they involve information that relates to a change in some variable. A typical question at this level would be: *What will happen if we brush teeth thrice a day instead of twice?* Randomized trials belong to this category. (Holland, 1986) even argues "No causation without manipulation", hinting that there needs to be some sort of manipulation to separate correlation from causation.

Counterfactuals: $P(y_x|x', y')$ Going back to the philosophy of causal thinking established by David Hume and Mill, this level of causal hierarchy deals with distributions that span multiple "parallel worlds" of which only one can ever be observed. A typical question at this level would be: *What would have happened if we had brushed teeth thrice a day instead of twice?* It is an extension of the above principle as it eliminates the implicit notion that interventional changes to a variable take time which might influence other time-dependent variables, whereas counterfactual theory examines the very same individual in different manifestations of reality.

As (Pearl and Mackenzie, 2018) state, each layer in the hierarchy has a syntactic signature characterizing statements admitted into that layer. For example, the association layer is characterized by conditional probability statements, e.g., $P(y|x) = p$ stating that: the distribution of event $Y = y$ given that it is observed that event $X = x$ is equal to p.
At the interventional layer, statements such as $P(y|do(x), z)$ are of interest, which means "The distribution of event $Y = y$ given that researchers intervene and set the value of X to x and observe event $Z = z$". Such expressions can be estimated experimentally.
Finally, at the counterfactual level, expressions of the type $P(y_x|x', y')$ are of interest which stand for "The distribution of event $Y = y$ had X been x, given that actually, X is observed to be x' and Y to be y'".

1.1.2 Causal Inference in controlled experiments

As enlisted in the previous chapter on the causal hierarchy, the second layer of the causal hierarchy postulated by Judea Pearl is concerned with "intervention". In practice, this layer encompasses the most straightforward way to inferring causality - experiments - manipulating a treatment variable (i.e. an intervention) to determine the effect on a dependent outcome variable. Experimentation is a powerful methodology that enables scientists to establish causal claims empirically by randomly assigning study units to treatment and control groups. Thereby, exchangeability is granted, i.e. the joint distribution of observations is invariant under permutations of the subscripts (Good, 2002). Changes in outcome can then be attributed to the treatment and an estimation of the average treatment effect is formed. Experiments vary greatly in scale and purpose - depending on the problem statement at hand, a multitude of designs is deemed appropriate. A comprehensive overview can be found in (Campbell and Stanley, 2015).
It will be argued in forthcoming chapters how these principles apply to any causal claim made, even if no a priori treatment assignment is possible.

1.1.3 Causal Inference in observational studies

The ancient commonplace stating that "correlation does not prove causation" has been used to reinforce the preference of experimental to observational studies for a long time all over the empirical literature. Obviously, correlation indeed does not prove causation, but it does not disprove it either. Due to strict regulations on experiments, legal and ethical reasons, most data that social science researchers have access to is observational, lacking random assignment of individuals to treatment[1]. Naive ways of analysis run into trouble here - lacking exchangeability leads to the impossibility of performing valid causal inference and to estimates being inherently biased (De Finetti, 1972; Lindley and Novick, 1981). In this section, common forms of bias will be presented, displaying how they affect estimates in naively estimated models.

The term *bias* is defined as a deviation of the expected value of the results from a "true" underlying quantitative parameter being estimated, stemming from errors in data collection, analysis, interpretation or publication. Avoiding bias in parameter estimates is "virtually impossible" if randomization is no viable option (Cochran and Rubin, 1973). Bias may result in inconsistent or wrong parameter estimates and eventually false claims. Therefore, it should be carefully considered when interpreting the results of such studies. The most prominent sources of bias include, but are not restricted to

- selection bias
- endogeneity
- information bias
- Simpson's paradox

Selection bias is a general term describing preferential exclusion of samples from sample data (either by self-selection or by decision of data analysts), thereby making the sample selected for analyses non-representative of the

[1] The division between experiments and observational studies is not clear as "natural experiments" are typically both experiments and observational studies with researchers lacking control over

population intended to draw conclusions on. It comes in different flavours (Berk, 1983) and constitutes a major obstacle to valid causal and statistical inferences and cannot be dealt with by neither randomized experiments nor observational studies. For example, conducting a survey in a dentists' practice may lead to unreliable conclusions as it relies on self-selection of individuals into answering a questionnaire, as these individuals are likely not representative of the population (some individuals may be embarrassed to respond since they do not visit the dentist regularly or the likes, also linguistic or health barriers may lead to non-random exclusion, commonly called non-response bias). Another example of selection bias is the very well documented *healthy worker bias* (McMichael, 1976), which describes the difficulty of comparing subgroups (such as healthy workers) with the entirety of the population. (Heckman, 1979) describes that in presence of selection bias, regression coefficients are confounded with regard to the function determining the probability that an observation makes its way into the non-random sample. In certain situations, selection bias can be mitigated using *Heckman correction* where self-selection is controlled using an additional predictor function. However, it has since been shown that this method only works in special scenarios (particularly in absence of multicollinearity) (Puhani, 2000). Therefore, it is vitally important for researchers to clarify possible sources of selection bias and restrictions that apply to any conclusion made.

Information bias refers to inexact or wrong measurements or classifications of outcomes, covariates or exposure in certain or all observations within a study, leading to different quality (accuracy) of information between comparison groups (a conclusive overview of types of information biases can be found in, for example, (Althubaiti, 2016)). The occurrence of information biases may not be independent of the occurrence of selection biases (Hartge, 2015).

Endogeneity refers to situations where an explanatory variable is correlated with the error term. In this case, a specified model is not reflective of causal situation that it tries to capture - of course, by nature of OLS, it will still correctly grasp mere correlations between all included variables.

For example, a simple depiction of a linear relationship between a dependent variable Y and an independent variable X, parametrized by a coefficient vector β, is

$$y = \beta X + \epsilon \tag{1.1}$$

This equation can be interpreted in a number of ways. One could think of it as a way of predicting y based on X's values (or even, after shuffling the coefficients, as a way of predicting X based on y's values) or as a way of conveniently modelling the conditional distribution $\mathbb{E}(y|X)$. In these cases, endogeneity is not an issue. However, once equation 1.1 is coerced to embodying causation, the equation suddenly becomes "directional" with X being interpreted as the cause and y as the effect (DAG representation $X \rightarrow y$). Then, β becomes the answer to the question "What would happen to y if X was increased by 1?" Using this interpretation, using OLS for estimation amounts to assuming that:

1. X causes Y
2. ϵ causes Y
3. ϵ does not cause X
4. Y causes X
5. Nothing which causes ϵ also causes X

Failure of any of (3-5) will generally result in $E(\epsilon|X) \neq 0$. A perfectly conducted randomized experiment actually forces (3-5) to be true (if X is picked randomly, it obviously is not caused by Y, ϵ or anything else).
This way, the methods used in this thesis can be contextualized once more - in so-called "natural experiments", researchers try to find real-world circumstances where (3-5) are somehow fulfilled. In the setting of instrumental variables, the fact that the causation is wrong is being corrected (by making another, different, causal assumption as will be argued in chapter 2.1).

In order to obtain an unbiased estimate of β, the exogeneity assumption $E\left(X^T \epsilon\right) = 0$ needs to be fulfilled. In observational studies, this assumption

may be violated in a number of ways, which all lead to endogeneity, i.e. $E\left(X^T \epsilon\right) \neq 0$:

measurement error in X If one of the independent variables within X is measured erroneously, endogeneity ensues. Assume only $X^\star = X + \tau$ is observed (with τ being arbitrarily distributed "measurement noise"). Then, the regression model 1.1 becomes

$$\begin{aligned} y = & \quad \beta X^\star + \epsilon \\ y = & \quad \beta\left(X + \tau\right) + \epsilon \\ y = & \quad \beta X + \epsilon + \beta\tau \\ y = & \quad \beta X + u \text{ (where } u = \epsilon + \beta\tau) \end{aligned}$$

This then fulfils the very definition of endogeneity, i.e. error term u and explanatory variables $X^\star$ being correlated (they obviously both are functions of τ).

reverse causality / simultaneous equations If two variables are co-determining each other, the exogeneity assumption also fails. There is an important distinction between reverse causality and simultaneity. Reverse causality entails a misidentification of cause and effect - the regressand X is hereby fully causing the regressand Y (DAG representation: $X \leftarrow Y$). As (Gerstman, 2013) states: "although one may be tempted to say that low social status causes schizophrenia, another plausible explanation is that shizophrenia causes downward social mobility (so that schizophrenics cannot maintain the normal social relations required to maintain a high socio-economic status)".

The latter entails a two-way causal relationship of X causing changes Y and Y causing X, likewise ($X \leftrightarrow Y$). It's unclear whether this situation even exists (discussions of this can become quite philosophical) as causality requires temporal succession - variables causing each other would then require concurrency. Therefore, examples thereof are

mostly constructed and of no practical use (like electric current) (Kline, 1980).

Preventing bias arising from reverse (or simultaneous) causality is done through “common sense” as these situations are logically improbable and require strong prior arguments or information encoded e.g. in causal graphs. In both of these cases, estimating the obvious regression equations leads to endogeneity.

omitted variables Omitted variable bias comes in many shapes and forms - omitted regressors, omitted interaction or polynomial terms, omitted selection and omitted fixed effects. If variables are omitted that explain part of the variation within the independent variable, the model will reflect this variable in the error term.

As a researcher, keeping track of all potential sources of bias (and measuring them) is typically impossible. To strengthen arguments in favour or causation, empirical literature provides frameworks that allow causal reasoning. In the following sections, some of the more common ones will be reviewed.

1.1.4 Natural Experiments / Quasi-experimental studies

While purely observational data can lead to situations prone to systematic bias as shown above, certain scenarios resemble experiments even though the researcher does not control the surroundings of the experimental implementation. Empirical literature subsumes these scenarios natural experiments, some of which will be subject of study in later chapters.

1.2 The Bradford-Hill Criteria

As has been debated in the previous chapters, neither experiments nor observational data can unveil causation in a metaphysical sense at all. Thus, argumentative strategies are deemed possible to support presumed causal connections between variables of interest, following the known phrase by James Whitcomb Riley "When I see a bird that walks like a duck and swims

like a duck and quacks like a duck, I call that bird a duck."
The “Bradford-Hill criteria”, also called “Hill Criteria for causality” are a set of nine “aspects of association”, i.e. minimum conditions that help build an argument for a supposed causal relationship of an observed association between variables. They are based on the inductive canons of John Stuart Mill and the rules given by Hume (Hume, 1739/1978; Mill, 1843; Mill, 2009) - the most renowned version that will be introduced below was formulated by the English epidemiologist Sir Bradford Hill (Hill, 1965b).

Strength The larger the magnitude of the association, the more likely a causal relationship is present, even if a small effect does not imply an absence of causality.

Consistency / Reproducibility Causality is more likely to be in place if an association has been observed across a variety of locations, populations, and methods. Also, Hill stressed the importance of reproducible findings because a single study, no matter how statistically sound, cannot be relied upon to prove causation due to enduring threats to internal validity.

Specificity If an exposure is specific to exactly one disease, there is no other conceivable explanation for the association, then causality is likely. It has been argued that this criterion is rather weak (from an epidemiological standpoint), as today typical exposure and health concerns at the forefront of research revolve around a plethora of risk factors such as complex chemical mixtures as well as low-dose environmental and occupational exposures, making them highly unspecific (Fedak et al., 2015).

Temporality Causality entails the temporal ordering of causes always preceding their effects in time. It is widely important to identify the valid temporal succession between variables to obtain unbiased estimates of their relationships. This condition is deemed “inarguable” in most practical settings, making study designs ensuring a temporal progression of exposure and disease more persuasive (Rothman and Greenland,

2005).

Biological gradient In the presence of causality, a larger dose leads to a larger effect. However, there are conceivable cases where either the mere presence of the cause leads to the effect or where there is an inverse relation, i.e. greater exposure to the cause leads to a diminished effect.

Plausibility A causal claim can justifiably be supported by the presence of a causal explanation describing possible pathways between cause and effect.

Coherence Previous findings (whether causal or associational, in the original paper it is termed "facts") explaining the relationship between cause and effect should not contradict causal explanations - coherence increases the likelihood of the presence of causality

Experiment When action has been taken on the basis of given evidence, for example reductions of dust in workshops, a change in lubricating oils or the stopping of smoking, strong support for causal hypothesis can be unveiled.

Analogy A causal claim can be supported by the existence of similar, but not equal causal connections.

In the current age of ever increasing capabilities of analytical computing for exploring potential cause-and-effect relationships, (Fedak et al., 2015) proposed an update to the Bradford-Hill criteria. In the case studies presented in later chapters, the above criteria will be used (mostly implicitly) to support or dismiss presumed causal connections therein taking into account respective data and interpretational background knowledge.

1.3 Modern models for causal inference

Before turning to methods capable of performing causal inference using observational data, the present section provides a succinct overview of common causal frameworks that have been widely used in empirical literature.

Hereby, the focus will lie on non-practical characterizations of the respective frameworks. In chapter 4.1, the applicability of these frameworks will be discussed in the context of actual scientific scenarios of causal inference.

As was already argued, causal inference is tightly linked with randomized experimentation. The three models of causality dominating the evaluation literature (i.e. Neyman-Rubin Causal Model (RCM), Campbell Causal Model (CCM) and Pearl Causal Model (PCM)) support the viability of causal inference from observational data when certain assumptions are met relating ex-post scenarios with controlled experiments.

1.3.1 The Neyman-Rubin Causal Model: The Potential Outcomes approach

The Neyman-Rubin causal model (RCM) (Rubin, 1974; Rubin, 1977; Rubin, 1978) is an approach to statistical analyses of cause and effect based on the notion of *potential outcomes* (therefore, it is also often called *the potential outcomes model*), allowing a rather straightforward definition of causal effects.
In a population under scrutiny of n units (whether a person, cohort, or population), each of these units is able to be exposed to either a treatment T or a control C. The treatment is given to a unit i and the outcome variable of interest $Y_i(T)$ is observed. Ideally, the control treatment C is given to the same participant at the same time and in the same context, and the so-called counterfactual outcome $Y_i(C)$ is observed.
The counterfactual outcome is a mental notion of what would have happened in a world where treatment assignment was different - in this way, the framework can be viewed as a missing data problem where for each individual, only one of the two potential outcomes is observed. Then, the individual-level causal effect is conceived to be the difference between actual and (hypothetical) counterfactual outcomes:

$$Y_i(T) - Y_i(C) \tag{1.2}$$

Counterfactual outcomes are inherently unobservable. In fact, being able to deal with counterfactuals corresponds to level 3 of Pearl's causal hierarchy (see chapter 1.1.1, (Pearl, 2009)). Humans make use of such thinking all the time and it intuitively makes sense that being able to answer such "What if?"-questions is pretty useful for intelligent behaviour.
However, this does in no way translate to useful properties for causal empirical research: in expression 1.2, for each individual i, either the outcome under treatment $Y_i(1)$ or the outcome under no treatment $Y_i(0)$ can be observed, but never both - observing both actual and counterfactual outcomes is inherently impossible. This is called the FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE. Therefore, instead of longing for the inferential goal of treatment effect estimation for an individual unit, counterfactual analysis in Rubin's sense aims at the easier target of calculating AVERAGE CAUSAL EFFECTS. The averaging, in this case, corresponds to averaging the (unobservable) individual causal effects across all n units in some well defined population, resulting in the AVERAGE TREATMENT EFFECT (ATE) (Holland, 1986)

$$\tau_{ATE} = \mathbb{E}(Y_i(T)) - \mathbb{E}(Y_i(C)) \tag{1.3}$$

However, the problem arises how to assign units i to either treatment or control groups. To that end, Rubin states that the fundamental problem of causal inference can be overcome by considering two assumptions, namely the independence assumption and the assumption of strong ignorability.
The independence assumption outlines a classical randomized experiment, where by assumption (and, of course, best practice), treatment assignment Z_i for a unit i is independent of the potential outcomes $(Y_i(T), Y_i(C))$ and all other potential confounding variables. Causal inference for randomized experiments is uncomplicated because when independence holds, the simple

$$\tau_{ATE} = \mathbb{E}(Y_i(T)|Z = T) - \mathbb{E}(Y_i(C)|Z = C) \tag{1.4}$$

holds. That's because due to random assignment, treatment and control groups are (on expectation) similar, and any difference in outcomes can be interpreted as a corresponding causal effect. Hereby, it is crucial that the assignment mechanism is the sole explanation for why some units received treatment and others control.
This, however, is a strong assumption which often does not hold in observational studies. In these cases, strong ignorability allows estimating the average treatment effect anyway. It holds when

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp X|Z \tag{1.5}$$

where X is a vector of covariates that measures the characteristics of some unit (e.g., gender, paraent's educational level, etc.) before the treatment assignment, and thus is not affected by the treatment. Then, the fundamental problem of causal inference can be overcome by utilizing additional knowledge on pretreatment variables - treatment effects can then be estimated without bias by adjusting (or "controlling") for the confounding variables Z[1].

In the empirical part of this work, Rubin's framework will be revisited to deduct statements about the validity of causal claims in different subpopulations. For future reference, it is sensible to introduce common forms of treatment effects.

1.3.1.1 An overview of treatment effects

As has been elucidated in the above sections, causal claims on the level of individuals are not obtainable, falling victim to the fundamental problem of causal inference. Therefore, causal inference invariably aspires to estimate causal effects on subpopulations. Even in study designs that are reknowned for allowing causal inference such as randomized controlled trials, IVs or DiD (this is assuming a "correct" study design) evidence of causation can

[1] An additional assumption that needs to be mentioned is *overlap* that ensures that for any covariate, there are units in both treatment and control groups: $0 < P(Z_i = 1|X_i = x) < 1$

only be drawn for subpopulations, thereby limiting what can be learned from the study.
The ATE has already been covered. In certain scenarios, other kinds of cumulative treatment effects are of interest or calculated. Generally, the value of interest is

$$[Y_i|D_i = 1] - [Y_i|D_i = 0]$$

As only one of these items is observable, the above is only a theoretical quantity. Therefore, one is often mostly interesting in the Average Treatment Effect (ATE), which compares outcomes across populations of treated and untreated units:

$$\mathbb{E}[Y_i|D_i = 1] - [Y_i|D_i = 0]$$

In certain situations, only the treatment effect on the subpopulation of treated individuals can be calculated, the Average Treatment Effect on the Treated (ATET).

$$\mathbb{E}[(Y_i|D_i = 1 - Y_i|D_i = 0)|D_i = 1]$$

In other settings, only the treatment effect for the subpopulation of units compliant with treatment assignment can be identified, called the Local Average Treatment Effect (LATE):

$$\mathbb{E}[(Y_i|D_i = 1 - Y_i|D_i = 0)|Z_i = 0]$$

The last variant of treatment effects is also known as the complier average causal effect (CACE). It always occurs when there is either one-sided or two-sided non-compliance, leading to decreased internal validity. This problem will be revisited in the chapter concerned with the quasi-experimental method of *instrumental variables*.

1.3.2 The Pearl Causal Model - Draw inferences from Causal Graphs

The use of graphical representations to display causal relationships began with the seminal work of (Wright, 1921) about interrelating factors in agriculture. The Pearl causal model (PCM)(Pearl, 2009) is based on a graphical representation of hypothesized causal relationships. Although it is natural for humans to interpret graphs causally (an arrow from X to Y representing the causal claim "X causes Y"), the graphical approaches first conveyed purely statistical relationships leading to so-called Bayesian networks or directed acyclic graphs (DAGs). Before turning to the PCM, DAGs will briefly be introduced.

The term directed acyclic graph (DAG) has its origins in graph theory, a discipline of computer science allowing the modelling of complex systems of relation. A directed graph G is a mathematical object describing a pair (V, E) (termed vertices and edges) of sets such that the set of edges E is composed of ordered pairs (a, b) of elements from the set of vertices V. The set of vertices V consists of structureless objects that are connected by edges - if E contains an edge (a, b), the vertices a and b are said to be connected or adjacent. Then, a is referred to be a parent of b and b is a child of a, respectively. A path in G is a sequence of pairwise distinct vertices $V1, \ldots, VN$ such that all consecutive vertices V_i and V_{i+1} are connected by edges. "Acyclic" implies that there is no way to start at any vertex v and follow a consistently-directed sequence of edges that eventually loops back into v again. Further, the notion of *d-separation* will prove useful when hypothesizing about deducing causal statements from assumptions encoded in DAGs. Two nodes X and Y in a graph are *d-separated*, if a node Z "blocks" each undirected path between X and Y. Pearl postulates a theory of causation based on Structural Causal Models described in (Pearl, 1995), subsuming and unifying other approaches to causation and providing a coherent mathematical foundation for the analysis of causes and counterfactuals.

Without any further reasoning, DAGs are just mathematical structures and d-separation and the Markov condition are just connecting DAGs and probability distributions without any causative assumptions or assertions

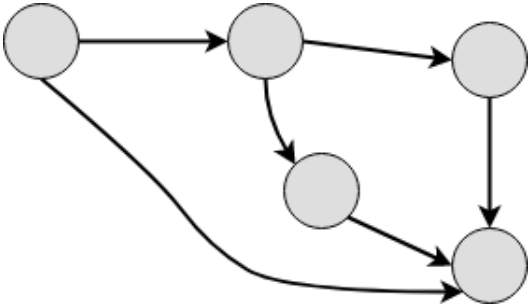

Figure 1.2: A depiction of a very generic Directed Acyclic Graph (DAG).

being present.

1.3.2.1 The interpretation of graphical models

There are, generally, three ways of applying directed graphs to statistical modelling - graphing the structure of a probability model, graphing a hypothesized causal pattern and graphing relations between real-world variables.

modelling probabilistic relationships If a graph is interpreted as to purely convey probabilistic relationships between underlying variables, a rather weak set of assumptions is needed: for this to be valid, the parents pa(X) of each variable X in the graph need to render X independent of all its non-descendants given pa(X). When a graph fulfils this condition, it is said to be *compatible* with the underlying joint probability distribution. In practice, compatibility is given if each parent-child family $\{X, pa(X)\}$ in the graph represents a distinct stochastic process by which randomness decides upon the values of a variable X as a function of the parents pa(X), independently of values previously assigned to variables other than the parents.

modelling causal relationships In more recent work, graphs have been used to represent causal relationships between variables (Spirtes et al., 1993). Numerous authors have proposed directed graphs to convey

causality from early 20th century (Wright, 1921) to more recent artificial intelligence research (Guo et al., 2018). However, typical causal models can't contain every single cause of a given effect due to our incomplete knowledge in interesting domains such as medicine, law, social science, economics and so forth- instead, as causal models are based on prior knowledge and assumptions, they can only display causal relationships with errors and at certain granularities. Still, on the supposition that (some of) these assumptions are correct, ignoring the reality of our ignorance, (Pearl, 2009) derived a rigorous mathematical notation of cause and effect, allowing the quantification of causal effects and using probability theory to quantify uncertainty as in statistical regressions. When DAGs are interpreted causally, the Markov condition and d-separation are in fact the correct connection between causal structure and probabilistic independence.

The copulative element of these approaches is the representation of variables as nodes while directed arrows represent direct cause-and-effect relationships. (Shpitser, 2008) notes that causal graphs also represent *modularity* meaning that full knowledge of all direct causes of a given effect determine the manifestation of the effect no matter all other variables in a model. Also, this modular structure to model how a PCM reacts to changes imposed from the outside. The simplest of these impositions is to set a variable X to a specific value x. This procedure, also referred to as an intervention, is denoted by the so-called *do-operator*. The model $M_{intervention}$ imposed by this intervention is a *submodel* of the original model M, resulting in an interventional distribution, which depicts another way to formalize the intuitive notion of counterfactuals. In the empirical part of this thesis, application of this concept will be attempted.

It is well known that identification of causal effects depends on the structure of the graph representing the causal information, the set of observable variables, the set of outcome variables (there is typically only one), and the set of variables that is intervened on (Pearl, 1995; Pearl, 2009). Us-

ing graphical conditions, most notably the concept of *d-separation* from the theory of directed graphs, one can show whether a causal effect (i.e. the joint response of any set S of variables to interventions on a set T of action variables), denoted PT (S) is identifiable or not.

In other words, dependencies among variables (purely probabilistic or causal depending on the nature of the graph) can be verified by checking if the "flow of dependence" is blocked along paths between variables. *D-separation* yields the precise way in which the flow of dependence can be blocked (Pearl, 1986), allowing the derivation of a strict mathematical calculus of causal effects when there exists a (conditional) probability distribution consistent with the given graphical causal model. Building upon this, Judea Pearl's *do-calculus*, introduced in his 1995 paper "causal diagrams for empirical research" (Pearl, 1995), establishes a mathematical language for connecting statistical and subject-matter information. In particular, the paper develops a non-parametric framework for causal inference using directed graphs to determine if available assumptions are sufficient for identifying causal effects from non-experimental data.
The *do-calculus* describes the conditional distribution one would learn from data collected in randomized controlled trials or A/B tests where the experimenter controls. A pitfall of this strict mathematical notation of causality is the availability of data and a priori knowledge of precise causal relationships that are often unclear in practice.

1.3.3 The Campbell Causal Model - Identify threats to Internal Validity

The third widely used causal model has been brought upon by Donald Campbell (Cambell and Stanley, 1963; Campbell, 1957), whose perspective on causal inference is the most widely used in social sciences, particularly in psychology, education and public health (Shadish, 2010). It revolves around the concept "validity" where "internal validity" describes whether a study supports a claimed cause-and-effect relationship of a given treatment and "external validity" describes to which extent the results of a study can be generalized to another population, time, or setting (in most cases, the whole

population of interest).

The approach of the Campbell Causal Model (CCM) is rather practical, taking into account all phases of pre-experimental, quasi-experimental and experimental designs. It revolves around the idea of identifying nine threats to validity plausibly undermining some aspect of the causal inference process in practical research settings: "We took the position that there could be lots of threats to validity that were logically uncontrolled but that one should not worry about unless they were plausible. The general spirit was that any interpretation of a body of data or research should be regarded as innocent until judged guilty for plausible reasons, as determined through the scientific method of mutual criticism." (Campbell et al., 1988)

For one of these plausible threats to be a problem that needs to be dealt with, they must entail operational differences between treatment (T) and control (C) groups. These nine threats defined in Campbell's approach are

History Events other than planned treatments influence results.

Maturation During study, changes may occur within subjects.

Testing Exposure to a pretest or intervening assessment influences performance on a post-test.

Instrumentation Measurement instruments may be inconsistent or may experience changes in calibration may produce unwanted changes.

Regression to the mean In measurements where randomness is involved, extremely high or low observations tend to regress towards the mean

Selection Treatment groups may entail systematic differences between subjects' characteristics.

Experimental mortality Study attrition of subjects may effect the results in unintended ways.

Diffusion of treatments When multiple treatments are given to the same subjects, it is difficult to control for any effects of prior treatments.

Different kinds of interaction effects This includes interaction effects between selection biases and the experimental variable, interaction effect

of testing or selection-maturation interaction.

Application of the CCM rests on a critical perspective of researchers on their own work both during the design of a study as well as during the evaluation and analysis, viewing "causality" as an additional property of a found association that can be claimed using an argumentative strategy.

Of course, it is impossible to attest that the above system of threats to validity is complete, but the approach has proven to be a thorough and practical tool for evaluating the validity of causal claims in applied research in the social sciences. In the case studies of this thesis, the CCM will be applied implicitly and explicitly to reason about causality in specific, concrete research settings.

1.4 Scope of this work

Any of the above causal models can be used to infer causality in both observational and experimental designs - these models are generic in that they explicitly include the formal synthesis of findings generated by research using "true" controlled experiments. In the empirical part of this thesis, however, these frameworks will be shone upon from the perspective of an applied researcher with access to observational data only. This work thereby contributes to the empirical literature by examining the three models of causal reasoning above - the RCM, CCM and the PCM - in the context of three case studies, showing their unique advantages and drawbacks by embedding them empirically. In section 4.2 of this work, the subjective applicability of them will be analyzed.

These case studies include the analysis of the implementation of a quality improvement framework in the UK primary care, the analysis of the impact of bearing children on oral health and the analysis of an implementation of altered provider incentives in the Danish dental care system. All of these make use of methods utilizing quasi-experiments, i.e. inherent randomization within the data.

In the following, this thesis will be concerned with introducing the setting these case studies take place in, presenting the results of this empirical work, ensued by a discussion about the validity of causal interpretations of their results, particularly with respect to the proposed causal models and lines of thinking along with their implications on common threats to studies based on observation data such as selection bias, confounding and endogeneity.

Chapter 2

Case studies and methods for causal inference regarding observational data

It has been argued in earlier chapters that reasoning about causality requires experiments. The crucial component in these experiments is randomization, ensuring that exchangeability is present (if treatment status of individuals had been reversed, the outcome would not have changed). However, in certain cases, there is no way of translating certain research questions into experimental settings. Some relationships are hard to observe outside of their natural environment (think about natural catastrophes such as hurricanes, nuclear power plant accidents etc.), some exposures can't be assigned to humans for ethical reasons (e.g. most adverse health behaviours such as smoking and drug abuse), policy interventions, participants not wanting to

be randomized or doubts about equipoise. Also, it is not uncommon that randomized experiments are "broken" - for example due to non-compliance (patients refusing treatment, actively seeking out alternative treatment or receiving partial treatment) or general attrition.
Elucidating causal relationships underlying these enigmatic cases of scientific uncertainty often requires either strong, largely untestable assumptions (as in "no selection bias") or, sometimes, a different kind of methodological approach. In some cases, researchers can leverage *observational studies*. Here, treatment assignment is neither manipulated nor randomized. This, however, does not imply that treatment assignment cannot be random – if it indeed is, i.e. if some plausible exogenous variation in the treatment assignment can be found, not all hope is lost. Minding some caveats, valid causal inference can still be performed in these cases. In this chapter, light will be shed on three common methods that allow to draw causal inference from observational data used during my time as a PhD student. The following chapter will introduce readers to the context of the three case studies and an in-depth overview of the methods utilized to analyze the respective ramifications.

2.1 Gain a child, lose a tooth - Using natural experiments to distinguish between fact and fiction using Instrumental Variables[1]

The first case study examines the old wife's tale which states "gain a child and lose a tooth". The idea that pregnancy causes tooth loss has been a wide-spread myth for hundreds of years, but there has been little evidence to deem serious countermeasures by expecting mothers necessary. This gap in literature is bridged by leveraging observational data from a recent large-scale European survey. A unique natural experiment allows for the derivation of causal effects using "instrumental variables".

[1] The corresponding paper has been published in the Journal of Epidemiology and Community Health.

2.1.1 Introduction

Dental conditions are among the most frequent diseases globally (Listl et al., 2016; Marcenes et al., 2013). The loss of permanent teeth imposes a significant burden on people's quality of life, (Gerritsen et al., 2010) yet disentangling the exact biological and behavioural pathways resulting in tooth loss remains a major challenge for research. Against this background, a particularly intriguing question is whether tooth loss is influenced by fertility. Until now, however, there is no causal evidence for or against a relationship between the number of biological children and their parents' number of missing natural teeth (a detailed overview of the related literature is provided in the appendix of (Gabel et al., 2018)).

To address this knowledge gap, this first case study relies on large-scale multi-country data and exploits random natural variation in family size resulting from (i) the birth of twins vs singletons, and (ii) the sex composition of the two first-born children (increased likelihood of a third child if the two first-born children have the same sex). A two-fold effect of fertility on the number of teeth in adults is hypothesized: first, biological effects during pregnancy influencing the oral health of women; and second, indirect effects related to having children (pregnancy and parenting stress, economic burden) which possibly affect both women and men.

Data Source The Survey of Health, Ageing, and Retirement in Europe (SHARE) (Börsch-Supan, 2019), contains data on health, socio-economic status, social and family networks for a total of over 120,000 older adults from 27 European countries and Israel. SHARE Wave 5, conducted in 2013, provides unique information about the number of natural teeth of individuals in Austria, Belgium, Czech Republic, Denmark, Estonia, France, Germany, Italy, Luxembourg, The Netherlands, Slovenia, Spain, Sweden, Switzerland, and Israel (for more details, see (Malter, F. and A. Börsch-Supan, 2015)[1]).
Inclusion and exclusion criteria The population under study consisted of

[1]To guarantee high-quality data across all countries, SHARE employs a centralized training program and rigorous quality control (Alcser KH, 2005; Börsch-Supan, 2019; Malter, F. and A. Börsch-Supan, 2015). Details about data collection are published elsewhere (Borsch-Supan et al., 2013).

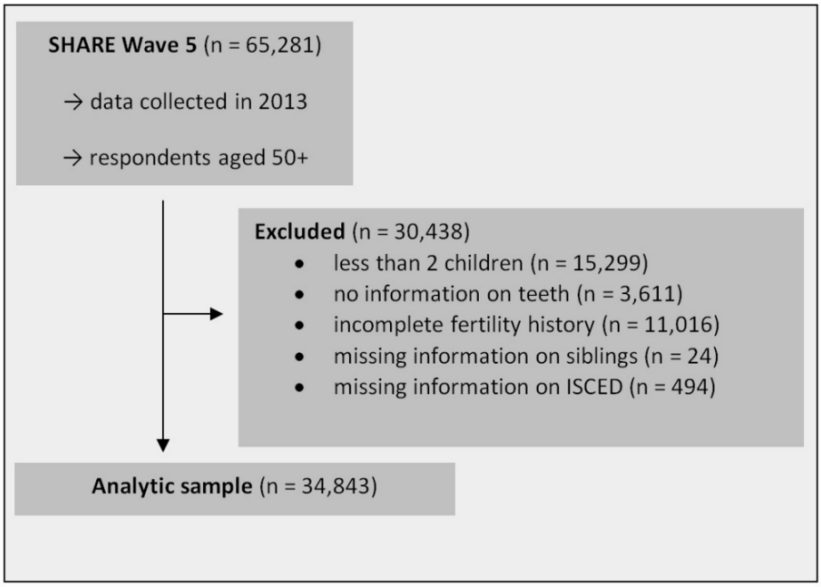

Figure 2.1: Study population and sample attrition.

individuals aged 50 years or older who were enrolled in SHARE Wave 5 unless they did not answer questions on key dimensions such as the number of remaining natural teeth or information on their fertility biography, i.e., the number, sex, and year of birth of their children. Further, for reasons related to the identification strategy explained below, the analytical sample was restricted to individuals with at least two children. After sample exclusions, the final analytical sample included 34,843 individuals aged 50+ with full fertility biographies and information on their number of teeth. Figure 2.1 illustrates the study population and sample attrition.

Dependent and independent variables Analyses are based on SHARE wave 5 data and each participant's number of missing teeth. Participants were asked: "Do you still have ALL your natural teeth (except wisdom teeth)?" (response options: "Yes" and "No"). Participants were informed that "Normally, a person has 28 teeth and 4 wisdom teeth. We are not interested in

wisdom teeth." Participants who reported not having all teeth were further asked: "About how many natural teeth are you missing?". Respondents' number of missing teeth were derived accordingly. It has been shown that self-reports are a valid means to report the number of teeth (Douglass et al., 1991; Gilbert et al., 1997; Ramos et al., 2013). In addition to each respondent's procreation history (number, sex, and birth date of own children), independent variables included each respondent's current age, country of residence, age at first birth, the number of siblings (to account for possible preferences regarding optimal family size acquired in childhood), and education as measured according to the International Standard Classification of Education (ISCED) (UNESCO Institute for Statistics, 2012).

The role of in-vitro-fertilization (IVF) A potential limitation of the twin births identification strategy is given by the recent rise in conception assisted by fertility treatments (IVF) as IVF has increased the probability of multiple births in a non-random fashion (Calhaz-Jorge et al., 2016; Pandian et al., 2015). However, since IVF became available only in the last 25 years and the study sample consists of individuals whose fertile period ended before the introduction of IVF, it seems reasonable to assume that fertility treatments are not responsible for most of the twin births. Robustness checks were performed by restricting the study sample to persons with children born before 1990. Further details hereof can be found in the appendix of (Gabel et al., 2018).

2.1.2 Basics and Estimation

As stated earlier, endogeneous regressors, i.e. unexplained variation between explanatory variables and error terms (for example due to unmeasured confounding) causes the key "exogeneity" assumption of OLS to fail, leading to inconsistent OLS parameter estimates. A classical strategy to encounter endogeneity in the applied literature are instrumental variables (Angrist and Krueger, 2001). The central strategy in IV estimation is to find "instrumental" variables, also simply called "instruments" that are correlated with the exposure of interest but not with the outcome. The variation induced by

instruments can then be used to cleanly estimate the relationship between the predictor and outcome (if the instrument is also not correlated with unobserved confounders).

The following chapter provides a formal overview of IV regression - a more thorough statistical approach can be found in the excellent book of (Angrist and Krueger, 2001). In traditional structural equation models, a linear and additive relationship between a dependent variable Y_i, an endogeneous regressor D_i, a set of exogeneous regressors $X_{1i}, \ldots, X_{ni}$, and an unobserved error term ϵ_i is alleged:

$$Y_i = \beta_0 + \beta_1 D_i + \gamma_1 X_{1i} + \cdots + \gamma_n X_{ni} + \epsilon_i \qquad \text{(OLS)}$$

This can only be estimated consistently using OLS if the covariance between X_i and ϵ_i is zero (*strict exogeneity*). When endogeneity is present, i.e. there exists a systematic relationship between X_i and unobserved causes of Y_i, OLS is generally biased - in such cases, IV estimation can help yield unbiased estimates. The IV estimator is premised on a two-equation model commonly known as "two-stage least squares". In the first stage, the relationship between an independent variable Y and the so-called instrument Z is estimated[1]:

$$X = \gamma + \delta Z + \epsilon \qquad \text{(First stage)}$$

In this equation, termed "first stage", changes in X due to exogenous variation are calculated. Using OLS, one can then estimate $\hat{\delta} = (Z^T Z)^{-1} Z^T X$ and use $\hat{\delta}$ to predict $\hat{X} = Z\hat{\delta}$: Under the IV assumptions, any variation in $\hat{X}$ is then caused by the instrument Z and can be used in the "second stage", resulting in an unbiased estimate of the causal effect of X on Y:

[1] Variable indices are omitted for brevity and readability.

$$Y = \alpha + \beta_{IV}\hat{X} + E \qquad \text{(Second stage)}$$

This approach yields a numerically identical IV estimate as in direct estimation of

$$\beta_{IV} = (Z'X)^{-1}Z'y \qquad \text{(IV estimate)}$$

In subsequent chapters, the internal and external validity of this estimate will be discussed.

2.1.2.1 Identification strategy

The above strategy can be abused to perform inference about the relationship between the number of natural children and the number of missing teeth from observational data which is complicated by the multitude of potential underlying mechanisms. When using Ordinary Least Squares (OLS) regression analysis, various common causes of both tooth loss and parity – some of them unobservable – can result in confounding and biased parameter estimates. For OLS to provide unbiased estimates of the causal effect of children on tooth loss, a "selection on observables" assumption has to be made (Dale and Krueger, 2002; Rothstein, 2009). However, due to the poorly understood mechanisms between fertility and dental health, it is highly unlikely that all variables that correlate with fertility and have an impact on tooth loss can be observed and controlled for in the regression. The "selection on observables" assumption is therefore not appropriate. A clean identification of the causal effect of an additional child on dental health is ideally provided by a randomized controlled trial, a setup that is obviously not available for this research question. Therefore, the instrumental variable approach is employed (estimated using a two-stage least squares (2SLS) regression model). Here, instruments relating to the number of natural children are harnessed, an approach which has previously been used to examine the effects of parity on physical and mental health later in life (Black et al.,

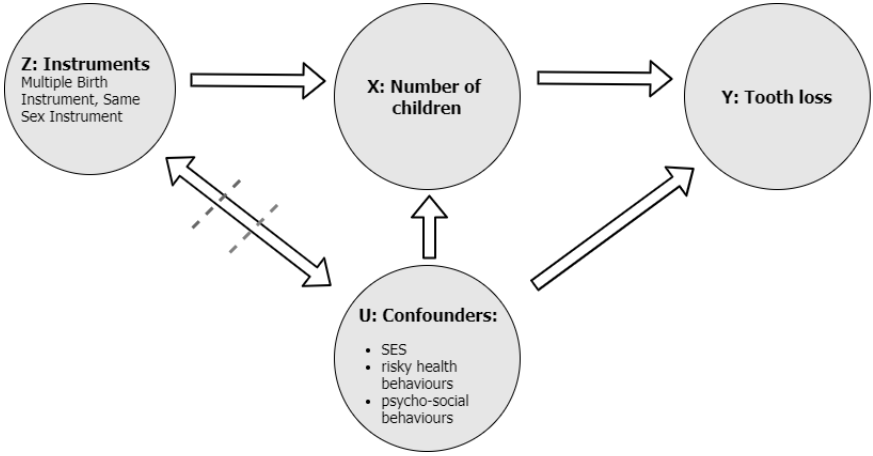

Figure 2.2: Stylised illustration of the instrumental variables approach.

2005; Caceres-Delpiano and Simonsen, 2012; Kruk and Reinhold, 2014a). These instruments are:

- The birth of twins vs singletons ("twin births")
- The sex composition of the two first-born children (assuming an increased likelihood of a third child if the two first-born children have the "same sex")

Figure 2.2 illustrates the principle of the 2SLS approach. The idea is that random variation in Z (the instruments) is directly associated with the predictor of interest X (the number of children). If the instruments Z are linked with the outcome variable Y (tooth loss) only through X and not linked with other confounders U (e.g. socio-economic status), causal inference can be established that uses only the (random) variation in X (number of children) attributable to variation in the instruments Z. Since the interest lies in comparisons between results using the "twin births" instrument and results stemming from the "same sex" instrument, the focus lies on the birth of twins vs singletons at the second birth; both the "twin births" and the "same sex" instrument are intended to primarily identify the effect of having

three instead of two children. For delineation of sex-specific pathways, all analyses were carried out separately for women and men. Besides descriptive statistics and balancing tests, the following estimations were carried out (controlling for the independent variables described laid out in Figure 2.2):

- **OLS regressions** of the number of missing teeth on the number of children
- **Intention to treat (reduced form) regressions (ITT)** of the number of missing teeth on the instruments
- **1st stage regressions (2SLS)** of the number of children on the instruments
- **2nd stage regressions (2SLS)** of the number of missing teeth on the number of children

This regression strategy allows to compare treatment effects from models with different sets of assumptions, aiding in discussing whether a causal connection is present and to which population results can be extrapolated. The IV estimates in particular have a very specific set of assumptions that will be discussed in the following.

2.1.3 Assumptions

Relevance: There exists a causal effect of the instrument Z on treatment status X. In this empirically verifiable assumption, the correlation between instrument and treatment status is calculated. The strength of this association is being evaluated using the F-value. Many researchers use an F-value of 10 to separate *weak instruments* (F-value < 10) from *strong instruments* (F-value > 10). Weak instruments might still be valid means of inserting exogenous variation, but result in wide confidence intervals in the second stage.

Exclusion restriction: There must be no direct effect of Z on potential outcomes Y. This assumption ensures that instruments affect the outcome only through X and not through other confounders (which

Table 2.1: Identifiable subgroups in an IV setup.

	$D_{0i} = 0$	$D_{0i} = 1$
$D_{1i} = 0$	never-taker	defier
$D_{1i} = 1$	complier	always-taker

would then be correlated with X) or the error term. Unfortunately, as the error term is unobservable by definition, this assumption is not empirically verifiable from data and subject-matter knowledge must be used to rule out possibilities for that.

Independence: Conditional on covariates, the instruments are as good as randomly assigned in being independent of potential outcomes and potential treatments. By comparing measured confounders across levels of the instruments Z, potential unbalances can be detected and the independence assumption can be empirically tested. This does obviously not include unmeasured confounders, making the independence assumption only partially testable. Commonly, a 4-way table is used that subsumes covariates across so-called *never-takers*, *defiers*, *compliers* and *always-takers*.

Monotonicity: The instruments affect everyone affected by them in the same way. If homogenous treatment effects were to be assumed, i.e. each individual is affected by the treatment in the same way, instrumental variable estimates would estimate the ATE[1]. However, in real-world scenarios, this assumption is rarely fulfilled and mostly implausible. Therefore, the *monotonicity* assumption has been brought forward, weakening the generalisability of effect estimates.

As one can only observe the expose under actual assignment, there is no way in real world scenarios to differentiate between these subgroups. That being said, within the subgroup of compliers, exchangeability is fulfilled. This fact also nicely displays the connnection between no

[1] As (Lousdal, 2018) points out, this does not imply that treatment effects can't vary, but it requires that the source of heterogeneity in the individual treatment effects is unrelated to observables.

defiers and monotonicity. If the subgroup of defiers is empty, only compliers will make a contribution to the causal effect of Z on Y. In other words, monotonicity assumes that for each subject, the level of the treatment that a subject would take if given a level of the IV is a monotonic increasing function if the level of the IV. For that reason, IV identifies the average treatment effect of compliers only (also termed Local Average Treatment Effect, LATE).

Implications thereof will be discussed in section 2.1.5 and the discussion chapter.

2.1.4 History

Interestingly, the history of instrumental variables entails a very instructive application, which is worth mentioning whenever possible: During the 1853-1854 Cholera epidemic in London, the English scientist John Snow believed that Cholera bacteria are waterborne (Snow, 1855), and the epidemic was linked with consumption of consuming water. A naïve way of analyzing this relationship would have been to analyze the correlation between drinking water quality (X) and Cholera incidence (y). However, those who drank impure water were more likely to be poor, to live in crowded tenements and to live in a surrounding contaminated in other ways, which impose a threat to analyses due to unmeasurable confounding. Valid instruments in this scenario would be strongly correlated with water quality but, at the same time, not correlated with other observed and unobserved determinants of Cholera incidence. Coincidentally, Snow (unknowingly) proposed such an instrument: the identity of the water company supplying households with drinking water (z). At that time, Londoners drew water directly from the Thames. One company, the *Lambeth water company*, took out water from the river upstream of the main wastewater discharge whereas the other company, the *Southwark and Vauxhall company*, took its water directly below the main discharge. The validity of this instrument has been discussed by Jon Snow himself: “the mixing of the supply is of the most intimate kind.

The pipes of each Company go down all the streets, and into nearly all the courts and alleys... The experiment, too, is on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and in most cases, without their knowledge; one group supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity." Thereby, John Snow was able to prove that the deaths were concentrated around a water pump in Broad Street (upstream from the *Southwark and Vauxhall company*, but downstream from the *Lambeth water company*). After the pump was shut down by removing its handle, the epidemic came to a halt. Interestingly, his theory was never accepted by scientists and doctors at the time and was only confirmed several years after his death (Fowke, 1885).

2.1.5 Limitations

The IV approach always rests on the validity of instruments found by researchers. This validity may, depending on the research design, be challenged on various grounds. First, if subjects are not randomly assigned to treatment, there may be doubts regarding the independence assumption. This is especially relevant in observational studies - as (Dunning, 2008) points out, instrumental variables may be classified along a spectrum ranging from "plausibly random" to "less plausibly random":

Also, it is often hard for empirical researchers to find valid instruments that strongly affect treatment, are independent of unmeasured confounders and affect the outcome only through its effect on the treatment.
If researchers resort to using IVs that are only weakly correlated with treatment status (so called weak instruments), it has been shown that estimates will have large standard errors, might be inconsistent and even biased in the same direction as OLS as the power goes towards 0 (Bound et al., 1995). In recent simulation studies, it has even been shown that IV reduces bias

as compared to OLS in ideal circumstances only - also, small sample sizes adversely affect the variance of the distribution of estimation errors which is compounded when the instrument is weak (Crown et al., 2011; Gennetian et al., 2005). As was mentioned in chapter 2.1.2.1, even when all of the above practical problems can be ruled out, IV is still only able to estimate treatment effects for compliers (those subjects who would take the treatment if encouraged to do so by the IV and not take the treatment if not encouraged). Keeping in mind that this particular subgroup cannot be identified from data (as only one of two counterfactual outcomes can be observed), questions about the usefulness of IV estimates have been raised.

2.2 Implementation of altered provider incentives for a more individual-risk-based assignment of dental recall intervals: evidence from a health systems reform in Denmark using Interrupted Time Series Analysis[1]

The second case study examines the impacts of 2015 regulatory changes in Danish dental care which aimed at effectuating a transition from six-to-twelve-monthly dental recall intervals, for every patient, towards a model where patients with higher need receive dental recalls systematically more frequently than patients with lower need. The implementation of this reform constitutes a unique natural experiment that allows the derivation of causal effects using “Interrupted Time Series Analysis” (ITS).

2.2.1 Introduction

In Denmark, dental care for adults is usually provided by private dental practitioners. Dental care expenses are partly covered by self-payment and from general taxation financed payments from the National Health Insurance. All adult citizens are eligible for compensation. For persons under the age of 18, dental care is provided in public dental clinics financed by general

[1] The corresponding paper has been published in Health Economics.

taxation and without additional out-of-pocket expenses (Danish Health Act, 2018). According to WHO criteria, the Scandinavian countries belong to the so-called very low and low-caries prevalence countries (Petersen, 2003). The use of dental services is comparatively high in these countries, with 64and 77Denmark are paid using the fee-for-service payment model in which each item of treatment is paid for separately, giving an incentive for dentists to provide more treatments because payment is dependent on the quantity, rather than quality of care.

In 2013, The Danish Health Authority issued new guidelines for dental recall intervals. From April 1, 2015, a new collective agreement was negotiated between the Danish Regions and the Danish Dental Association, incorporating the 2013 guidelines (Regionernes Lønningsog Takstnævn, 2014). The collective agreement describes the dental services delivered in adult dental care and sets the level of remuneration paid from the Danish National Health Insurance. In this paper, this is designated as the "2015 reform". Since then, dentists have been required to risk-classify their patients into three distinct classes according to their current oral health status and the assessed risk of future oral disease. Healthy patients (free from active oral disease and free from risk factors for future oral disease) should be categorized as "green", at-risk patients (active oral-disease and/or presence of risk factors for oral disease which are modifiable, for instance poor oral hygiene) should be categorized as "yellow" and high-risk patients (active oral disease and/or risk factors for oral disease, which are not modifiable, for instance chronic general disease with known influence on oral health) should be categorized as "red". The recommended dental recall intervals vary across these risk-groups. Patients categorized as either "yellow" or "red" are advised to attend for check-ups more frequently while healthy patients are incentivized to attend for check-ups less frequently (Figure 2.3). Additionally, in part, the risk classification determines which treatments can be remunerated. Most notably, remunerating "Individual Preventive Treatment (IPT)" in diagnostic check-ups is now restricted to patients characterized as either yellow or red. Also, claiming remuneration for newly created codes concerning "focused

examination (FE)" is only possible if patients are classified as yellow or red. This way, at-risk patients should both undergo a more thorough treatment (through IPT) and visit the dentist more frequently (through FE). The "Status Examination (SE)" is to be performed regularly (every 12-24 months) for all patients. Further details of the Danish treatment approach are shown in Figure 2.3 and in Table 6.3. From the dentists' perspectives, the reform was anticipated as likely to result in reductions in revenues from treating patients in the low risk group (green category) but in increases in earnings from treating patients in higher risk groups for whom provision of preventive care (IPT) during follow-up examinations became mandatory. Yet dentists who exceeded a maximum limit of health insurance reimbursements were also subject to restitution of payments exceeding the respective threshold. Unique administrative data with patient-level information on the services provided by dental practitioners in Denmark over a 5-year period from 2012 to 2016 were obtained. The data comprise all treatment claims achieved in the Danish National Health Insurance database between 2012 and 2016. Usable variables included the treatment performed, patient age and sex, date of treatment, municipality of both patient and dental practice as well as cost of treatment. In total, 72,155,539 claims from a total of 3,759,721 unique patients in 25,533,311 distinct treatment sessions were investigated. A single observation was formed by a claim handed into the health insurance by a dentist. The raw data presented as very homogenous and were not indicative of missing data. Several sense checks to exclude observations with typing errors or missing commas were performed. A detailed description of the raw dataset and variables can be found in the Appendix (Table 6.3). The data used were pseudonymized in accordance with Danish jurisdiction and no ethical clearance was required for purposes of this research project.

2.2.2 Basics and Estimation

In the introductory chapter, it was argued that causes always precede their effects temporally, justifying the general preference of data containing a time period over cross-sectional data in the applied literature on causal inference

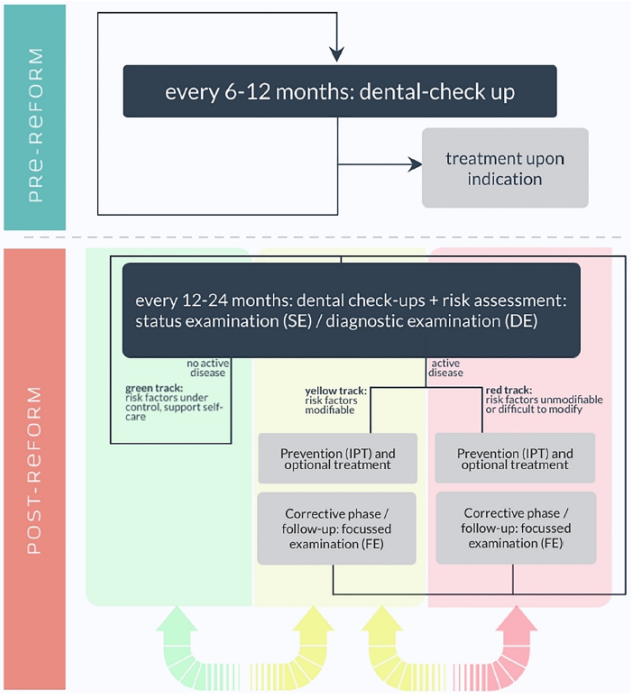

Figure 2.3: Treatment pathways according to risk grouping. Patients can switch between risk groups when their disease status changes. The dental recall intervals are now 12-24 months in all tracks with additional sessions pertaining to yellow and red risk groups.

(Wunsch et al., 2010).

ITS analysis is a quasi-experimental design leveraging the longitudinal nature of data typically used to evaluate the longitudinal effects of interventions using standard regression techniques.

Hereby, a time series is a consecutive sequence of observations on a population, taken repeatedly over time (Shumway, 1988). In ITS studies, a particular time series is "interrupted" by an event at a known, clearly defined point in time. Then, a counterfactual scenario in Rubin's sense can be defined, vindicated by the hypothetical scenario under which the intervention

had not taken place and the trend had continued as before (in other words: the "expected" trend, given the pre-intervention trend). Using this scenario as a counterfactual, one can compare the effect of the intervention by examining any changes occurring in the post-intervention period (Cambell and Stanley, 1963).

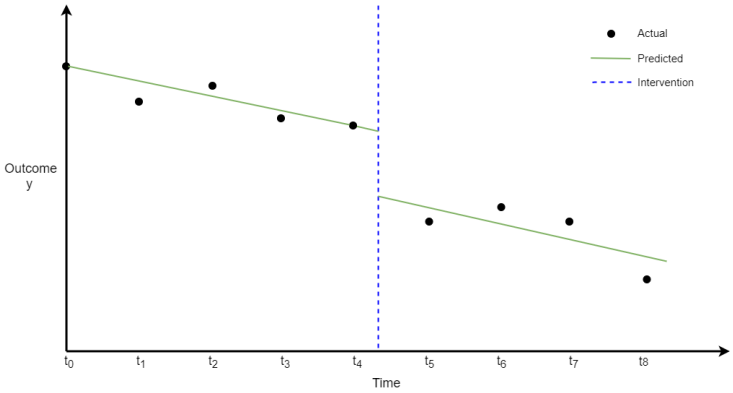

Figure 2.4: Stylized illustration of interrupted time series designs

Figure 2.4 shows a typical ITS scenario where an intervention effectuates a change in the variable of interest. ITS allows for modelling individual effects, possibly including unobservable, time-invariant characteristics which may be correlated with the observable variables. Considering these individual effects becomes especially interesting when there are unmeasured confounders causally affecting the outcome variable and, at the same time, being correlated with observed explanatory variables. If these unmeasured confounders additionally are time-invariant, unbiased estimation using panel-regression is warranted. Depending on assumptions imposed on the individual effects (fixed or stemming from random variation), literature discerns two types of panel data regression (Wooldridge, 2010) - *fixed effects* and *random effects* models.

For the purpose of this thesis, a national reform to the dental system implemented in Denmark is used as a case study. Given the nature of these data, particular considerations are deemed necessary that will be scrutinized

in chapter 2.2. To that end, the utilization of dental services thought to be affected by this reform was analyzed. The dental services under scrutiny were divided into several treatment baskets: preventive, diagnostic, scaling, X-ray, periodontal and fillings. For each of these categories, a binary variable was set to 1 for a particular session (here, one session consists of several treatments, all performed within one particular day) if at least one corresponding code was remunerated in that session. Otherwise, it was set to 0.

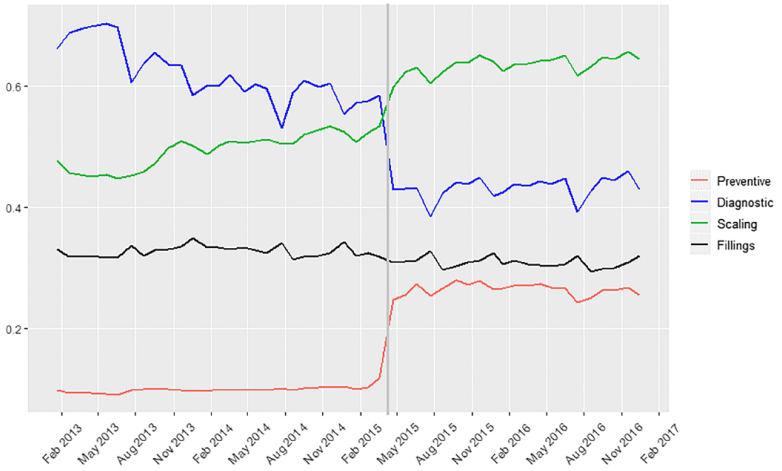

Figure 2.5: Stylized illustration of interrupted time series designs

The identification strategy for the reform in 2015 rests on an interrupted time series design on session level (more technically, estimated using OLS with binary treatment variables, see equation below), comparing the utilization of dental services before and after the introduction of the reform. By binning treatments in different “baskets” and running separate ITS analyses on each of them, a better view of the effects of implementing reforms using individual risk classes is achieved, multiple interrupted time series analyses on utilization of different treatment patterns were performed, which is a valid way of evaluating large-scale implementations in analyzing data known or thought to be affected by interventions. As argued above, if patient characteristics are not fully contained in the explanatory variables, OLS

regression models may be biased. In order to deal with potential problems of this kind, the longitudinal character of our data was exploited by estimating a fixed effects model (on patient level), giving the most complete control for unobserved heterogeneity. The following equations enable capturing both time trends, intervention effects, the interaction of them and covariates. Very generally, ITS can be written as

$$Y_t = \beta_0 + \beta_1 T + \beta_2 X_t + \beta_3 T X_t$$

In the scenario described here where X describes whether a reform is present or not, this amounts to

$$Y_{it} = \beta_1 reform_{it} + \beta_2 (reform \times time)_{it} + \epsilon_{it}$$

Hereby, Y denotes the binary outcome variable at hand at time point T (within an individual i), i.e. whether a particular session contained a treatment code of interest or not. The variables year, gender and municipality constitute categorical control variables and the binary variable reform is 1 if an observation pertains to the period after the reform and 0 otherwise. The variable *(reform × time)* captures possible changes in trend after the reform was put into place.

2.2.3 Assumptions

ITS models are typically linear models estimated using Ordinary Least Squares - the usual assumptions related to OLS apply (see, for example (Angrist and Pischke, 2008)). With the outcome variable above being a binary variable determining the conduct of a particular treatment in a particular session, without loss of generality, a logistic regression model was used. However, other types of regression models have also been used, e.g. linear regression and, for modelling count data, Poisson regression models. Regardless of the distribution of the outcome variable, the longitudinal nature of data usable for ITS requires additional methodological considerations (many properties of the typical regression approaches are shared). (Bernal

et al., 2016) lists these methodological subtleties:

Seasonality Seasonality is characterized by periodic patterns (often representing seasons) in time series data which are typically detected using graphical criteria (Hylleberg, 1992). When seasonality is present, the linearity assumption of OLS is violated, leading to biased estimates. Another related methodological issue is seasonal noise which is both hard to recognize and hard to correct (Sims, 1974).

Time-Varying Confounders ITS models are typically robust with respect to confounders that are constant over time (e.g. education, socioeconomic status etc.). However, confounders that have the potential to change rapidly over the course of time may pose problems in ITS studies (e.g. natural events, risk factors etc.), especially if these confounders are other events targeting the same outcome (e.g. simultaneous changes in reimbursement schemes or treatment code composition).

Use of controls and other more complex ITS designs Enhancements to ITS techniques (in the design stages) have been proposed to mitigate effects of time-varying counfounders. In *Controlled ITS* (Cummins et al., 2018a), different types of controls are used, each of which has associated strengths and limitations. Researchers undertaking controlled ITS should carefully consider a priori what confounding events may exist and whether different controls might be able to exclude these or if they could even introduce other sources of bias to the study. Multiple-baseline designs consist of introducing the reform in different places to different times.

Autocorrelation Conventional models estimated using OLS are only unbiased if the error terms are independent. In time series, errors are often correlated over time (autocorrelated). This does not cause bias, but OLS estimators don't fulfil the minimum variance paradigm anymore, leading to an underestimation of the MSE and standard errors of regression coefficients (Andrews, 1991). Thus, there is a need to model

the time component in regressions as otherwise, the error term would be subject to endogeneity.

2.2.4 Limitations

The most striking limitation of ITS is the non-existence of a real control group which is assumed only implicitly as co-interventions or other effects occuring at the time of intervention cannot be ruled out. A possible countermeasure is the addition of a control series, introducing both a before-after comparison as well as an intervention-control group comparison (Cummins et al., 2018b). In the discussion section of this thesis, the implications of this shortcoming on causal statements deducted from ITS analyses will be noted.
A second limitation is imposed by the need of having multiple pre- and post-intervention observations - naive pre- and post-intervention based on single time points have poor internal validity as they cannot exclude underlying trends as a cause for found changes. The required number of observations to correctly identify trends is debatable, but having at least 5 time points seems to be a sensible middle ground between data availability and robustness of the model (Soumerai et al., 2015). Even then, it is not always clear whether the linearity assumption is even sensible - it is not easy to test and mostly requires a qualitative inspection of time series.
Finally, ITS cannot typically be used to reason about individual effects as data relate to population rates. Although it is tempting to make such inferences, any interpretation needs to be wary of the ecological fallacy of deducing inference about the nature of individuals from statements about the collective they belong to (Penfold and Zhang, 2013; Selvin, 1958). The consequences of these shortcomings and their interaction with causal statements will be subject of discussion in subsequent chapters.

2.3 An evaluation of a multifaceted, local Quality Improvement Framework for long-term conditions in UK primary care using Differences-in-Differences [1]

The third case study evaluates the effects of a local, multifaceted large pay-for-performance scheme in general practice in Stoke-on-Trent introduced in 2009 in the context of the national Quality and Outcomes Framework that operated from 2004. The implementation of this reform constitutes a unique natural experiment that allows the derivation of causal effects using "Differences-in-Differences" (DiD).

2.3.1 Introduction

Stoke-on-Trent is an industrial conurbation with a ceramics, mining and steel heritage and a registered population of 285,000. Of the 326 local authorities in England, Stoke-on-Trent is ranked the 16th most deprived, with large areas in the city ranked among the top 10% most deprived in the whole of England.

Across a range of health and lifestyle indicators, outcomes in Stoke-on-Trent are poor. Male life expectancy at birth in 2012 was 76.5 years compared with 79.4 years in England; female life expectancy was 80.6 and 83.1 years in England (on Trent Clinical Commissioning Group, 2015). As a response to poor health indicators, a local Quality Improvement Framework (QIF) commenced in primary care in 2009. The important context for QIF was that in 2004 as part of a new contract for GPs, the UK introduced a large, national P4P scheme—the Quality and Outcomes Framework (QOF). This article describes the evaluation of the local QIF up to 2015, in the context of the continuing national QOF.

This article describes the evaluation of the local QIF up to 2015, in the context of the continuing national QOF.

In the analyses, seven indicators relevant to the long-term conditions included within the QIF analysed. For each indicator, Stoke-on-Trent was

[1] The corresponding paper has been published in Family Practice.

compared with (i) national, (ii) regional (West Midlands) and (iii) a basket of localities with similar population demographics and other characteristics relevant to the determinants of health (peer localities).

In 2004, as part of a new contract for GPs, the UK government introduced a pay-for-performance scheme with 136 indicators. The population included in the indicators is defined by practice-based disease registers [e.g. patients with coronary heart disease (CHD)] and the indicator measures the achievement of evidence-based targets (e.g. "the percentage of patients with coronary heart disease in whom the last blood pressure reading measured in the preceding 12 months is 150/90 mmHg or less""). The indicators covered the management of chronic disease, practice organization and patients' experiences with respect to care. Electronic clinical records, which were already used in many practices, became universal because they were needed to support payment for work undertaken, though GPs employed more administrative staff to collect the required data, and there was an acceleration of existing trends to shift care for chronic physical conditions to nurse-led clinics. Practices required more intensive internal and external management support to ensure they achieved the targets. Periodic revisions to the scheme added or removed indicators and topics depending on local priorities. Payments make up 25% of general practice income, and 99.6% of general practices participated in the scheme, which remains voluntary. The scheme continues in England but has been replaced in Scotland, Wales and Northern Ireland.

The quality improvement approach used in the local Quality Improvement Framework The team leading the local QIF programme in Stoke-on-Trent designed and delivered a wide-ranging approach to quality improvement in all practices. The QIF had a local implementation strategy, which is a close fit with the evidence on the best approaches to spread good practice (Greenhalgh et al., 2004a).

The aims were to identify patients with long-term conditions currently undiagnosed, to improve the management and treatment of people with those conditions and to reduce health inequalities both within localities in the

city and between the Stoke-on-Trent population and other areas in England. The QIF was much more than a pay-for-performance scheme; a multifaceted design included data feedback on achievement of locally agreed chronic disease management standards, and an educational programme comprising (i) individual support as bursaries, (ii) multidisciplinary learning events for primary care teams and (iii) QIF-focussed practice visits from clinical leaders and managers to encourage sharing of approaches between practices (Cox, 2012).

Pre-requisites for annual review of acceptance of each practice into the QIF programme included thresholds for numbers of registered patients per whole time equivalent practice clinicians, prevalence rates for specific long-term conditions versus those expected, minimum QOF attainment of clinical indicators, completion of clinical audits and progress with addressing clinical indicators of unwarranted clinical variation. All of these were designed to, and became, more challenging over time.

A panel of local stakeholders including patients was convened each year to review attainment of progress with existing QIF indicators. Quality improvement support was individualized to each practice with annual practice-related comparative reports covering 50 key indicators. These included the practice attainment in addressing adverse lifestyle issues such as smoking cessation quit rates, conversion rates for urgent cancer referrals, location of diagnosis of cancer, as well as comparison with peer practice populations and England average rates. Each year the practices that generated most concerns about attainment of the quality indicators were visited by the QIF team who agreed a regularly monitored development plan.

Practice income derived from the QIF was supplementary to practice's funding derived from national contracts. Payments were set at £6 per patient (an additional 4.4% of average, gross practitioner income) if all standards were achieved and gradually less if only part of the standards were achieved. The patient population registered with the 55 general practices in Stoke-on-Trent was 265000 in 2009. Over the first 7 years of the QIF scheme, there was 100% participation of all practices; this includes on average three practices each year that failed to match the pre-requisite criteria for participation

at the start of the year—for example, being able to meet data-availability criteria. All such practices remained engaged via quality and performance development in order to achieve the criteria for participation the following year, but did not receive in-year direct funding for that year.

2.3.2 Basics and Estimation

For the analyses of the QIF and QOF, Difference-in-differences analyses are used. Thereby, the longitudinal character of data is utilized to obtain valid treatment and control groups whose outcomes before and after an intervention has taken place is compared to obtain causal estimates of the intervention.

It relies on scenarios where both treated and untreated units are observed during an intervention of interest. In case both groups follow the same time trends (where only the treated group is exposed to the intervention), one can compare the outcomes of the groups before and after the intervention. This approach allows the separation of factors that affect both groups from the actual effects of the intervention. Figure 2.6 displays the general strategy.

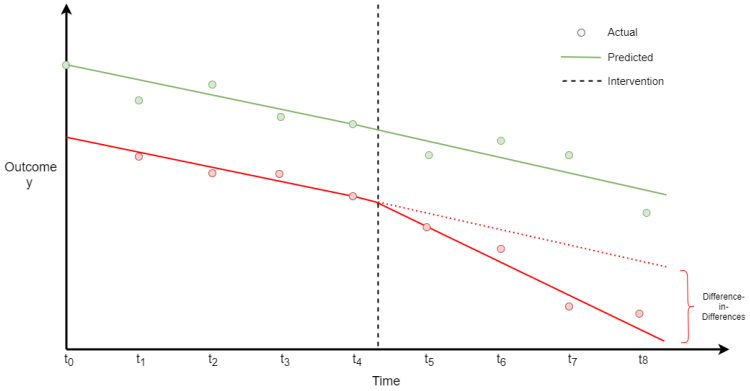

Figure 2.6: Stylized illustration of a Difference-in-Differences designs

Under certain assumptions that will be discussed in section 2.3.2.1, counfounders are then eliminated by design. DiD is most typically used in sce-

narios “at scale” such as the impact of the passage of laws or the enactment of public interventions.
In order to show the basic methodological concept, a minimal example displaying in what kind of setting DiD is applicable and sensible is introduced in the following. This example is entirely non-parametric - empirical literature often imposes a linear model which comes with different identifying assumptions which will be covered in the chapter about assumptions and our case study.
Let D be a binary treatment variable with realisations $\{0,1\}$ and a binary temporal variable T with realisations $t \in \{0,1\}$. $t = 0$ represents the pre-treatment period (the outcome at some specific time therein, an average of outcomes at several pre-treatment time points) and $t = 1$ represents the post-treatment period (or, again, the outcome at some specific time therein, an average of outcomes at several pre-treatment time points). y represents the outcome variable. Then, the quantity of interest is the difference in differences

$$\hat{\delta} = \big(y(T=1,D=1) - y(T=1,D=0)\big) - \big(y(T=0,D=1) - y(T=0,D=0)\big)$$

Albeit impossible, if the two groups are exactly identical, this exactly corresponds to the counterfactual in Rubin’s sense. When the two groups are just similar in some aspects that will be discussed below, the average treatment effect can be estimated.

Hereby, groups are often identified by geography and time period. However, other scenarios are also conceivable such as product categories and geography, age groups and geography or age groups and time. The Difference-in-Difference (DiD) approach exploits natural experiments and is predominantly used in social sciences and empirical economics, where it is used to make statements about the effects of policy interventions or changes that are not assumed to affect everybody at the same time and in the same way. Figure 2.6 displays a very simple, generic scenario where the outcome of two comparable groups follows the same trend. The decisive part is that it is credible to deal with "group non-equivalence" through differencing (which,

again, assumes that the two groups are comparable up to a constant). Thus, the design compares de facto four different groups of objects (post-treatment treated, pre-treatment treated, post-treatment nontreated, pre-treatment nontreated) where three of these groups are not affected by the treatment. The empirical idea of DiD rests on the assumption that the two treated and non-treated follow the same time trends, but only the intervention has an effect on the treated group. Then, the additional difference between treated and non-treated after the intervention took place can be used to remove the effect of confounding factors to which a comparison of post-treatment outcomes of treated and non-treated may be subject to.
DiD analysis is problematic if there are confounders affecting treated and non-treated differently. This can be remedied persuasively by employing regression-based DiD analyses as confounding influences can then be controlled for. In the following, the empirical strategy used to analyze the impact of quality improvement frameworks in the NHS in Stoke-on-Trent will be exemplified.

For DiD analyses, directly standardized mortality rates from a total of 326 local authorities in England were used. Data were available as three-year-rolling averages. The time frame consisted of yearly observations from 1995 to 2013, totalling a balanced panel of 5542 observations for each condition without any missing data.

The following key conditions were analysed: CHD, stroke, diabetes, chronic obstructive pulmonary disease (COPD), asthma, chronic kidney disease (CKD) and epilepsy. Data from four age bands (all age groups, <65 years, <75 years and 65–74 years) were available for several of the above conditions, helping to increase the validity of analyses.

To determine any impact of the 2009 QIF and the 2004 QOF, a DiD setup was used - in particular, estimating a fixed-effects linear regression model with an interaction effect and a linear time trend (Dimick and Ryan, 2014):

$$y_{it} = \beta_0 + \beta_1(year) + \delta(place_{treatment} \times after_{treatment})_{it} + \gamma_i + \epsilon_{it} \quad (2.1)$$

The coefficient of interest is δ, representing the effect of being in the treatment group (Stoke-on-Trent) after the treatment went into place (2004/2009). Here, $place_{treatment}$ and $after_{treatment}$ are the corresponding categorical variables. γ_i is a one-hot-encoded variable representing a regional fixed effect, and ϵ_{it} is an error term. Note that this approach also enables to check for an effect of the 2004 national QOF in Stoke-on-Trent; if there is no effect in Stoke-on-Trent, the interaction effect is insignificant.

This obtained coefficient is only valid under the common trend assumption (see next section). In the absence of this assumption, the above model yields biased estimates. Since this was the case for several of the models, this problem was circumvented by merging four respective pre- and post-treatment years, looking at the mean number of yearly deaths pre- and post-treatment. This procedure is indicated in the results by an asterisk in the corresponding tables.

To test for significant changes in mortality rates following the 2004 national QOF, an interrupted time-series regression with a linear time trend was used:

$$y_t = \beta_0 + \beta_1(year) + \delta(after_{treatment})_t + \epsilon_{it} \tag{2.2}$$

The variables have similar meanings as described earlier; however, as a national reform with no differential local implementation is dealt with, there is neither a possible comparison of treated and untreated authorities nor a possibility of applying authority-level fixed effects. The validity of this secondary approach rests on the assumption that the slope, had there been no reform, would have continued to follow the same slope.

In order to account for eventual correlation in the data, authority-level cluster-robust standard errors were used in all regressions. Statistical analyses were performed with R version 3.2.1.

To determine any impact of the 2009 QIF and the 2004 QOF, a differences-in-differences setup was used, estimating a fixed-effects linear regression

model with an interaction effect and a linear time trend (5):

$$y_{it} = \beta_0 + \beta_1(year) + \delta(place_{treatment} \times after_{treatment})_{it} + \gamma_i + \epsilon_{it} \quad (2.3)$$

The coefficient of interest is δ, representing the effect of being in the treatment group (Stoke-on-Trent) after the treatment went into place (2004/2009). Here, $place_{treatment}$ and $after_{treatment}$ are the corresponding dummy variables. γ_i is a one-hot-encoded variable representing a regional fixed effect, and ϵ_{it} is an error term. Note that this approach also enables to check for an effect of the 2004 national QOF in Stoke-on-Trent; if there is no effect in Stoke-on-Trent, the interaction effect is insignificant.

This obtained coefficient is only valid under the parallel slopes assumption. In the absence of this assumption, the above model yields biased estimates. Since this was the case for several of our models, this problem was circumvented by merging four respective pre- and post-treatment years, looking at the mean number of yearly deaths pre- and post-treatment. This procedure is indicated in the results by an asterisk in the corresponding tables.

To test for significant changes in mortality rates following the 2004 national QOF, an interrupted time-series regression with a linear time trend was used:

$$y_t = \beta_0 + \beta_1(year) + \delta(after_{treatment})_t + \epsilon_{it} \quad (2.4)$$

The variables have similar meanings as described earlier; however, as the reform is national with no differential local implementation, there is neither a possible comparison of treated and untreated authorities nor a possibility of applying authority-level fixed effects. The validity of this secondary approach rests on the assumption that the slope, had there been no reform, would have continued to follow the same slope.

In order to account for eventual correlation in the data, authority-level cluster-robust standard errors were used in all regressions. Statistical analy-

ses were performed with R version 3.2.1.

2.3.2.1 Assumptions

The validity of the approach to identify causal effects as outlined above rests on a number of assumptions. First and foremost, all assumptions impeding OLS also apply to DiD. As briefly discussed above, the identification strategy of DiD consists of comparing differences in average pre-treatment outcomes between treatment and control groups with differences in average post-treatment outcomes before and after a treatment has been performed or a reform has taken place. Under a certain set of assumptions that will now be discussed, this strategy will identify an average causal effect by mimicing an actual experiment.

Common Trend (CT) The Common Trend Assumption is the most crucial one of the DiD approach. It is assumed that in absence of treatment the difference between control and treatment groups would be constant or "fixed" over time. This assumption states that the differences in the expected potential non-treatment outcomes over time (conditional on X) are unrelated to belonging to the treated or control group in the post-treatment period. When dealing with linear models, this can be represented geometrically by "parallel trends"" in outcome levels between treatment and control groups in absence of a treatment. It has been argued above already that the common trend assumption essentially ensures exchangeability and thereby allows estimation of the ATE in Rubin's sense.

Stable Unit Treatment Value Assumption (SUTVA) This assumption actually consists of **non-interference** and **stability**. Non-interference means that treatment status of one subject of study does not exert influence on the treatment status of other subjects of study. Treatment stability means that treatment is the same for all subjects under study (for example, no patients receive a different dose). In most cases, SUTVA is fulfilled by design - especially in experimental studies. How-

ever, for example, if a researcher aims to analyze the ramifications of vaccinating people in a geographically close environment, SUTVA may be violated. In the discussion section, strategies will be enumerated on how to counter possible violations.

Exogeneity/Ignorability (EXO) Usually, DiD models are interpreted using linear dependencies between independent variables and dependent variables. Thereby estimated using standard OLS techniques and allowing the use of covariates - exogeneity, as often, assumes that the components of covariates X are not influenced by the treatment or are independent of treatment assignment.

In the following, limitations that may arise from these assumptions are discussed.

2.3.3 Limitations

There are a number of issues possibly arising that lead to a biased DiD estimate $\hat{\delta}_{DD}$. Checking for these deviations is often difficult and sometimes impossible as they are made about unobservable quantities.
In particular, the conventional DiD estimator requires that, in absence of treatment, the average outcomes for the treated and control groups follow parallel paths over time. If pre-treatment characteristics that are thought to be associated with the outcome variable are unbalanced between treated and untreated, this assumption may be implausible. This would happen, for example, if selection for treatment (for each individual) is influenced by past outcomes. (Abadie, 2005) proposed a semi-parametric estimators mitigating the effects of non-parallel trends, allowing a consistent estimation of the ATOT.
Also, limitations applying to randomized controlled trials in general such as non-compliance, results only relating to limited populations (e.g. convenience sampling) or non-blinded participants (Deaton and Cartwright, 2018), (Krauss, 2018) can also be translated to DiD and predominantly jeopardize the external validity of results arising from DiD studies.

2.4 A word on panel data

In this piece of work, most data used entails a time component (particularly case studies using ITS and DiD), i.e. units are observed over time. This temporal nature of data can be exploited by using panel regression methods. The two panel regression approaches described in the literature vary from each other with respect to assumptions related to the "individual effects", where fixed effects are constant, time-invariant attributes of individuals and random effects are stochastic attributes of individuals stemming from a probability distribution (where these variables are uncorrelated with other explanatory variables).
In presence of a balanced longitudinal dataset of N units and T time periods, a simple linear *unobserved effects* regression model in the following way can be parametrized (for each unit i at time t, the outcome variable Y_{it} and the binary treatment variable $X_{it} \in \{0, 1\}$) is observed:

$$Y_{it} = \alpha_i + \beta X_{it} + \epsilon_{it}$$

Here, Y_{it} is the dependent variable, α_i are the unobservable group-specific effects for unit i, X_{it} denotes the $K \times 1$ column vector of explanatory variables and ϵ_{it} is a disturbance term for unit i at time t with $\mathbb{E}(\epsilon_{it})$. In this model, the unit fixed effect α_i can be written as $\alpha_i)h(U_i)$, where U_i is a vector of unobserved time-invariant confounders within a certain unit i and h is an unknown function.
The unobservable coefficients α_i determine the type of model - in case they are group-specific (where group often equals individuals) fixed quantity, then fixed effects is the correct estimation procedure, whereas in case they are drawn from an (unknown) distribution and are therefore stochastic.

In the case studies examined in this dissertation, it is assumed that unobserved heterogeneity within individuals is constant, justifying the use of fixed effect model estimation. This line of thought will be justified individually in the methods section of the respective case studies.

Nevertheless, limitations also apply to panel data analysis, as selection bias

might still occur and the implicit assumption of fixed effects (confounders don't vary over time) might be violated (Hsiao, 2014).

Chapter 3

Results of case studies

After having discussed the methodological approaches to causal inference using observational data (while introducing three case studies from the research done during my time as a PhD student), this chapter subsequently introduces readers to the results thereof. In the first section, causal relationships between the dental health of parents as measured by the number of remaining teeth and the number of their children using **Instrumental variables** are explored.

The second section and third section will be concerned with investigating the ramifications of reforms of the Health system in two European countries. First, the impact of the introduction of the "Quality Improvement Framework" in Stoke-on-Trent on mortality rates as compared to peer regions unaffected by the reform using a **Differences-in-differences** framework is analyzed. Second, the impact of the introduction of a patient-risk-classification system in Denmark, aimed at improving the alignment of need and supply, on the utilization of dental services using an **Interrupted Time Series** design will be scrutinized.

In the greater context of this work, this section provides the contextual background for embedding the scientific argumentations within the frameworks

of causal inference introduced in the above chapter.

3.1 Gain a child, lose a tooth? Using natural experiments to distinguish between fact and fiction

The mean age of the individuals in the sample is 67.4 years, and on average, 10.4 teeth were reported missing at the time of the interview. On average, the birth of the youngest child happened 35.5 years ago. Thus, the majority of individuals in the sample have completed their fertile period, allowing to examine the long-term effects of childbearing on oral health.

Table 3.1 presents average numbers of missing teeth by the various independent variables used in subsequent analyses. The average number of missing teeth mostly differed by age (values ranging from an average of 6.8 missing teeth for women aged 50–65 years to an average of 19.2 missing teeth for men aged 80+) and educational attainment (on average 6.3 missing teeth for women with post-secondary education; on average of 15.2 missing teeth for women with (pre-)primary education) and less notably by other independent variables.

3.1.0.1 Results from regression analyses

The upper part of table 3.2 shows the results from OLS and intention to treat regressions. The OLS estimates indicate that women have an average of 0.57 (95%-CI: 0.45 to 0.69) fewer teeth per additional child; men have an average of 0.26 (95%-CI: 0.12 to 0.40) fewer teeth per additional child. The intention to treat estimates indicate that women have an average of 0.36 (95%-CI: 0.11 to 0.60) fewer teeth if their first two children had the same sex instead of different sexes; women have an average of 0.88 (95%-CI: -0.25 to 2.02)fewer teeth if they gave birth to multiples rather than a singleton; men have an average of 0.24 (95%-CI: -0.53 to 0.05) fewer teeth if their first two children had the same sex instead of different sexes; and men have an average of 0.01 (95%-CI: -1.26 to 1.23) fewer teeth if they had a multiple birth instead of the birth of a singleton.

Table 3.1: Mean number of missing natural teeth by covariates

	Number of missing teeth mean (std.dev.)		
	Women	Men	% of sample
Age			
50 to 65 years old	6.8 (8.6)	6.9 (8.6)	0.47
66 to 80 years old	12.4 (10.7)	11.5 (10.5)	0.42
81 years and older	16.9 (10.8)	19.2 (10.5)	0.11
Sex			
Women	10.7 (10.6)		0.57
Men		10.1 (10.3)	0.43
Educational attainment			
(Pre-)primary (ISCED 0 and 1)	15.2 (11.0)	13.3 (10.9)	0.22
Secondary (ISCED 2 and 3)	10.6 (10.5)	10.8 (10.3)	0.51
Post-secondary (ISCED 4 and 5)	6.3 (8.5)	7.1(9.0)	0.27
Age at first birth			
Up to 25 years old	11.2 (10.7)	10.8 (10.4)	0.65
26 to 30 years old	9.2 (10.3)	9.7 (10.2)	0.26
31 years and older	8.3 (10.2)	8.9 (10.2)	0.09
Number of children			
Two children	9.7(10.2)	9.6 (10.1)	0.56
Three or more children	11.8 (11.0)	10.7 (10.5)	0.44
Number of siblings			
No siblings	13.1 (11.1)	11.6 (10.6)	0.17
One sibling	10.1 (10.6)	9.9 (10.3)	0.28
Two siblings	9.8 (10.4)	9.4 (10.0)	0.22
Three or more siblings	10.4 (10.4)	9.9 (10.2)	0.33
Ever had a multiple birth			
Yes	10.9 (11.0)	9.9 (10.4)	0.04
No	10.7 (10.6)	10.1 (10.3)	0.96
First two children			
Same Sex	10.5 (10.6)	10.2 (10.3)	0.50
Different Sexes	10.8 (10.7)	10.0 (10.3)	0.50
observations	19 970	14 873	34 843

Table 3.2: Results of regression analysis of the number of children on oral health

	Women		Men	
OLS	0.57		0.26	
	Same-sex instrument	Multiple-birth instrument	Same-sex instrument	Multiple-birth instrument
ITT	0.36	0.88	-0.25	-0.01
	[0.11; 0.60]	[-0.25; 2.02]	[-0.53; 0.05]	[-1.26; 1.23]
1^{st} stage (2SLS)	0.084	1.04	0.076	1.15
	($F = 31.69$)	($F = 235.50$)	($F = 20.20$)	($F = 251.53$)
2^{nd} stage (2SLS)	4.27	0.85	-3.12	-0.01
	[1.08; 7.46]	[-0.23; 1.93]	[-7.17; 0.93]	[-1.09; 1.07]
observations	19 970		14 873	

The lower part of Table 3.2 shows the results from 2SLS regressions:

1st stage: on average, individuals whose first two children have the same sex have 0.084 (women) and 0.076 (men) more children than individuals whose first two children have different sexes. The F-statistics relate to the statistical significance of the instruments. With values of 31.69 and 20.20, respectively, they are higher than the rule-of-thumb critical value of 10 (Bound et al., 1995), indicating sufficient strength of the same-sex instrument. Multiple as compared to singleton births cause a change in the number of children with parents having coefficients of 1.04 (women) and 1.15 (men). This means that individuals hardly compensate for additional children due to twin births by reducing subsequent fertility. Again, the F-statistics of the multiple birth instruments indicate sufficient strength.

2nd stage: owing to identification via smaller subsamples of the study population, confidence intervals are generally wider than in OLS regressions. An additional child caused by the first two children having the same

sex has a large causal effect on the number of teeth in women; these women have an average of 4.27 [95%-CI: 1.08; 7.46] fewer teeth than women who did not have another child while their first two children have different sexes. For men, the same sex instrument indicates a smaller effect of 3.12 [95%-CI: -7.17; 0.93] more teeth. Having one additional child in response to having twins at second birth leads to an average of 0.85 [95%-CI: -0.23; 1.93] fewer teeth among women. For men, the point estimate is small, suggesting no causal effect of multiple births on the number of teeth in men.

3.2 Implementation of altered provider incentives for a more individual-risk-based assignment of dental recall intervals: evidence from a health systems reform in Denmark

The following section deals with results gained from ITS analyses on the 2015 reform to the Danish dental health care system.

Table 3.3 shows summary statistics for dependent and explanatory variables. The mean age of patients was 49.4 years. Overall, in most sessions, scaling and diagnostic codes were utilized. Fillings and periodontal treatments were performed in 32% and 22% of sessions, respectively, while the preventive code was used in about 15% of all sessions. Radiographs were taken in about every 6th session. Table 3.4 shows descriptive statistics regarding changes in dental utilization and recall interval characteristics before and after the reform. The average number of dentist visits and recalls per patient was shown to have increased slightly. The average number of days between dental visits per patient reduced by about seven days and that of dental recalls reduced by about six days. The proportion of patients visiting the dentist every 6 months or more often increased by about 0.9%, whereas the proportion of patients with 6-12-monthly or more than 12-monthly dental visits decreased somewhat. In comparison to the pre-reform period, the proportion of patients with recall intervals of up to 6 months was by 1.2%-points larger post-implementation; that of patients with 6-12-monthly

Table 3.3: Summary statistics of dependent and independent variables.

sample size 25,533,311 claims	description	mean(std.dev.)
age	in years	49.4 (16.9)
sex	Women/men in percent	53.8 / 46.2
total number of sessions	mean number of yearly sessions	2.41 (0.74)
preventive (IPT)	1 if a preventive code was remunerated in a session	0.15
diagnostic (SE, DE, FE)	1 if a diagnostic code was remunerated in a session	0.75
scaling	1 if scaling was remunerated in a session	0.58
fillings	1 if fillings were remunerated in a session	0.32
periodontal	1 if periodontal codes were remunerated in a session	0.22
radiographs	1 if x-rays were remunerated in a session	0.15
surgical treatments	1 if operations were remunerated in a session	0.07

A detailed description of the presented variables can be found in the Appendix (Table A2).

recalls increased by 0.7%-points; that of patients with more than 12-monthly dental recalls decreased by 1.9%-points.

Figure 3.1 illustrates the trajectories of treatment codes over time for the years 2012-2016. The proportion of treatment sessions including preventive items or scaling is shown to have become larger after the regulatory changes in April 2015; the proportion of diagnostic items was shown to have become lower. Table 3.5 displays results from regression analyses on the utilization of certain treatment codes represented by independent, mostly binary variables (except for "total number of sessions" which is a count variable) following the 2015 reform. The second column contains point estimates from OLS regressions. A total of 5,420,552 sessions were included in this regression. The total number of sessions did only change marginally (point estimate: -0.003 [95%-Confidence Interval: -0.005; -0.002] as compared to the pre-reform baseline). The proportion of sessions containing preventive codes and the proportion of sessions with scaling increased by 0.301

[0.300, 0.302] %-points and 0.225 [0.224; 0.226] %-points, respectively. At the same time, fewer sessions contained diagnostic codes (-0.298 [-0.299; 0.297] %-points). The two columns on the right of Table 4 show parameter estimates from fixed-effects regressions when considering both the 12 months before and after the reform (third column) and, as a robustness check, both 2014 and 2016 excluding 2015 (fourth column). In the third column (12 months pre/after reform), the results indicate an increase in the proportion of sessions including the preventive code by about a third (0.310 [95%-Confidence-Interval: 0.309; 0.311] %-points).

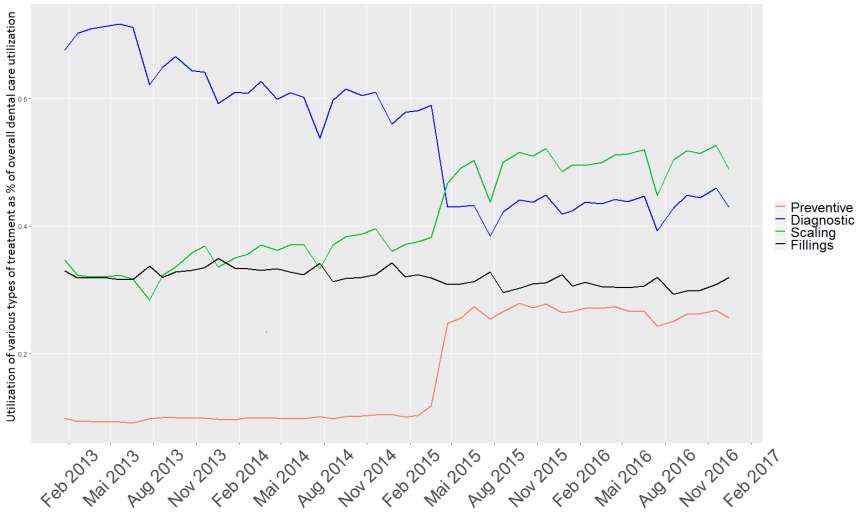

Figure 3.1: Trajectories of the proportion of treatment claims containing codes for preventive care, diagnostic care, scaling, and filling treatment sessions taking place in the period of 2012 to 2017. The vertical grey line depicts April 1, 2015, when the reform commenced.

The proportion of diagnostic sessions decreased by 0.345 [-0.346; -0.344] %-points and scaling experienced a sizable increase as well (0.241 [0.240; 0.242] %-points). Treatment codes apart from preventive and diagnostic codes and scaling did not exhibit large variations: fillings slightly decreased while the number of radiographs remained relatively stable. Results in the fourth column (19 months pre-/post-reform) are similar, with exceptions being different signs for the total number of sessions and periodontal treatment

Table 3.4: Summary statistics of the frequency of dental visits recalls before and after the refor

sample size: 3,759,721 unique patients	beform the 2015 reform	after the 2015 reform
Average number of dental visits per patient (per year)	1.83	1.85
Average number of dental recalls* per patient (per year)	1.29	1.30
Average number of days between dental visits (per patient)	148.5	141.6
Average number of days between dental recalls* (per patient)	230.6	225.3
% of patients with a 6-month interval between dental visits (or more frequently)	20.4	21.3
% of patients with a 6-month interval between dental recalls* (or more frequently)	16.3	17.5
% of patients with a 6-12-month interval between dental visits	33.4	32.6
% of patients with a 6-12-month interval between dental recalls	36.1	36.8
% of patients with a more than 12-month interval between dental visits	46.2	46.1
% of patients with a more than 12-month interval between dental recalls	47.6	45.7

Observations: 5,420,552 sessions

*dental recalls were defined as treatment sessions which only included only SE and/or DE (see Figure 1). Patients with 0 visits to the dentist were not considered as their recall intervals can't be recovered with claims-data based observations.

Table 3.5: OLS and Fixed Effects regression results for effects of the 2015 reform

	OLS	FE (Patient Level)	
		12 months pre vs. 12 months post	19 months pre vs. 19 months post
total number of sessions	-0.003	-0.012	0.018
	[-0.005; -0.002]	[-0.014; -0.010]	[0.016; 0.020]
Preventive	0.301	0.310	0.315
	[0.300; 0.302]	[0.309; 0.311]	[0.314; 0.316]
Diagnostic	-0.298	-0.345	-0.364
	[-0.299; -0.297]	[-0.346; -0.344]	[-0.365; -0.363]
Scaling	0.225	0.241	0.282
	[0.224; 0.226]	[0.240; 0.242]	[0.281; 0.283]
Fillings	-0.029	-0.041	-0.041
	[-0.030; -0.028]	[-0.042; -0.039]	[-0.043; -0.040]
Periodontal treatments	-0.005	-0.032	0.008
	[-0.006; -0.004]	[-0.033; -0.031]	[0.007; 0.009]
Radiographs	-0.008	-0.041	-0.016
	[-0.010; -0.007]	[-0.042; -0.040]	[-0.017; -0.015]
Surgical treatments	-0.003	-0.012	-0.018
	[-0.004; -0.002]	[-0.013; -0.011]	[-0.020; -0.017]
Patient fixed effects	NO	YES	YES
	[-0.004; -0.002]	[-0.013; -0.011]	[-0.020; -0.017]
N (*observations = sessions*)	5,420,552	3,181,824	3,259,848
	[-0.004; -0.002]	[-0.013; -0.011]	[-0.020; -0.017]
Observations	3,181,824 sessions (12 mo.)		3,259,848 sessions (19 mo.)

OLS (second column) and linear individual fixed effects regression (third and fourth column) using the number of sessions containing preventive, diagnostic, scaling, fillings and periodontal codes as well as codes related to surgical procedures and extractions ("surgical treatments") resp. as independent variables. In our OLS model, age, sex and municipality of patients was used as confounders. In our Fixed Effects model, no additional controls were included as no time-varying confounders could be identified (controlling for age did not add explanatory power). 95% Confidence Intervals in brackets.

(relative to preventive, diagnostic and scaling codes, however, these effect sizes are much smaller). By and large, our results are robust to changes in the observation periods and the general tendency of variations in utilization remains. When the sample is split according to various proxies for dental disease risk, the results were mixed. Parameter estimates differed relatively little when differentiating between persons with high vs. low income although effects for codes related to preventive and diagnostic services were stronger for the high-risk group. There was more variation between parameter estimates when differentiating between persons in young vs. old age; again, the estimated effects of regulatory changes on utilization of preventive, diagnostic and scaling items are larger for persons with higher risk (older age) than for persons with lower risk (younger age) but all have the same sign (more prevention and scaling but less diagnostics after regulatory changes). When using previous treatment experience as dental disease risk proxy, the most substantial differences in parameter estimates were found for the total number of treatment sessions (after regulatory changes: fewer sessions for high risk patients; more sessions for low risk patients), scaling (after regulatory changes: nearly four-fold more sessions with scaling for low-risk than for high-risk patients), and fillings (after regulatory changes: fewer fillings for high-risk patients but more fillings for low-risk patients). By and large, the robustness checks neither provide clear evidence in support of, or against treatment patterns having become more risk-oriented in response to regulatory changes.

3.3 An evaluation of a multifaceted, local Quality Improvement Framework for long-term conditions in UK primary care

The following section will deal with the results of an evaluation of a multifaceted, local Quality Improvement Framework for long-term conditions in UK primary care, utilizing DiD.

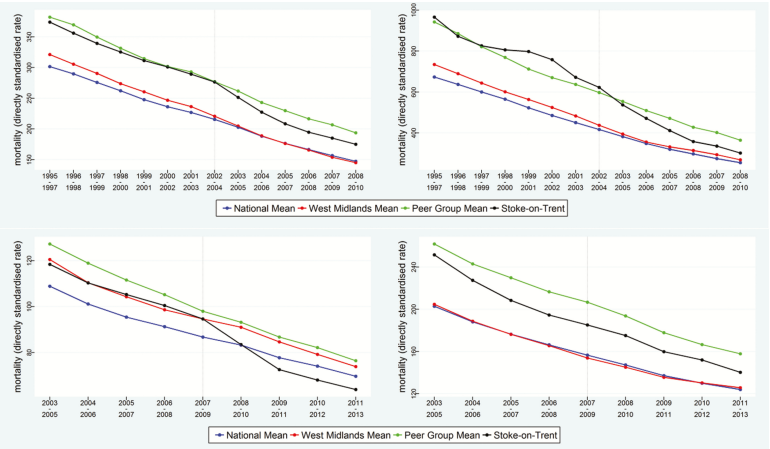

Figure 3.2: Time series of mortality rates from 1995 to 2013 for different geographic localities in England and for the 2004 Quality and Outcomes Framework and 2009 Quality and Outcomes Framework interventions. (a) The dashed black line refers to the 2004 national Quality and Outcomes Framework introduction— mortality time series are for cardiovascular heart disease in all age groups (left) and 65- to 74-year-old age groups (right). (b) The dashed black line refers to the 2009 local Quality Improvement Framework introduction—mortality time series are for stroke (left) and cardiovascular heart disease (right).

Table 3.6 (left panel) displays the mean mortality rates per 100000 people in England before any intervention took effect. It was found that from 2004 the downward trend in the national mean mortality rate for the conditions analysed increased by an additional 3.72 fewer deaths for CHD and 5.53 fewer deaths per 100 000 general population per annum for stroke (see Table 3.6, Figure 3.2). This came in addition to a yearly trend indicating a reduction of 11.07 deaths per 100 000 general population per annum for CHD and 4.37 deaths per 100 000 general population per annum for stroke. In relation to the comparison groups 'national' and 'West Midlands', pre-2004 mean mortality rates in Stoke-on-Trent (Table 3.7, italic) for all the conditions and age bands were considerably higher. The mortality rates of the peer group of local authorities show a mixed picture across conditions and age groups, and are mostly similar to Stoke-on-Trent (see Fig. 3.2).

A statistically significant greater benefit in Stoke-on-Trent on CHD mortality associated in time with the 2004 introduction of the national QOF with an

Table 3.6: Interrupted time series analysis of mortality rates in England following the 2004 Quality and Outcomes Framework introduction based on data from 1998 to 2014

Absolute mortality rate before 2004	Linear time trend [95% confidence interval]	Intervention [95% confidence interval]
CHD		
All age groups	**-11.07**	**-3.72**
256.93	[-11.32, -10.82]	[-5.46, -1.97]
<65 years	**-1.85**	**0.39**
39.45	[-1.92, -1.77]	[-0.11, 0.89]
<75 years	**-5.05**	**-0.03**
97.68	[-0.20, -4.91]	[-0.87, 0.81]
65-74 years	**-29.63**	**-3.22**
544.07	[-30.46, -28.79]	[-8.76, 2.33]
Stroke		
All age groups	**-4.37**	**-5.53**
129.89	[-4.51, -4.22]	[-6.59, -4.47]
<65 years	**-0.42**	**-0.44**
11.19	[-0.45, -0.39]	[-0.71, -0.17]
<75 years	**-1.18**	**-1.40**
28.53	[-1.23, -1.13]	[-1.78, -1.02]
65-74 years	**-6.97**	**-8.75**
161.43	[-7.29, -6.64]	[-11.35, -6.16]
Diabetes		
All age groups	**-0.35**	**0.30**
13.61	[-0.38, -0.32]	[0.09, 0.50]
Epilepsy		
All age groups	**0.02**	**-0.11**
1.80	[0.01, 0.03]	[-0.21, -0.00]
COPD		
All age groups	**-0.60**	**-0.73**
56.48	[-1.29, 0.09]	[-0.82, -0.65]
Asthma		
All age groups	**-0.08**	**-0.09**
2.98	[-0.09, -0.06]	[-0.23, 0.05]
CKD		
All age groups	**-0.04**	**0.17**
3.16	[-0.06, -0.02]	[0.00, 0.33]

additional reduction of 36 deaths per 100 000 general population per annum in Stoke-on-Trent compared with the national mean (see Table 3.7) was found. This effect occurred in all age groups and is especially relevant for the 65- to 74-year-old age group with an excess reduction of 166 deaths per 100 000 per annum in this population group (see Figure 3.2, upper right). When compared with the West Midlands, this effect consistently becomes less (for all age bands, see Figure 3.2, upper left) over time; when compared with peer localities, the reduction marginally fails to be statistically significant for all age groups but is significant for all three age subgroups. For stroke, a significant benefit on mortality for the 65- to 74-year age groups and <75-year age groups (13 fewer deaths per 100 000 per annum for the 65- to 74-year age group; -1 per 100 000 per annum for the <75-year age group) when compared with the national mean and the West Midlands was found. There was also a significant benefit in the 65- to 74-year age group when the comparison group was peer local authorities. Results in the all age group and <65 years showed small but statistically significant increases against the national comparator (see Table 3.7). Analyses of the other conditions show a mixed picture with small reductions in CKD in Stoke-on-Trent and small adverse trends for deaths from diabetes, chronic obstructive kidney disease and asthma. Effects of the introduction of the 2009 Quality Improvement Framework The pre-2009 mean mortality rates were higher in Stoke-on-Trent across all conditions and age bands compared with the regional and national means, with smaller absolute differences than in 2004. With some exceptions, the mortality rates for the conditions analysed were generally lower in Stoke-on-Trent in 2009 than the mean mortality rates of the peer local authorities. Mortality rates for most conditions and age groups showed a clear reduction associated in time with the introduction of the 2009 QIF in Stoke-on-Trent (see Table 3.8). Compared with the national mean, there was an additional reduction of about 9 deaths per 100 000 people for CHD (see Figure 3.2, bottom left) and a reduction of 14 deaths per 100 000 people for stroke (see Figure 3.2, bottom right). This effect remains when compared with the regional mean, but there was no significant difference when compared with the mean of the peer regions. Analyses of other condi-

tions showed a small reduction in mortality from diabetes, asthma and CKD consistent across all comparison group means. On the other hand, epilepsy and COPD showed small increases.

Table 3.7: Interrupted time series analysis of mortality rates in England following the 2004 Quality and Outcomes Framework introduction based on data from 1998 to 2014

	National	West Midlands	Peer localities
	Intervention [95% CI] Mean difference to Stoke (pre 2004)	Intervention [95% CI] Mean difference to Stoke (pre 2004)	Intervention [95% CI] Mean difference to Stoke (pre 2004)
CHD			
All age groups	**-35.85** [-37.87; -33.82] 63.01	**-24.69** [-30.73; -18.65] 53.35	**-13.58** [-28.24; 1.07] -6.95
<65 years	**-14.11** [-14.65; -13.57] 23.60	**-12.69** [-14.36; -11.03] 21.47	**-5.10** [-7.27; -2.93] 2.00
<75 years	**-31.64** [-32.74; -30.55] 48.16	**-27.86** [-31.06; -24.66] 42.65	**-14.69** [-18.63; -10.75] 5.52
65-74 years	**-166.05** [-172.27;-159.84] 236.44	**-144.13** [-161.36;-126.90] 205.02	**-88.25** [-117.63;-58.88] 32.54
Stroke			
All age groups	**3.60** [2.76; 4.44] 5.97	**6.21** [2.73; 9.70] -5.72	**6.12** [-0.01; 12.25] -14.29
<65 years	**0.62** [0.41; 0.83] 1.96	**0.03** [-0.57; 0.63] 1.74	**1.96** [0.87; 3.05] -2.74
<75 years	**-0.96** [-1.27; -0.65] 6.35	**-1.33** [-2.54; -0.12] 4.49	**1.13** [-1.53; 3.79] -3.99
65-74 years	**-13.10** [-15.30; -10.90] 39.98	**-11.78** [-19.90; -3.66] 25.54	**-5.27** [-27.14; 16.62] -13.51
Diabetes			
All age groups	**1.90** [1.63; 2.17] 0.87	**1.39** [0.16; 2.62] -2.00	**1.27** [-0.63; 3.18] -0.54
Epilepsy			
All age groups	**0.06** [-0.02; 0.14] 0.44	**-0.12** [-0.45; 0.20] 0.33	**0.14** [-0.34; 0.63] -0.18
COPD			
All age groups	**3.34** [2.75; 3.93] 21.31	**4.54** [2.37; 6.71] 22.57	**2.27** [-1.25; 5.79] -8.00
Asthma			
All age groups	**2.18** [2.07; 2.28] 0.73	**1.74** [1.33; 2.14] 0.39	**2.66** [2.03; 3.30] 0.16
CKD			
All age groups	**-1.29** [-1.41; -1.16] 2.01	**-1.51** [-1.95; -1.07] 1.14	**-1.21** [-1.74; -0.67] 1.23

Table 3.8: Interrupted time series analysis of mortality rates in England following the 2004 Quality and Outcomes Framework introduction based on data from 1998 to 2014

	National	West Midlands	Peer localities
absolute mortality rate in Stoke pre-intervention	Intervention [95% CI] Mean difference to Stoke (pre 2009)	Intervention [95% CI] Mean difference to Stoke (pre 2009)	Intervention [95% CI] Mean difference to Stoke (pre 2009)
CHD			
All age groups	**-8.85** [-10.11; -7.60]	**-10.77** [-15.47; -6.07]	**4.60** [-1.82; 11.01]
321.52	29.21	31.65	-22.99
<65 years	**-4.98** [-5.36; -4.60]	**-4.90** [-6.40; -3.40]	**-2.32** [-4.72; 0.09]
63.30	12.43	11.59	-1.39
<75 years	**-7.90** [-8.59; -7.20]	**-7.15** [-9.55; -4.74]	**0.06** [-4.46; 4.58]
147.14	21.36	19.32	-7.89
65-74 years	**-30.25** [-34.42; -26.07]	**-24.36** [-37.44; -11.29]	**18.27** [-5.79; 42.34]
789.96	89.84	78.59	-57.76
Stroke			
All age groups	**-13.61** [-14.46; -12.77]	**-10.21** [-12.87; -7.55]	**-6.81** [-14.82; 1.21]
132.56	7.97	-1.22	-8.15
<65 years	**-0.06** [-0.25; 0.14]	**0.59** [0.01; 1.16]	**1.10** [-0.27; 2.47]
13.64	1.87	1.27	-1.44
<75 years	**-3.03** [-3.34; -2.72]	**-1.33** [-2.45; -0.22]	**-0.66** [-2.37; 1.04]
35.41	4.44	4.44	-3.12
65-74 years	**-23.24** [-25.24; -21.25]	**-16.33** [-24.00; -8.67]	**-11.83** [-23.71; 0.05]
202.36	21.62	12.65	-18.40
Diabetes			
All age groups	**-3.59** [-3.82; -3.37]	**-2.08** [-3.02; -1.13]	**-2.63** [-3.26; -2.00]
14.49	2.18	-1.22	0.76
Epilepsy			
All age groups	**0.69** [0.62; 0.77]	**0.44** [0.12; 0.76]	**0.99** [0.48; 1.50]
2.57	0.20	-0.10	-0.35
COPD			
All age groups	**5.59** [5.07; 6.11]	**5.74** [3.74; 7.74]	**5.35** [1.60; 9.10]
76.32	27.09	29.15	-3.01
Asthma			
All age groups	**-0.88** [-0.97; -0.79]	**-0.62** [-1.00; -0.24]	**-0.93** [-1.32; -0.54]
3.66	1.76	1.05	1.44
CKD			
All age groups	**-1.10** [-1.22; -0.98]	**-0.78** [-1.31; -0.25]	**-0.85** [-1.80; 0.09]
5.30	1.10	0.08	0.59

Chapter 4

Discussion

The objective of this chapter is to discuss the above results with particular focus on the causal approaches and frameworks introduced in earlier chapters. To that end, it is structured as follows. The major findings of each of the studies will be critically discussed with respect to what implications the respective methods provide intrinsically and also how other frameworks of causality, before turning to reasoning about causality in a one-by-one manner.
Using the combined findings from all three studies, the chapter is concluded by general remarks and limitations concerning causality in observational studies.

4.1 Separate Causal Discussions

4.1.1 Case study 1

In the following section, implications of the results of the first case study will be covered, discussing conceivable causes and considering implications on causality based on the causal frameworks introduced in chapter 1. Using

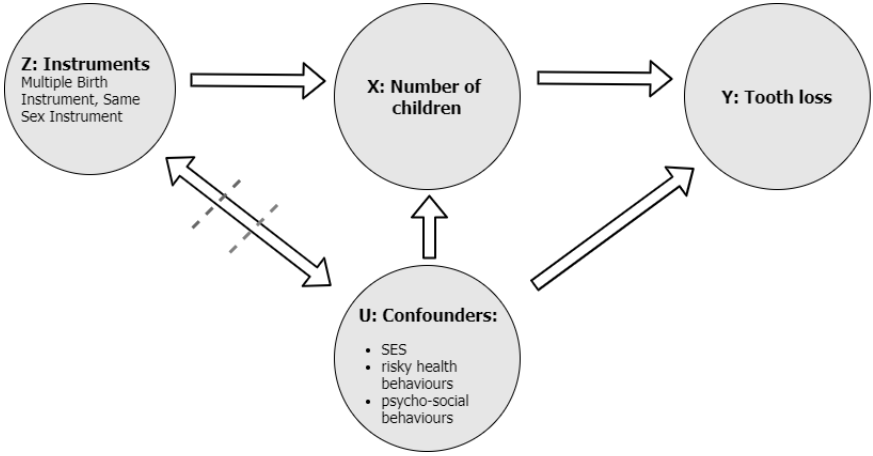

Figure 4.1: A directed acyclic graph depicting the presumed relationships between the number of children and tooth loss. The methodological setup deemed both the inclusion of relevant confounders as well as instrumental variables necessary.

quasi-experimental methods (instrumental variables / 2SLS) and unique survey data from SHARE (Börsch-Supan, 2016), the relationship between the number of biological children and the number of missing natural teeth among these children's parents was investigated. Thereby, random natural variation in family size resulting from (i) the birth of twins vs singletons, and (ii) the sex composition of the two first-born children was investigated, relying on an increased likelihood of a third child if the two first-born children have the same sex.

Regressions detected a strong relationship between the number of children and teeth for women when an additional birth was given after the first two children had the same sex. Women then had an average of 4.27 [95%-CI: 1.08; 7.46] fewer teeth than women without an additional birth whose first two children had different sexes. In contrast, sizeable effects for the relationship between children and missing teeth for men or when using natural variation in twin births as an instrument could not be identified. The choice of 2SLS as a estimation strategy provides a strong argument for a

causal interpretation of these effects.
Therefore, the above study examined causal links between fertility and the number of missing natural teeth. This relationship is generally plausible - tooth loss is a relevant health outcome, representing a frequent endpoint of dental diseases, particularly untreated caries (the most prevalent disease worldwide according to the Global Burden of Disease study) and severe periodontitis (the 6th most prevalent disease worldwide according to the Global Burden of Disease study). It is worth noting that plain OLS is problematic in this case: there is a multitude of often unknown factors that both contribute to the probability of parents to get children and tooth loss, some of which change over time and some of which maturate. As argued above, this jeopardizes internal validity as effects might be captured that are of no interest. To strengthen causal arguments, the PCM offers to introduce a priori knowledge (or assumptions) of causal connections. Figure 4.1 displays a conceived graph. OLS only considers the right three nodes where it is unclear whether all possible confounders are included or if data to correct them are available.
Therefore, another strength of the study is the use of instrumental variables, which offers a remedy to this problem and in combination with unique and large-scale survey data, identifies causal effects for a field of research in which there was previously no causal evidence. This is particularly relevant given that controlled experiments on the relationship between the number of biological children and the number of missing natural teeth seem impossible. As such the present study is, to our knowledge, the first to provide causal evidence for the question whether tooth loss is influenced by fertility. On the other hand, our study also has some limitations. The external validity of our results is limited because our findings only concern narrow subpopulations ("compliers") with certain fertility patterns (experience of multiple birth; additional child because the first two children had the same sex). Moreover, since the complier-subpopulations are relatively small, most of our results suffer from large standard errors, which is a typical "price to pay" (statistically) for the theoretically "clean" causal identification via instrumental variables. Nevertheless, the plausibility and usefulness of the analytical approach used

in the present study is endorsed by the existing and growing literature in support of quasi-experimental methods. The features of the instrumental variables used in this study correspond closely with previous literature (for example, see previous publications using the same sex instrument (Angrist and Evans, 1996; Kruk and Reinhold, 2014b)).

OLS regressions indicated that, per additional child, women have an average 0.57 fewer natural teeth and men have an average 0.26 fewer natural teeth. By and large, these results are consistent with previous non-experimental evidence on the relationship between the number of children and teeth. Although sparse, most of the extant literature suggests that individuals with more children have more missing teeth than individuals with fewer children. For example, 70-year-old Swedish women without children were found to have 5.0 to 6.6 more teeth than their counterparts with five or more children. Similarly, a study from Japan found an age-adjusted difference of 2.97 teeth between women with no and women with more than four children (Ueno et al., 2013). However, given that unobserved confounders may affect both oral health and fertility, such results should be interpreted with caution. A solution to this problem is to exploit two different natural experiments which give raise to exogenous variation in family size: (i) the birth of twins vs multiples; and (ii) the increased likelihood of a third child if the two first-born children had the same sex (as compared to different sexes). As is common practice in the quasi-experimental literature, 2SLS regressions for computation are used. 1st stage estimates are in line with studies using similar instruments (Angrist and Evans, 1998; Kruk and Reinhold, 2014a), hence corroborating this empirical approach. In 2nd stage regressions, a large causal effect of 4.27 fewer teeth for women with an additional birth after their first two children had the same sex was identified (as compared to women without additional birth whose first two children had different sexes); this is a larger effect size than that identified via OLS regression or that previously reported in the non-causal literature on the association between fertility and (missing) teeth. One potential explanation for the larger effect size in 2SLS regression is that healthier women are generally less susceptible of tooth loss and tend to have more children than less healthy

women; as health is largely multifactorial and likely not fully observable, OLS regression and other non-causal approaches may underestimate the effect of children on the number of missing teeth due to omitted variable bias.

In the present study, a large causal effect of fertility on tooth loss could only be identified for women and only via the "same sex" instrument but not for men and not via the "twin birth" instrument. Differences between regressions with different instruments are to be expected if treatment effects vary across population subgroups (Imbens and Angrist, 1994). Individuals with changes in fertility because their first two children had the same sex are likely different from individuals with changes in fertility because of a multiple birth. Causal effects of fertility can thus differ for various reasons. First, twin births cause two children to be born and grow up at the same time; whereas births induced by the parents´ sex preference on their offspring occur consecutively. Second, twin pregnancies are more demanding than singleton pregnancies; evidence indicates the total birth weight of twins is nearly twice that of a singleton birth weight (Min et al., 2000). But having an additional child because of the sex imbalance of the first two children implies an additional pregnancy (with increased risk of gum disease) and an additional cycle of (time-consuming) parenting. Hence, two singleton motherhoods might have a different (as our results suggest: more detrimental) effect on the oral health of the mother than one twin motherhood. Yet the precise role of pregnancy-related vs parenting-related factors requires further deciphering. On basis of our findings, enhanced promotion of oral hygiene, tooth-friendly nutrition, and regular (preventive) dental attendance – specifically targeted at expecting and parenting mothers – seem to be sensible strategies for clinicians and health policy. This study provides unique and novel evidence for causal links between the number of natural children and missing teeth. While no sizable effects for men could be identified, two motherhoods with singletons seem to be more harmful to a mother's oral health than one motherhood with twins. Still, the role of pregnancy-related vs parenting-related impacts on individuals' oral health need to be examined in more detail. For example, the changing roles of modern day fathers might

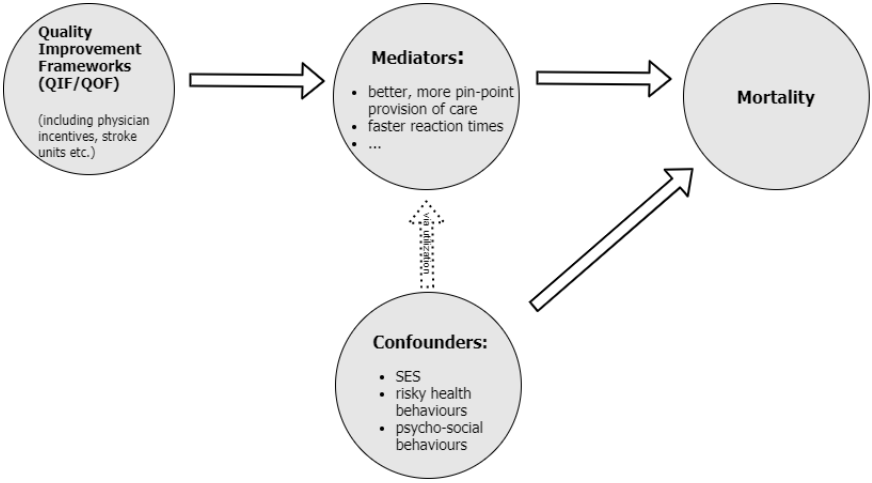

Figure 4.2: A directed acyclic graph depicting the presumed relationships between the Quality Improvement Frameworks introduced to the UK NHS and mortality. This presumption deemed the methodological approach of Differences-in-differences necessary and possible.

provide an interesting analytical setup: if causal research would detect no relevant effects for fathers which are younger than those observed in our study, this might provide evidence against parenting-related impacts on oral health. Further research is needed to establish refined interventions against tooth loss.

4.1.2 Case study 2

In the following section, the results of the second case study will be covered, discussing conceivable causes and consider implications on causality based on the causal frameworks introduced in chapter 1.

The cardiovascular health of the population of Stoke-on-Trent improved faster from 2004, with statistically significant greater improvements seen in Stoke-on-Trent when compared with most other populations. These were associated in time with the 2004 QOF and the 2009 QIF. The national

improvement was a reduction of 10 deaths per year per 100000 of the general population; the additional effects associated with the QOF in Stoke-on-Trent per annum were 36 CHD deaths per 100000 general population and 166 CHD deaths per 100000 population of 65-74 years old.

Figure 4.2 displays a DAG that encodes assumptions about causal relationships contributing to the proclaimed association. The graph deliberately contains only a sparse selection of variables deemed contributing - reflecting both the fact that actual causal pathways are unclear and that only very few confounders were measurable. The intervention scrutinized in this scenario entails a variety of sub-interventions and influential factors of which many likely interact and essentially none are available for analyses. The choice of ITS as a methodological strategy therefore represents a valid means of encountering such issues.

Stroke mortality in the 65- to 74-year age group showed that in addition to the national effect of the 2004 QOF introduction of about -8.75 deaths per 100000 population, there was an additional reduction of around -5 deaths per 100000 per year in Stoke-on-Trent.

Whether these changes are causally connected to the introduction of QOF/QIF respectively is unclear. In chapter 2, it has been debated under which assumptions DiD can be used to estimate the ATE. Most prominently, estimation can only be unbiased if the common trend assumption is met which can only be verified graphically. Figure 3.2 displays three time series corresponding to outcomes in different localities, showing that the CTA is a reasonable assumption. Remarkably, the strength of the associations is large, as the estimates correspond to significant reductions in mortality rates. According to the BHC, this strengthens claims of causality. These claims are also strengthened by the coherent nature of the findings. Following the 2009 local introduction of the QIF, there were further significant reductions of mortality rates for most conditions measured, again largest for CHD and stroke. These effects remain when compared to the West Midlands but are not detectable in comparison with peer localities. A possible explanation for this is that when the QIF commenced, Stoke-on-Trent had improved its implementation of evidence-based interventions in response to the QOF to

improve cardiovascular health better than those peer localities; therefore, the ability to further achieve a statistically significant reduction mortality was reduced because much of the available benefits had already been achieved. An alternative explanation is this might have occurred due to the statistical impreciseness of the coefficients. The likely explanation of the failure to detect a statistically significant reduction in stroke in <65-year age group from 2009 is the low event rate at baseline and therefore the small number of potentially preventable events in that age group, especially in relatively small, sub-group population samples.

Benefits were greatest for the high-prevalence conditions amenable in the short term to evidence-based interventions-blood pressure lowering and lipid-lowering medicines, and the existing smoking cessation services and support. Self-reported, short-term smoking cessation rates were high in Stoke-on-Trent during this time (NHS England, 2013), and sample data showed that hypertension and cholesterol levels improved locally during the relevant time period. By 2014/2015, detection and control of hypertension were better than comparable localities, while overall smoking, diet and activity indices in Stoke-on-Trent continued to be adverse. A differential increase in the effectiveness of the acute, secondary care treatment of myocardial infarction and stroke in Stoke-on-Trent compared with other localities is a possible but unlikely explanation for our findings. It seems far more plausible that the mortality of high prevalence chronic diseases such as CHD and stroke is more amenable to primary prevention interventions than secondary care interventions.

That differences in mortality rates were detectable and associated with the QIF that started in 2009, after 5 years of QOF, is a very interesting finding. For there to be detectable, small mortality benefits across several conditions, including diabetes, asthma and CKD where simple short-term therapeutic interventions are less likely to result in detectable improvements in mortality data-is notable. These consistent excess reductions in mortality further increase the likelihood of a causal connection.

The results demonstrate that some important outcomes that health care quality improvement schemes seek to address can be satisfactorily assessed

using publicly available mortality data. However, there are well-known limitations to mortality data (notably diagnostic imprecision), and local authority populations do not map directly to patients registered with practices in clinical commissioning groups. This constraint, might explain part of the effects found, even though the direction of error is unclear. Further, given the large year-to-year variability in the data of the lesser prevalent conditions, the corresponding results should be treated with care, since eventual effects of any change could be concealed by random variation.

A detailed review of the literature evaluating pay-for-performance schemes was undertaken to inform this evaluation. The evidence that large, complex, pay-for-performance schemes improve the health of populations is mixed, and no examples of local schemes similar to the Stoke-on-Trent QIF with its multifaceted approach combining P4P, professional and managerial support and monitoring, and educational co-initiatives were found. In summary, the concerns with P4P schemes are a lack of evidence of benefits associated with the schemes, loss of focus on conditions outwith schemes, schemes not being relevant to local health priorities, mechanistic approaches to individual care as clinicians 'follow the rules' irrespective of whether the intervention is appropriate for that patient (including their values and preferences), and the sheer burden of administration and management on the workforce. Perhaps the most important finding in the many evaluations of the UK QOF is that it was associated with a reduction in health inequalities (Doran et al., 2008); this analysis supports that finding.

It seems plausible that both the QOF from 2004 and QIF from 2009 may have contributed to reducing premature mortality from some important conditions in this specific locality. Given the limitations of large, national, pay-for-performance schemes, the question is what now replaces large-scale, complex, invasive, mandatory measurement as the dominant approach in some health systems to reduce unwarranted variation in provided care (Berwick, 2016). Despite several inherent analytical limitations, a local, multifaceted scheme incorporating P4P alongside other locally agreed strategies may improve the health of populations. In the short term, benefits may only occur for common conditions for which there are simple, safe, effective,

acceptable interventions in localities with high event rates. Benefits may be more difficult to achieve when disease-specific pathophysiology is more complex, and when event rates in the targeted diseases drop over time, presumably in part due to early gains resulting from the more consistent adoption of interventions in vulnerable populations. Nevertheless, local approaches, if they are well led and managed, may overcome many of the drawbacks of national schemes.

4.1.3 Case study 3

In the following section, the results of the third case study, utilizing **Interrupted Time Series Analysis** will be put into perspective, discussing conceivable causes and consider implications on causality based on the causal frameworks introduced in chapter 1. The findings of this study indicate significant and quantitatively large shifts in treatment compositions following regulatory changes to provider incentives in Denmark in 2015. By having dentists classify patients into three distinct risk groups, these changes were intended to effect a transition from six-to-twelve-monthly dental recall intervals for every patient towards a more patient-centered model in which patients with higher need should receive dental recalls systematically more frequently than patients with lower need.

In comparison to the pre-reform period, our findings suggest that the proportion of patients with dental recalls every 6 months or more often increased by 1.2%-points, the proportion of patients with 6-12-monthly recalls increased by about 0.7%-points and the proportion of patients with more than 12-monthly dental recalls decreased by about 1.9% points. While this distribution of recall intervals changed only to a relatively small extent, the composition of utilized care items shifted substantially. For the time period following regulatory changes, substantial increases in preventive services and scaling as well as a substantial decrease in diagnostic services were observed. Given the comparably low dental disease burden in Denmark, these findings may appear to be somewhat against expectations.

In international comparisons, the Danish population has comparatively

good oral health. According to WHO criteria, the Scandinavian countries belong to the so-called very low and low-caries prevalence countries (Petersen, 2003; Silveira Moreira, 2012), but there is still room for improvement due to disparities related to social inequalities (Rosing, 2015).

However, in a recent review, the Danish Health Authority reported that following the 2015 reform, only a minority of Danish patients were classified as having a low oral disease risk. It was remarkable that - despite good oral health - the majority of examined individuals were categorised as belonging to the at-risk groups. However, following regulatory changes, most patients (79%) were classified as either being at-risk or high-risk (Sundhedsstyrelsen, 2017). Following the reform, there were many payments for the FE-code, pertaining to patients with active disease status only. This code also made it mandatory to perform IPT (unlike before the reform, where preventive treatment was to be performed only when found to be necessary by the dentist), contributing to the steep rise in the usage of this code. Also, the number of "scalings" increased in a similar manner, while the use of diagnostic codes decreased - as most patients were assigned to the yellow and red tracks, there were fewer basic examinations where diagnostics were being performed routinely (see Figure 2.3). The volume of other types of treatment changed only marginally.

It's plausible that treatments vary in response to altered incentives as has previously been reported (Brocklehurst et al., 2013; Chalkley and Listl, 2018), the type and extent of variation observed in the present paper may still appear intriguing. It is relevant to note that facilitative problems have been reported with respect to the regulatory changes examined in this study. A recently published report pointed at misalignments in treatment codes and care delivery, that is dentists were unable to receive a remuneration for a filling that needed replacement unless the patient was classified as having active disease (*Healthcare in Denmark: An overview* 2016). Apparently, dentists categorised patients in the yellow (moderate) risk group in order to get any remuneration. This was reported to have given rise to approximately 20% of yellow risk group categorizations. Also, the use of the initially introduced criteria for diagnosing gingivitis seemed to be affected

by inaccuracy, resulting in a relatively large number of patients who were diagnosed with mild gingivitis (yellow risk category). After revision of the guideline, the criteria for mild gingivitis were changed so that individuals with bleeding gums at 15% of sites or less could be classified as green instead of yellow, leading to a 7%-point decrease in patients classified within the yellow category (down from a baseline of 33% before the revision).

While a considerable proportion of patients might have been misclassified, this may have happened for various reasons. Dentists may have exploited the high-risk classes by increasing demand for their services for a financial gain, taking into account that patients classified as "green" are inclined to visit the dentist less often. If such behaviour was also linked to changes in clinical activity, it may be questioned whether such changes were in the best interests of the patient. In addition to provider incentives, other factors might prevent movement towards a more individual-risk-based approach to recalls for dental check-ups. First, assessing risk is not a trivial task (Cagetti et al., 2018), especially for high-risk patients (Twetman et al., 2013). There have been doubts regarding the usefulness of popular risk assessment tools (Clough et al., 2016). Second, given the scarcity of conclusive evidence regarding fixed vs. variable recall intervals (see above), it could be hypothesised that dental care professionals may be reluctant to extend recall intervals for risk of causing harm or potential (perceived) threats of legal action for alleged malpractice. Some dentists may believe that the prolonged recall intervals for the “green” category may be too long. Third, demand-side influences may also play a role. For example, patients may have developed preferences for a frequency of services that they have become familiar with over a longer period of time or dentists might feel pressured by patients to categorize them into the yellow risk group instead of the green one because patients might have expectations or preferences to be seen semi-annually as they have been before the reform. Following this logic, the "yellow" and "red" risk classes might support both physician profit and (perceived) patient care.

These presumed causes can be nicely subsumed graphically (see Figure 4.3). However, it is deemed inconceivable to derive mathematical notation of causality from this graph - much of these confounders are unavailable to

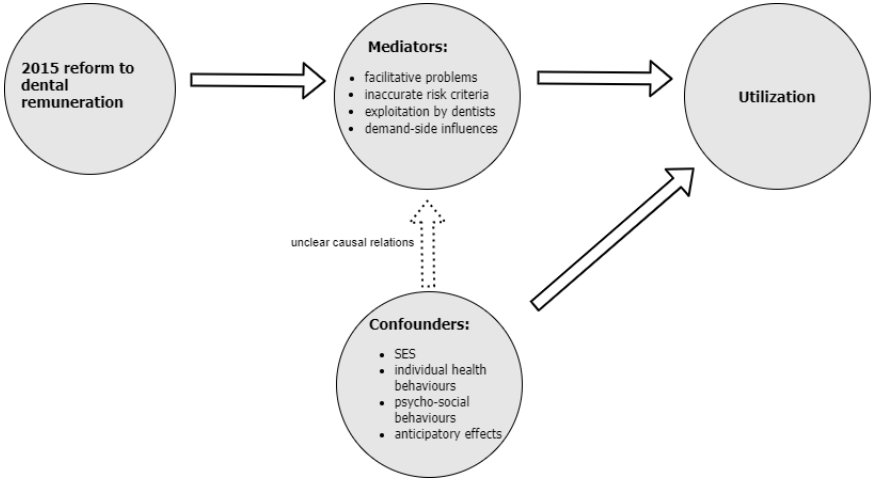

Figure 4.3: A directed acyclic graph depicting the presumed relationships between the introduction of a reform to remuneration in dental care in Denmark.

analyses. By and large, rationally explainable distortions in classifications of patients into risk groups due to dentist or patient interests or preferences may constitute only one possible explanation for the observed treatment patterns. It is not always clear how best to roll out and sustain innovations to health care systems while taking into account the perspective of all relevant stakeholders and, at the same time, achieve the desired changes in health care providers' behaviour (Greenhalgh et al., 2004b). Given their standards, beliefs and expectations, dentists and patients may react differently than anticipated by health policymakers, leading to unpredictable consequences (possibly caused by, e.g., utility maximization or rational/irrational behavior).

Our study has limitations. First, there is no way of postulating intrinsic claims of causality as conducting a sufficiently randomised experiment is not possible at this scale (which contradicts Hill's "experiment" criterion of causality). As the calculations are based on unique patient-level and population-representative data composed of all 72 million treatments car-

ried out in Denmark from 2012 to 2016, statistical analyses of this study do not suffer from pitfalls related to sampled data such as selection bias, entirely eliminating the need of extrapolating statements to a larger population. Also, the multifactorial nature of the data allowed an exact matching of patients and treatments, giving rise to the application of fixed effect methods and thus the modeling of unobserved inter-patient heterogeneity. In total, the statistical approach used is highly robust and not very prone to error and bias. However, our study suffers from the fact that detailed individual risk groupings were not available for scientific evaluation. Despite performing some robustness checks and using proxy variables for dental disease risk (socio-economic status, age, previous treatment experience), no more detailed analyses within actual classifications of risk groups could be performed. This tarnishes causal claims as incorporating risk groupings would make a causal connection more "plausible" in Hill's sense (of course, assuming no qualitative change in results). Another weakness of our study was the lack of suitable outcome measures – while the DHA regulations require dentists to monitor the number of carious and missing teeth as well as the number of teeth with fillings, these data are limited to very narrow age groups, hence they do not provide sufficient longitudinal character and rendering the usefulness of the data for purposes of assessing oral health outcomes non-applicable to the present study (see Appendix for details)). Note that all treatment codes were changed with the introduction of the regulatory changes (only exception: code for preventive treatment). While this means that, by design of the regulatory changes, there was no possibility for dentists to be slow in changing towards the new taxonomy of treatment codes, data entry mistakes could still be a relevant issue. However, given the absence of reliable reference values for dental diagnostics or treatment needs, the extent of such coding issues is difficult to determine.

Successful implementation of healthcare reforms is a function of the dynamic interaction between evidence, context and facilitation (Cohen et al., 2015). While it is often hard to translate research findings into modified provider behavior (Grimshaw et al., 2001), the problems encountered

throughout the studied reform have mostly been contextually and facilitative.

Moreover, implementing health system and practice changes through incentives is highly complex and many various influencing factors can determine success or failure of payment reforms (for a more detailed review of the relevant literature see the introduction section above). Therefore, the question arises how future implementations of similar systems can try to prevent these problems. For example, a sensible amendment could be the use of a monitoring system such as regular checks and inspections of dentists' assessment of oral disease risk scores by an independent regulatory institution. This way, the use of risk categories might be better aligned with actual disease prevalence rates. Another conceivable scenario would be to connect a risk-classification system and a Pay-for-Performance (P4P) implementation based on the powerful assumption that individuals and organizations are motivated to perform better by incentives (Witter et al., 2013). Literature regarding previous implementations of P4P systems in dentistry is sparse, but it has been suggested that P4P in dentistry may not be a viable option before progress is achieved in the development of reliable indicators for quality of dental care (Voinea-Griffin et al., 2010). Other conceivable options include introductions of risk group quotas based on available evidence to regulate the share of patients in each group. This approach was used as an ad-hoc solution in the actual reform by the DHA in 2016 (Sundhedsstyrelsen, 2017). In addition, the existing literature about program implementation suggests it could be sensible to precede large-scale implementations with localised, controlled implementation trials to estimate possible ramifications (Bauer et al., 2015) as adaptations to implementations of health care reforms are not uncommon (Escoffery et al., 2018). This would allow for a closely supervised roll-out during which changes can be implemented smoothly and in a co-productive manner (considering the perspectives of all relevant stakeholders) in advance of large scale (and expensive) implementation. Not least, it is important to bear in mind that health care payment reform may be shaped by social and learning processes that can affect all stakeholders involved (Conrad et al., 2015).

It remains to be discussed how Campbell's approach to causal inference can be used to analyze causality in this study. In this study, the reform was, to the knowledge of the researchers, not accompanied by events other than planned treatment (i.e. history). Also, the risk subsumed as "maturation" is not very prevalent as the reform was not incrementally introduced, rendering changes during intervention impossible. The "instrumentation" risk, however, poses a greater risk to internal validity for two reasons. First, as the study is based on claims data, these claims may not be consistent across time, within practices and ETC. Second, the reform was accompanied by changes in remuneration codes, making comparisons between practices before and after the reform challenging. As this study is based on a complete sample, there is neither selection nor attrition effect in place.

Another way of reasoning about causation is presented by the Bradford-Hill criteria. The first criterium, "strength of association", demands that the larger an association between exposure and disease, the more likely it is to be causal. As it's never objectively possible to call an association "strong" or "weak" - arguments based on contextual information are deemed necessary to do so, for example by comparing changes with baseline values. In the current study, the share of sessions containing preventive items increased from 0.153 by 0.310 (on average), from 0.753 by -0.345 for diagnostic items, from 0.584 by 0.241 for sessions including scaling and from 2.413 by -0.012 regarding the total number of sessions. Considering that all these point estimates have been calculated on non-sampled data, these comparisons bolster the "strength of association"-argument, thereby strengthening the causality argument in Hill's sense.

The study does not uphold the "consistency" criterion as there are no similar studies with a variety of locations, populations and methods showing the same association. As also mentioned in the second case study, this is often the case in studies analyzing impacts of policy reforms - the more overarching and the more specific the reform, the less likely it is to find comparable results reinforcing causality claims.

The "specificity" criterion is also hard to justify - there may be multiple causes contributing to the found effect. In the above discussion, several of

these possible causes were discussed. Figure 4.3 also displays this relationship.

That being said, “temporality” is obviously given - as can be seen in figure 3.1, the steep changes in utilization begin right after the reform has taken place. Interestingly, one could argue that there is a “biological gradient” in place, i.e. a dose-response relationship, even though this term refers to epidemiological or biological circumstances mostly. More accurately, some of the treatment baskets (such as preventive items) are more affected by the reform than others, and they also show the largest response to the reform. This train of thought is complicated by the fact that these “doses” are only assumed (in real-world biological scenarios, one could objectively increase the dose and measure the response).

4.2 Putting things together - discussion on causal claims

In the earlier chapters of this thesis, three case studies relating to causal inference on observational data were presented. While it was insinuated that intrinsic causal inference is only possible when conducting randomised experiments (leading to an estimate of treatment effects), it was shown how observational data in conjunction with causal frameworks can be used to deduce causal statements when certain identifying assumptions are met. Matters are made easier by scenarios where randomized experiments can be emulated.

In general, performing randomised experiments is often challenging in social, political and life sciences where proper randomization is unfeasible for ethical, juristic or practical reasons. Thus, researchers are forced to resort to either naively extrapolating statements from associational information (attained typically from a linear regression design or a variation thereof) or, more reasonable when applicable, using a method for causal inference on observational data. The common discrepancy between the inferential goal of causality and the reality of neither being able to capture causal effects by design or analytically contributes to the mentioned methodological crisis. In

this concluding section of the discussion chapter, the empirical applicability of these causal frameworks will be discussed by contemplating how they can be integrated in the case studies enlisted above and also research conducted by others.

Tightly embedding the three causal models PCM, CCM and RCM within the context of case studies utilizing quasi-experiments showed that each of them can be used to better integrate causal thinking into scientific arguments. The PCM forces researchers to wrap a priori assumptions of causal relationships into causal graphs, the CCM enforces contemplating about internal validity and the RCM encourages counterfactual thinking. This section will start by enlisting the practicability of these frameworks based on the three case studies discussed above.

The PCM in particular could unfortunately not be used for calculating causal effects as suggested by Pearl, but rather for displaying presumed causal relationships graphically. There were mainly three reasons for this. First, the assumed causal graphs were so simple that the rules of *d-separation* were not applicable. Second, there are always doubts regarding assuming the correctness of priors such as contrived causal graphs - subsequent deduction depend on them being correct and third, data availability gravely constricted any inference possibly drawn from such graphs as many confounders deemed necessary were simply not available for analyses.

As such, a direct application of the PCM requires benevolent data scenarios which are mostly not given. The CCM, however, is easily applicable as it is based on arguing in favour of internal validity. It was found that these arguments are somewhat helpful during interpretation as they indicate inappropriate conclusions from results stemming from selection biases, regression-to-the-mean, instrumentation and so on. Still, it is likely that researchers are aware of these limitations, anyway. Also, it is conceptually similar to the Bradford-Hill criteria, which systematically cover principles that subsume evidence for a presumed causal relationship between variables (Hill, 1965a). These are implicitly used in established scientific discussion structures: *plausilibity* and *coherence* ensure that results are consistent with

the body of knowledge, *strength*, *specificity*, *biological gradient* and *temporality* ensure internal validity and *experiment* and *analogy* occasionally help uncover similarities to experimental evidence or similar previous studies.
Similarly, the RCM is a powerful framework of causal thinking that forces researchers to formulate counterfactuals ("What if"-questions) and allows derivation of different types of treatment effects depending on the study population. For example, as has been argued, instrumental variables are only able to identify treatment effects for compliers. For this group, however, these counterfactuals can be calculated, which is a powerful tool for causal inference. Many of the econometric methods ultimately ensure exchangeability of units (and thereby calculation of counterfactuals) under study by imposing identifying assumptions that then allow straightforward comparison of treated and untreated units. In essence, the RCM is an overarching principle that all methods for causal inference can be condensed into - using the case studies above, the generality of the RCM can be displayed. Any causal inference can be ascribed to some form of counterfactual question: these questions were "What would be the oral health of parents had they not had children? (when in fact they had children)?", "What would mortality have developed like, had there not been quality improvement schemes in a local NHS implementation?" and "How would the utilization in Danish dental care system have developed had there been no re-design of the remuneration system?".

Also, by design, a very prominent facilitator of causal inference in observational studies is data with a time component: temporal precedence of the cause over the effect has been exemplified as a necessity for causation several times above. This precedence, however, also constitutes a chance to perform valid causal inference (rather than just a necessity). Therefore, the following section enlists, based on experiences drawn from the case studies above, arguments for the usability of longitudinal and cross-sectional data in causal inference research.
It seems natural that putative causes and effects should indeed be ordered in time. By that logic, longitudinal data (where the same individuals are followed over time either prospectively or retrospectively), are required for

testing causal hypotheses. Such data can vastly help regression analyses as they allow capturing "individual effects", which are defined as per-individual, unobservable, time-invariant characteristics (*Fixed Effects*). If these individual effects are assumed to be the result of random variation, the model of *Random Effects* may be applied. In general, application of these models (where possible) is beneficial as they allow controlling for unobserved confounders correlated with observed explanatory variables and consistent estimation of their effects.

Two of the case studies above utilised methods (DiD and ITS) make extensive use of these counterfactual arguments by comparing units of observation before and after an acclaimed cause has taken effect. The lack of such temporal features in data disallows a straightforward application of counterfactual arguments as suggested in the RCM - elapsing time between measurements of some outcome is contrary to the idea of counterfactual outcomes that were measured at the same time and under the same circumstances.
It's interesting that the identifying assumptions of DiD and ITS assume such similarity of circumstances explicitly (e.g. by means of the common trends assumption).

The study has limitations. The data acquired to conduct the studies above originated from insurers incentivised through the European research project ADVOCATE and the pan-European survey SHARE. It is well-established and makes sense intuitively that data containing a large number of covariates (as possible causal mediators) makes causal inference more plausible (Imai et al., 2010). However, data availability turned out to be a major obstacle in performing valid causal inference in observational studies. Complications related to data availability were four-fold. First, it was found that in practice, there is a significant gap between what some of the causal models implicitly assume regarding availability of covariates as well as data in general and what is typically available to researchers. This discrepancy particularly relates to causal models that require significant covariate adjustment to construct valid models for causal inference. For example, the PCM assumes knowledge of a causal graph in which every mediating causal factor (i.e. node) is covered by underlying data. If that is not the case as in our practical

examples, only a subset of possible causal factors can be accounted for, resulting in partly restricted, partly unclear interpretability of the remaining graph.

Second, as was already discussed in the case studies, non-experimental data available to researchers is often prone to selection bias, which may eventually lead to entirely spurious associations and false inference (Ellenberg, 1994). Meanwhile, detecting selection bias is hard and avoiding it typically requires a careful planning of studies (Hammer et al., 2009). In presence of selection bias, regression coefficients are confounded with regard to the function determining the probability that an observation makes its way into the non-random sample (Heckman, 1979). In certain situations, selection bias can be mitigated using *Heckman correction* where self-selection is controlled using an additional predictor function. However, it has been since shown that this method only works in special scenarios (particularly in absence of multicollinearity) (Puhani, 2000).

The third practical obstacle encountered in this thesis were missing data due to incomplete observations. (Rubin, 1976) points out that inferences from such data generally depend on the observed pattern of missing data where it is only appropriate to ignore missing data when they are "missing at random" and the observed data are "observed at random". Meanwhile, elaborate techniques have been developed to identify missing data mechanisms and to statistically reflect this information, mostly related to the well-known resampling technique of multiple imputation (Sterne et al., 2009).

Another major impediment for the analyses conducted in ADVOCATE were data protection regulations. High standards of data privacy, requiring personal data to be processed lawfully and fairly based on the subject's consent, are valuable in modern societies and have been subject to intensive debates in the recent past. Some of its requirements are, however, fundamentally incompatible with the demands of scientific research, especially those that require data processors to disclose the purpose of data processing and minimize their use of data as it is not always clear to researchers which data are needed and which covariates might be useful for their future models.

The digitation of medicine has made a pledge towards tapping vast public health databases for research purposes. Data protection regulations such as the GDPR are threatening to derail projects such as ADVOCATE that aim to utilize such knowledge. For this reason, some data providers were only able to provide aggregated data that were found mostly unusable for meaningful analyses, let alone considerations for sensibly inferring causality. Further, possible erasure of individuals' data from health care records or health insurance databases introduces another possibility of selection bias in analyses if this selection is non-random and some characteristics of individuals who have their data removed are correlated with the outcome of interest.

Further, even when suitable data is at researchers' disposal, applying quasi-experimental methods ultimately rest on the availability of suitable interventions. Often, these are very hard to find as unlike in controlled, randomised experiments, there is no active conduct of studies. Also, quasi-experimental methods have narrowing premises that restrict external validity and generalizability - for example, IV only identifies causal effects for a subgroup of the study population having been affected by the instrument. In the case study above, any statements therefore only apply to parents who both had two children and also decided to get another one due to them having the same sex. Generalization beyond this subgroup is then a matter of reasoned argumentation. The causal frameworks can help support such claims by providing causal graphs (as seen in Figure 4.1), argumentation using counterfactual claims ("Would parents with two children of different sexes plausibly have reacted similarly to the same sex instrument?") and using structured arguments as Campbell suggested. Lastly, it's not always clear in quasi-experiments if all identifying assumptions are met. As they (or rather, some of them) are inherently untestable, researchers have to make an a-priori effort to plausibly reason about the applicability of these methods (this is very much similar to designing a randomised study).

Scientific work will always benefit from arguing in favour of causality. This has become particularly relevant due to the ongoing methodological crisis that states that many studies are not replicable. Partly, this is due to researchers finding spurious correlations, selective reporting, or "p-hacking"

(Pashler and Wagenmakers, 2012). It was shown that 14 % of researchers alleged others of having fabricated and falsified data (at least 2 % conceded own scientific misconduct) (Fanelli, 2009). Thus, future research should make sure that they don't fall prey to this by utilizing appropriate methods, common sense and principles from *reproducible science*.

Chapter 5

Conclusion

In this dissertation, observational data were used in conjunction with methods for causal inference to deduct causative relationships between variables of interest. Disregarding causality in empirical research altogether exposes results to the risk of enormous biases and false interpretations. This is arguably problematic as expressing the simple causal relationship between two variables X and Y statistically is not possible in an explicit manner - a problem known as the "fundamental problem of causality" where only one of multiple possible outcomes manifests in the real world.

Meticulously conducted randomised experiments allow approaching an "as-close-as-possible" replication of these multiple outcomes, thereby constituting science's greatest tool of attaining real causal inference. However, it was argued that observational research, a major constituent of scientific progress, does not possess this luxury. Researchers have to resort to methods designed to emulate experiments (quasi-experimental methods) or make strong arguments (in the form of, for example, grounded causal graphs) in favour of possible causal interpretations. Thus, this work contributed to preventing these shortcomings by attempting to integrate methods for causal inference on observational data with three frameworks to conceptually represent

causal relationships - the frameworks of Neyman-Rubin, Pearl and Campbell. It was found that these frameworks' extensive causal implications are based on strong theoretical assumptions that are hardly met in practice. While this often forbids direct application of such frameworks, there is still merit for scientists to attempt and integrate the general ideas of these frameworks into their research: constructing strong counterfactuals (inspired by Neyman-Rubin), building causal graphs (inspired by Pearl) and check-listing research strategies to preemptively exclude threats to internal validity (inspired by Campbell and Bradford-Hill).

The daunting replication crisis, publication bias and fake science have wrongfully dented researchers' reputations all over the globe. This world full of alternative realities needs the neutral voice of science to articulate strong facts more than ever.

> Facts are stubborn things; and whatever may be our wishes, our inclinations, or the dictates of our passions, they cannot alter the state of facts and evidence.
>
> (John Adams)

Summary

This dissertation reflects the use of various methods of causal inference using observational data based on three articles in health economics and epidemiology. These case studies encompass analyses of the implementation of a quality improvement framework in UK primary care (using Difference-in-differences), the analysis of the impact of bearing children on oral health (using Instrumental Variables) and the analysis of an implementation of altered provider incentives in the Danish dental care system (using Interrupted Time Series Analysis). All of these make use of methods utilizing quasi-experiments that allow post-hoc derivation of causal effects if certain identifying assumptions are met. Disregarding causality in empirical research altogether exposes results to the risk of enormous biases and false interpretations. This is arguably problematic as expressing the simple causal relationship between two variables X and Y statistically is not possible in an explicit manner - a problem known as the "fundamental problem of causality" where only one of multiple possible outcomes manifests in the real world. This thesis offers a remedy to this problem by showcasing how the causal frameworks of Pearl, Campbell and Rubin may contribute to causal argumentation and ultimately more robust research.

Zusammenfassung

Diese Dissertation reflektiert die Anwendung verschiedener Methoden kausaler Inferenz unter Verwendung von Beobachtungsdaten auf der Grundlage von drei Artikeln aus der Gesundheitsökonomie und Epidemiologie. Diese als Fallstudien konzipierten Artikel umfassen Analysen der Implementierung eines Rahmenwerks zur Qualitätsverbesserung in der Primärversorgung im Vereinigten Königreich (unter Verwendung von Differenzen-in-Differenzen-Verfahren), die Analyse der Auswirkungen des Gebärens von Kindern auf die Mundgesundheit (unter Verwendung von Instrumentalvariablen) und die Analyse einer Implementierung veränderter Anbieteranreize im dänischen Zahngesundheitssystem (unter Verwendung der Analyse von Zeitreihen). Bei all diesen Methoden werden Quasi-Experimente verwendet, die eine post-hoc-Ableitung kausaler Effekte ermöglichen, wenn bestimmte identifizierende Annahmen erfüllt sind. Die Missachtung der Kausalität in der empirischen Forschung setzt die Ergebnisse insgesamt dem Risiko enormer Verzerrungen und Fehlinterpretationen aus. Dies ist insofern problematisch, als es nicht möglich ist, den einfachen kausalen Zusammenhang zwischen zweier Variablen X und Y statistisch explizit auszudrücken - ein Problem, das als "fundamentales Problem der Kausalität" bekannt ist, bei dem sich in der realen Welt nur eines von mehreren möglichen Ergebnissen manifestiert. Diese Arbeit soll daher aufzeigen, wie die Kausalmethoden von Pearl, Camp-

bell und Rubin zur kausalen Argumentation und letztlich zu einer robusteren empirischen Forschung beitragen können.

Bibliography

Abadie, A. (2005). "Semiparametric Difference-in-Differences Estimators." In: *The Review of Economic Studies* 72.1, pp. 1–19. URL: http://www.jstor.org/stable/3700681 (cit. on p. 67).

Alcser KH, B. G. (2005). "The SHARE Train-the Trainer Program: The survey of health, aging, and retirement in Europe – methodology." In: *Mannheim, Germany: Mannheim Research Institute for the Economics of Aging (MEA)*, pp. 70–81 (cit. on p. 39).

Althubaiti, A. (May 2016). "Information bias in health research: definition, pitfalls, and adjustment methods." In: *Journal of Multidisciplinary Healthcare*, p. 211. URL: https://doi.org/10.2147/jmdh.s104807 (cit. on p. 19).

Andrews, D. W. K. (1991). "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." In: *Econometrica* 59.3, pp. 817–858. URL: http://www.jstor.org/stable/2938229 (cit. on p. 56).

Angrist, J., W. Evans (Sept. 1996). *Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size*. Tech. rep. URL: https://doi.org/10.3386/w5778 (cit. on p. 90).

Angrist, J. D., W. N. Evans (1998). "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." In: *The American Economic Review* 88.3, pp. 450–477 (cit. on p. 90).

Angrist, J. D., A. B. Krueger (Nov. 2001). "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." In: *Journal of Economic Perspectives* 15.4, pp. 69–85. URL: https://doi.org/10.1257/jep.15.4.69 (cit. on pp. 41, 42).

Angrist, J. D., J.-S. Pischke (Dec. 2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press (cit. on p. 55).

Bauer, M. S., L. Damschroder, H. Hagedorn, J. Smith, A. M. Kilbourne (Sept. 2015). "An introduction to implementation science for the non-specialist." In: *BMC Psychology* 3.1. URL: https://doi.org/10.1186/s40359-015-0089-9 (cit. on p. 101).

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al. (2018). "Redefine statistical significance." In: *Nature Human Behaviour* 2.1, p. 6 (cit. on p. 14).

Berk, R. A. (1983). "An Introduction to Sample Selection Bias in Sociological Data." In: *American Sociological Review* 48.3, pp. 386–398. URL: http://www.jstor.org/stable/2095230 (cit. on p. 19).

Bernal, J. L., S. Cummins, A. Gasparrini (June 2016). "Interrupted time series regression for the evaluation of public health interventions: a tutorial." In: *International Journal of Epidemiology*, dyw098. URL: https://doi.org/10.1093/ije/dyw098 (cit. on p. 55).

Berwick, D. M. (2016). "Era 3 for medicine and health care." In: *Jama* 315.13, pp. 1329–1330 (cit. on p. 95).

Black, S. E., P. J. Devereux, K. G. Salvanes (2005). "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education." In: *The Quarterly Journal of Economics* 120.2, pp. 669–700 (cit. on p. 43).

Börsch-Supan, A. (2016). *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5* (cit. on p. 88).

Börsch-Supan, A. (2019). *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5*. eng. URL: http://www.share-project.org/data-documentation/waves-overview/wave-5.html (cit. on p. 39).

Borsch-Supan, A., M. Brandt, C. Hunkler, et al. (2013). "Data Resource Profile: the Survey of Health, Ageing and Retirement in Europe (SHARE)." In: *International journal of epidemiology* 42.4, pp. 992–1001 (cit. on p. 39).

Bound, J., D. A. Jaeger, R. M. Baker (1995). "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak." In: *Journal of the American Statistical Association* 90.430, p. 443 (cit. on pp. 48, 74).

Brocklehurst, P., J. Price, A.-M. Glenny, M. Tickle, S. Birch, E. Mertz, J. Grytten (2013). "The effect of different methods of remuneration on the behaviour of primary care dentists." In: *The Cochrane database of systematic reviews* 11, p. CD009853 (cit. on p. 97).

Caceres-Delpiano, J., M. Simonsen (2012). “The toll of fertility on mothers’ wellbeing.” In: *Journal of health economics* 31.5, pp. 752–766 (cit. on p. 44).

Cagetti, M. G., G. Bontà, F. Cocco, P. Lingstrom, L. Strohmenger, G. Campus (July 2018). “Are standardized caries risk assessment models effective in assessing actual caries status and future caries increment? A systematic review.” In: *BMC Oral Health* 18.1. URL: https://doi.org/10.1186/s12903-018-0585-4 (cit. on p. 98).

Calhaz-Jorge, C., C. de Geyter, M. S. Kupka, et al. (2016). “Assisted reproductive technology in Europe, 2012: results generated from European registers by ESHRE.” In: *Human reproduction (Oxford, England)* 31.8, pp. 1638–1652 (cit. on p. 41).

Cambell, D. T., J. Stanley (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally & Company (cit. on pp. 32, 53).

Campbell, D. T. (1957). “Factors relevant to the validity of experiments in social settings.” In: *Psychological Bulletin* 54.4, pp. 297–312. URL: https://doi.org/10.1037/h0040950 (cit. on p. 32).

Campbell, D., E. Overman, U. of Chicago. Press (1988). *Methodology and Epistemology for Social Sciences: Selected Papers*. University of Chicago Press. URL: https://books.google.de/books?id=m-E1dCzVFRYC (cit. on p. 33).

Campbell, D., J. Stanley (2015). *Experimental and Quasi-Experimental Designs for Research*. Ravenio Books. URL: https://books.google.de/books?id=KCTrCgAAQBAJ (visited on 05/05/2020) (cit. on p. 17).

Carver, R. (1978). “The case against statistical significance testing.” In: *Harvard Educational Review* 48.3, pp. 378–399 (cit. on p. 14).

Chalkley, M., S. Listl (2018). “First do no harm - The impact of financial incentives on dental X-rays.” In: *Journal of health economics* 58, pp. 1–9 (cit. on p. 97).

Chow, S. L. (1997). *Statistical Significance: Rationale, Validity and Utility (Introducing Statistical Methods)*. SAGE Publications Ltd. URL: https://books.google.de/books?id=0DWeqYsehDsC (cit. on p. 14).

Clough, S., Z. Shehabi, C. Morgan (2016). “Medical risk assessment in dentistry: Use of the American Society of Anesthesiologists Physical Status Classification.” In: *British dental journal* 220.3, pp. 103–108 (cit. on p. 98).

Cochran, W. G., D. B. Rubin (1973). “Controlling Bias in Observational Studies: A Review.” In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 35.4, pp. 417–446. URL: http://www.jstor.org/stable/25049893 (cit. on p. 18).

Cohen, A. N., A. B. Hamilton, M. Ritchie, B. S. Mittman, J. E. Kirchner, G. E. Wyatt, J. C. Fortney, G. Hellemann, H. Liu, G. M. Curran, F. Whelan, A. M. Eccles, L. E. Parker, K. McNagny, C. S. Hutchinson, A. B. Teague, C. Reist, A. S. Young (2015). "Improving care quality through hybrid implementation/effectiveness studies: Best practices in design, methods, and measures." In: *Implementation Science* 10.S1, p. 1154 (cit. on p. 100).

Conrad, D. A., M. Vaughn, D. Grembowski, M. Marcus-Smith (Nov. 2015). "Implementing Value-Based Payment Reform." In: *Medical Care Research and Review* 73.4, pp. 437–457. URL: https://doi.org/10.1177/1077558715615774 (cit. on p. 101).

Cox, T. (2012). *Using a Quality Improvement Framework to Make Local Population Health Gains*. URL: https://www.nice.org.uk/sharedlearning/using-a-quality-improvement-framework-to-make-local-population-health-gains (cit. on p. 60).

Crown, W. H., H. J. Henk, D. J. Vanness (Dec. 2011). "Some Cautions on the Use of Instrumental Variables Estimators in Outcomes Research: How Bias in Instrumental Variables Estimators Is Affected by Instrument Strength, Instrument Contamination, and Sample Size." In: *Value in Health* 14.8, pp. 1078–1084. URL: https://doi.org/10.1016/j.jval.2011.06.009 (cit. on p. 49).

Cummins, S., J. Lopez Bernal, A. Gasparrini (July 2018a). "The use of controls in interrupted time series studies of public health interventions." In: *International Journal of Epidemiology* 47.6, pp. 2082–2093. eprint: http://oup.prod.sis.lan/ije/article-pdf/47/6/2082/27015921/dyy135.pdf. URL: https://doi.org/10.1093/ije/dyy135 (cit. on p. 56).

– (July 2018b). "The use of controls in interrupted time series studies of public health interventions." In: *International Journal of Epidemiology* 47.6, pp. 2082–2093. eprint: http://oup.prod.sis.lan/ije/article-pdf/47/6/2082/27015921/dyy135.pdf. URL: https://doi.org/10.1093/ije/dyy135 (cit. on p. 57).

Dale, S. B., A. B. Krueger (2002). "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables." In: *The Quarterly Journal of Economics* 117.4, pp. 1491–1527. URL: http://www.jstor.org/stable/4132484 (cit. on p. 43).

De Finetti, B. (1972). "Probability, induction, and statistics." In: (cit. on p. 18).

Deaton, A., N. Cartwright (2018). “Understanding and misunderstanding randomized controlled trials.” In: *Social Science & Medicine* 210, pp. 2–21. URL: https://doi.org/10.1016/j.socscimed.2017.12.005 (cit. on p. 67).

Dimick, J. B., A. M. Ryan (2014). “Methods for evaluating changes in health care policy: the difference-in-differences approach.” In: *Jama* 312.22, pp. 2401–2402 (cit. on p. 63).

Doran, T., C. Fullwood, E. Kontopantelis, D. Reeves (2008). “Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework.” In: *The Lancet* 372.9640, pp. 728–736 (cit. on p. 95).

Douglass, C. W., J. Berlin, S. Tennstedt (1991). “The validity of self-reported oral health status in the elderly.” In: *Journal of public health dentistry* 51.4, pp. 220–222 (cit. on p. 41).

Dunning, T. (2008). “Model Specification in Instrumental-Variables Regression.” In: *Political Analysis* 16.3, 290–302 (cit. on p. 48).

Ellenberg, J. H. (Mar. 1994). “Selection bias in observational and experimental studies.” In: *Statistics in Medicine* 13.5-7, pp. 557–567. URL: https://doi.org/10.1002/sim.4780130518 (cit. on p. 107).

Escoffery, C., E. Lebow-Skelley, R. Haardoerfer, E. Boing, H. Udelson, R. Wood, M. Hartman, M. E. Fernandez, P. D. Mullen (2018). “A systematic review of adaptations of evidence-based public health interventions globally.” In: *Implementation science : IS* 13.1, p. 125 (cit. on p. 101).

Fanelli, D. (2009). “How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data.” In: *PLoS ONE* 4.5. Ed. by T. Tregenza, e5738. URL: https://doi.org/10.1371/journal.pone.0005738 (cit. on p. 109).

Fedak, K. M., A. Bernal, Z. A. Capshaw, S. Gross (2015). “Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology.” In: *Emerging Themes in Epidemiology* 12.1. URL: https://doi.org/10.1186/s12982-015-0037-4 (cit. on pp. 23, 24).

Fowke, F. (Mar. 1885). “On the First Discovery of the Comma-Bacillus of Cholera.” In: *BMJ* 1.1264, pp. 589–592. URL: https://doi.org/10.1136/bmj.1.1264.589 (cit. on p. 48).

Gabel, F., H. Jürges, K. E. Kruk, S. Listl (Mar. 2018). “Gain a child, lose a tooth? Using natural experiments to distinguish between fact and fiction.” In: *Journal*

of Epidemiology and Community Health 72.6, pp. 552–556. URL: https://doi.org/10.1136/jech-2017-210210 (cit. on pp. 39, 41).

Gennetian, L., P. Morris, J. Bos, H. Bloom (2005). *Constructing instrumental variables from experimental data to explore how treatments produce effects*. Russell Sage Foundation, pp. 75–111 (cit. on p. 49).

Gerritsen, A. E., P. F. Allen, D. J. Witter, et al. (2010). "Tooth loss and oral health-related quality of life: a systematic review and meta-analysis." In: *Health and quality of life outcomes* 8, p. 126 (cit. on p. 39).

Gerstman, B. B. (2013). *Epidemiology Kept Simple: An Introduction to Traditional and Modern Epidemiology*. Wiley-Blackwell (cit. on p. 21).

Gilbert, G. H., R. P. Duncan, A. M. Kulley (1997). "Validity of self-reported tooth counts during a telephone screening interview." In: *Journal of public health dentistry* 57.3, pp. 176–180 (cit. on p. 41).

Good, P. I. (Nov. 2002). "Extensions Of The Concept Of Exchangeability And Their Applications." In: *Journal of Modern Applied Statistical Methods* 1.2, pp. 243–247. URL: https://doi.org/10.22237/jmasm/1036110240 (visited on 05/05/2020) (cit. on p. 17).

Greenhalgh, T., G. Robert, F. Macfarlane, P. Bate, O. Kyriakidou (2004a). "Diffusion of innovations in service organizations: systematic review and recommendations." In: *The Milbank Quarterly* 82.4, pp. 581–629 (cit. on p. 59).

– (2004b). "Diffusion of innovations in service organizations: systematic review and recommendations." In: *The Milbank quarterly* 82.4, pp. 581–629 (cit. on p. 99).

Grimshaw, J. M., L. Shirran, R. Thomas, G. Mowatt, C. Fraser, L. Bero, R. Grilli, E. Harvey, A. Oxman, M. A. O'Brien (2001). "Changing provider behavior: an overview of systematic reviews of interventions." In: *Medical care* 39.8 Suppl 2, pp. II2–45 (cit. on p. 100).

Guo, R., L. Cheng, J. Li, P. R. Hahn, H. Liu (2018). "A Survey of Learning Causality with Data: Problems and Methods." In: eprint: arXiv:1809.09337 (cit. on p. 31).

Hammer, G. P., J.-B. du Prel, M. Blettner (Oct. 2009). "Avoiding Bias in Observational Studies." In: *Deutsches Aerzteblatt Online*. URL: http://doi.org/10.3238/arztebl.2009.0664 (cit. on p. 107).

Hartge, P. (Mar. 2015). "A Dictionary of Epidemiology, Sixth EditionEdited by Miquel Porta." In: *American Journal of Epidemiology* 181.8, pp. 633–634. URL: https://doi.org/10.1093/aje/kwv031 (cit. on p. 19).

Healthcare in Denmark: An overview (2016). Ministry of Health (cit. on p. 97).

Heckman, J. J. (1979). “Sample Selection Bias as a Specification Error.” In: *Econometrica* 47.1, pp. 153–161. URL: http://www.jstor.org/stable/1912352 (cit. on pp. 19, 107).

Hill, A. B. (May 1965a). “The Environment and Disease: Association or Causation?” In: *Proceedings of the Royal Society of Medicine* 58.5, pp. 295–300. URL: https://doi.org/10.1177/003591576505800503 (cit. on p. 104).

Hill, S. A. B. (1965b). “The Environment and Disease: Association or Causation?” In: *Proceedings of the Royal Society of Medicine* 58.5. PMID: 14283879, pp. 295–300 (cit. on p. 23).

Holland, P. W. (1986). “Statistics and Causal Inference.” In: *Journal of the American Statistical Association* 81.396, pp. 945–960. URL: http://www.jstor.org/stable/2289064 (cit. on pp. 16, 26).

Hsiao, C. (2014). *Analysis of Panel Data*. 3rd ed. Econometric Society Monographs. Cambridge University Press (cit. on p. 69).

Hume, D. (1739/1978). *A treatise of human nature*. Oxford: Oxford University Press (cit. on p. 23).

Hylleberg, S., ed. (1992). *Modelling Seasonality*. Oxford University Press. URL: https://EconPapers.repec.org/RePEc:oxp:obooks:9780198773184 (cit. on p. 56).

Imai, K., L. Keele, D. Tingley (2010). “A general approach to causal mediation analysis.” In: *Psychological Methods* 15.4, pp. 309–334. URL: https://doi.org/10.1037/a0020761 (cit. on p. 106).

Imbens, G. W., J. D. Angrist (1994). “Identification and Estimation of Local Average Treatment Effects.” In: *Econometrica* 62.2, p. 467 (cit. on p. 91).

Kline, A. D. (1980). “Are There Cases of Simultaneous Causation?” In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1980, pp. 292–301. URL: http://www.jstor.org/stable/192573 (cit. on p. 22).

Krauss, A. (2018). “Why all randomised controlled trials produce biased results.” In: *Annals of Medicine* 50.4, pp. 312–322. URL: https://doi.org/10.1080/07853890.2018.1453233 (cit. on p. 67).

Kruk, K. E., S. Reinhold (2014a). “The effect of children on depression in old age.” In: *Social science & medicine (1982)* 100, pp. 1–11 (cit. on pp. 44, 90).

– (Jan. 2014b). “The effect of children on depression in old age.” In: *Social Science & Medicine* 100, pp. 1–11. URL: https://doi.org/10.1016/j.socscimed.2013.09.003 (cit. on p. 90).

Lindley, D. V., M. R. Novick (1981). “The Role of Exchangeability in Inference.” In: *The Annals of Statistics* 9.1, pp. 45–58. URL: http://www.jstor.org/stable/2240868 (cit. on p. 18).

Listl, S., H. Jürges, R. G. Watt (2016). “Causal inference from observational data.” In: *Community dentistry and oral epidemiology* 44.5, pp. 409–415 (cit. on p. 39).

Loken, E., A. Gelman (2017). “Measurement error and the replication crisis.” In: *Science* 355.6325, pp. 584–585. eprint: https://science.sciencemag.org/content/355/6325/584.full.pdf. URL: https://science.sciencemag.org/content/355/6325/584 (cit. on p. 14).

Lousdal, M. L. (Jan. 2018). “An introduction to instrumental variable assumptions, validation and estimation.” In: *Emerging Themes in Epidemiology* 15.1. URL: https://doi.org/10.1186/s12982-018-0069-7 (cit. on p. 46).

Malter, F. and A. Börsch-Supan (2015). “SHARE Wave 5: Innovations & Methodology. Munich: MEA, Max Planck Institute for Social Law and Social Policy.” In: (cit. on p. 39).

Marcenes, W., N. J. Kassebaum, E. Bernabé, et al. (2013). “Global burden of oral conditions in 1990-2010: a systematic analysis.” In: *Journal of dental research* 92.7, pp. 592–597 (cit. on p. 39).

McMichael, A. J. (1976). “Standardized mortality ratios and the "healthy worker effect": Scratching beneath the surface.” In: *J Occup Med* 18.3, pp. 165–168 (cit. on p. 19).

Mill, J. S. (1843). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence, and methods of scientific investigation*. London: J. W. Parker (cit. on p. 23).

Mill, J. S. (2009). *A System of Logic, Ratiocinative and Inductive*. Cambridge University Press. URL: https://doi.org/10.1017/cbo9781139149839 (cit. on p. 23).

Min, S.-J., B. Luke, B. Gillespie, et al. (2000). “Birth weight references for twins.” In: *American Journal of Obstetrics and Gynecology* 182.5, pp. 1250–1257 (cit. on p. 91).

NHS England (2013). “Commissioning for Value.” In: URL: https://www.england.nhs.uk/rightcare/wp-content/uploads/sites/40/2017/01/cfv-nene-jan17.pdf (cit. on p. 94).

Pandian, Z., A. Gibreel, S. Bhattacharya (2015). “In vitro fertilisation for unexplained subfertility.” In: *The Cochrane database of systematic reviews* 11, p. CD003357 (cit. on p. 41).

Pashler, H., E. Wagenmakers (2012). “Editors’ Introduction to the Special Section on Replicability in Psychological Science.” In: *Perspectives on Psychological Science* 7.6, pp. 528–530. URL: https://doi.org/10.1177/1745691612465253 (cit. on p. 109).

Pearl, J. (1986). “Fusion, Propagation and Structuring in Belief Networks.” In: *Artificial Intelligence* 29.3, pp. 241–288 (cit. on p. 32).

Pearl, J. (Dec. 1995). “Causal diagrams for empirical research.” In: *Biometrika* 82.4, pp. 669–688. eprint: http://oup.prod.sis.lan/biomet/article-pdf/82/4/669/698263/82-4-669.pdf. URL: https://doi.org/10.1093/biomet/82.4.669 (cit. on pp. 29, 31, 32).

– (Sept. 2009). *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press (cit. on pp. 15, 26, 29, 31).

Pearl, J., D. Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. 1st. New York, NY, USA: Basic Books, Inc. (cit. on pp. 15, 17).

Penfold, R. B., F. Zhang (Nov. 2013). “Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements.” In: *Academic Pediatrics* 13.6, S38–S44. URL: https://doi.org/10.1016/j.acap.2013.08.002 (cit. on p. 57).

Petersen, P. E. (2003). “The World Oral Health Report 2003: continuous improvement of oral health in the 21st century–the approach of the WHO Global Oral Health Programme.” In: *Community dentistry and oral epidemiology* 31 Suppl 1, pp. 3–23 (cit. on p. 97).

Popper, K (1959). *The Logic of Scientific Discovery*. Hutchinson (cit. on p. 14).

Puhani, P. (2000). “The Heckman Correction for Sample Selection and Its Critique.” In: *Journal of Economic Surveys* 14.1, pp. 53–68. URL: https://doi.org/10.1111/1467-6419.00104 (visited on 05/05/2020) (cit. on pp. 19, 107).

Ramos, R. Q., J. L. Bastos, M. A. Peres (2013). “Diagnostic validity of self-reported oral health outcomes in population surveys: literature review.” In: *Revista brasileira de epidemiologia = Brazilian journal of epidemiology* 16.3, pp. 716–728 (cit. on p. 41).

Rosing, K. (June 2015). “The Danish dental health monitoring system for adults.” PhD thesis (cit. on p. 97).

Rothman, K. J., S. Greenland (July 2005). “Causation and Causal Inference in Epidemiology.” In: *American Journal of Public Health* 95.S1, S144–S150. URL: https://doi.org/10.2105/ajph.2004.059204 (cit. on p. 23).

Rothstein, J. (Oct. 2009). “Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables.” In: *Education Finance and Policy* 4.4, pp. 537–571. URL: https://doi.org/10.1162/edfp.2009.4.4.537 (cit. on p. 43).

Rubin, D. B. (1974). “Estimating causal effects of treatments in randomized and nonrandomized studies.” In: *Journal of Educational Psychology* 66, pp. 688–701 (cit. on p. 25).

Rubin, D. (Dec. 1976). “Inference and missing data.” In: *Biometrika* 63.3, pp. 581–592. eprint: http://oup.prod.sis.lan/biomet/article-pdf/63/3/581/756166/63-3-581.pdf. URL: https://doi.org/10.1093/biomet/63.3.581 (cit. on p. 107).

Rubin, D. B. (1977). “Assignment to Treatment Group on the Basis of a Covariate.” In: *Journal of Educational Statistics* 2.1, pp. 1–26. URL: http://www.jstor.org/stable/1164933 (cit. on p. 25).

– (1978). “Bayesian Inference for Causal Effects: The Role of Randomization.” In: *The Annals of Statistics* 6.1, pp. 34–58. URL: http://www.jstor.org/stable/2958688 (cit. on p. 25).

Selvin, H. C. (1958). “Durkheim’s Suicide and Problems of Empirical Research.” In: *American Journal of Sociology* 63.6, pp. 607–619. eprint: https://doi.org/10.1086/222356. URL: https://doi.org/10.1086/222356 (cit. on p. 57).

Shadish, W. R. (2010). “Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings.” In: *Psychological Methods* 15.1, pp. 3–17. URL: https://doi.org/10.1037/a0015916 (cit. on p. 32).

Shpitser, I. (Jan. 2008). “Complete identification methods for causal inference.” PhD thesis (cit. on p. 31).

Shpitser, I., J. Pearl (Sept. 2008). “Complete Identification Methods for the Causal Hierarchy.” In: *Journal of Machine Learning Research* 9, pp. 1941–1979 (cit. on p. 16).

Shumway, R. (1988). *Applied statistical time series analysis. Applied statistical time series analysis / Shumway, Robert H.. - Englewood Cliffs, NJ : Prentice-Hall, 1988 [Hauptbd.]* Prentice-Hall series in statistics. Englewood Cliffs, NJ: Prentice-Hall. XV, 379. URL: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+125595492&sourceid=fbw_bibsonomy (cit. on p. 52).

Silveira Moreira, R. d. (2012). “Epidemiology of Dental Caries in the World.” In: *Oral Health Care - Pediatric, Research, Epidemiology and Clinical Practices*. Ed. by M. Virdi. InTech (cit. on p. 97).

Sims, C. A. (1974). “Seasonality in Regression.” In: *Journal of the American Statistical Association* 69.347, pp. 618–626. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1974.10480178. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10480178 (cit. on p. 56).

Snow, J. (1855). *On the mode of communication of cholera*. London: John Churchill (cit. on p. 47).

Soumerai, S. B., D. Starr, S. R. Majumdar (June 2015). “How Do You Know Which Health Care Effectiveness Research You Can Trust? A Guide to Study Design for the Perplexed.” In: *Preventing Chronic Disease* 12. URL: https://doi.org/10.5888/pcd12.150187 (cit. on p. 57).

Spirtes, P., C. Glymour, R. Scheines (1993). *Causation, Prediction, and Search*. Springer New York. URL: https://doi.org/10.1007/978-1-4612-2748-9 (cit. on p. 30).

Sterne, J. A. C., I. R White, J. B Carlin, M. Spratt, P. Royston, M. G Kenward, A. M Wood, J. R Carpenter (June 2009). “Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.” In: *BMJ* 338.jun29 1, b2393–b2393. URL: https://doi.org/10.1136/bmj.b2393 (cit. on p. 107).

Sundhedsstyrelsen (2017). “Evaluering af National Klinisk Retningslinje for fastlæggelse af intervaller mellem diagnostiske undersøgelser i tandplejen.” In: (cit. on pp. 97, 101).

The Significance Test Controversy: A Reader (2006). Aldine Transaction (cit. on p. 14).

Trent Clinical Commissioning Group, S. on (2015). *Stoke-on-Trent Clinical Commissioning Group Annual Report*. URL: https://www.stokeccg.nhs.uk/stoke-governance/annual-reports/2015-3/157-afd-1149-stoke-on-trent-annual-report-28-05-15-v10-final-for-publishing/file (cit. on p. 58).

Twetman, S., M. Fontana, J. D. B. Featherstone (2013). “Risk assessment - can we achieve consensus?” In: *Community dentistry and oral epidemiology* 41.1, e64–70 (cit. on p. 98).

Ueno, M., S. Ohara, M. Inoue, et al. (2013). "Association between parity and dentition status among Japanese women: Japan public health center-based oral health study." In: *BMC public health* 13, p. 993 (cit. on p. 90).

UNESCO Institute for Statistics (2012). *International standard classification of education: ISCED 2011*. Montreal, Quebec: UNESCO Institute for Statistics (cit. on p. 41).

Voinea-Griffin, A., D. B. Rindal, J. L. Fellows, A. Barasch, G. H. Gilbert, M. M. Safford (2010). "Pay-for-performance in dentistry: what we know." In: *Journal for healthcare quality : official publication of the National Association for Healthcare Quality* 32.1, pp. 51–58 (cit. on p. 101).

Wasserstein, R. L., N. A. Lazar (2016). "The ASA Statement on p-Values: Context, Process, and Purpose." In: *The American Statistician* 70.2, pp. 129–133. eprint: https://doi.org/10.1080/00031305.2016.1154108. URL: https://doi.org/10.1080/00031305.2016.1154108 (cit. on p. 15).

Witter, S., J. Toonen, B. Meessen, J. Kagubare, G. Fritsche, K. Vaughan (2013). "Performance-based financing as a health system reform: mapping the key dimensions for monitoring and evaluation." In: *BMC health services research* 13, p. 367 (cit. on p. 101).

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Mit Press. URL: http://www.jstor.org/stable/j.ctt5hhcfr (cit. on p. 53).

Wright, S. (1921). "Correlation and causation." In: *Journal of Agricultural Research* 20, pp. 557–585 (cit. on pp. 29, 31).

Wunsch, G., F. Russo, M. Mouchart (Apr. 2010). "Do We Necessarily Need Longitudinal Data to Infer Causal Relations?" In: *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 106.1, pp. 5–18. URL: https://doi.org/10.1177/0759106309360114 (cit. on p. 52).

Chapter 6

Appendix

Individual Preventive Treatment (IPT)

Applicability:	IPT can be used in the presence of active caries, gingivitis, mucositis around implants, marginal periodontitis and periimplantitis. It may also be used in the presence of other oral disorders which require preventive treatment. Diagnoses underlying the preventive treatment must be recorded in patients clinical documents. In order to perform this service, patients need to be classified as being in either the yellow or the red risk group.
Details:	• Detection of the extent of the observed disease incidence. It is explained to the patient how the condition is known by symptoms and changes on the tooth surfaces, in the gums and/or mucous membranes. • Tailored instruction in preventive measures for the detected condition and presentation of available treatment options. • Detection and demonstration of disease-causing plaque and of general plaque retaining factors. • Instruction in self-care, with special attention to individual needs appropriate for the individual patient. • Information should be provided on risks related to tobacco use and on specific damages tobacco can cause in the oral cavity and on the importance of a heathy diet. • Provision of fluoride treatments of active caries lesions (maximum four times a year). Caries lesions must be cleaned professionally beforehand (dental floss or polish). • Scale and/or polish for removal of plaque - if not contained in other dental service within the same dental visit, to support an understanding of good oral hygiene.

Table 6.1: An overview of the most important remuneration codes in the Danish dental health care system. A detailed version of treatment items can be found in the Appendix (Table A1).

Status examination (SE)

Applicability:	The SE service forms the basis for the planning of necessary preventive and treatment efforts until the next SE or FE service can be carried out.
Details:	• Update of medical history. • Removal of plaque. • Clinical examination of teeth, periodontium, oral cavity, mucosal surfaces, tongue and jaws. • Screening for occlusal interferences. • Update of charting tooth restorations and replacements. • Assessment of disease progression. • Treatment planning including patient information and involvement. • Identification of risk factors. • Diagnostics • Determining the interval until the next examination based on risk assessment and the individual need of the patient. • Categorization of the patient according to the Health Authorities´ guidelines from 2013 into green, yellow or red risk-category.

Table 6.2: An overview of the most important remuneration codes in the Danish dental health care system. A detailed version of treatment items can be found in the Appendix (Table A1).

Focused examination (FE)

Applicability:	A focused examination is a follow-up examination focused on a current disease problem.
Details:	• Update of medical history. • Removal of plaque. • Update of clinical examination focusing on progression of previously diagnosed disease. • Update of diagnoses. • Re-instruction in self-care, if necessary. • Update of treatment planning. • Determining the interval for the next focused examination or status examination based on individual risk.

Table 6.3: An overview of the most important remuneration codes in the Danish dental health care system. A detailed version of treatment items can be found in the Appendix (Table A1).

EIGENANTEIL AN DATENERHEBUNG UND -AUSWERTUNG UND EIGENE VERÖFFENTLICHUNGEN

Teile dieser Arbeit entstanden im Rahmen des EU-Projekts ADVOCATE (Added Value for Oral Care).

Die Datenauswertung und -interpretation aus Kapitel 3 wurden vollständig von mir durchgeführt. Die Datenerhebung der Fallstudie aus Kapitel 2.1 wurde von mir selbst durchgeführt, die Datenerhebung für Fallstudie 2.2 war Teil des im Universitätsklinikum Heidelberg verorteten Arbeitspakets und wurde in Zusammenarbeit mit europäischen Krankenversicherungen zu Projektbeginn von mir selbst durchgeführt und die Datenerhebung für Fallstudie 2.3 wurde von Prof. Neal Maskrey durchgeführt.

Teilergebnisse der vorliegenden Arbeit wurden in folgenden Aufsätzen vorab publiziert:

1. **Gabel F.**, Chambers R., Cox T., Listl S., Maskrey N. (2019). An evaluation of a multifaceted, local Quality Improvement Framework for long-term conditions in UK primary care. *Family Practice* 36/5, 607–613.
2. **Gabel F.**, Jürges H., Kruk K.-E., Listl S. (2018). Gain a child, lose a tooth? Using natural experiments to distinguish between fact and fiction. *Journal of Epidemiology and Community Health* 72, 552-556.
3. **Gabel F.**, Kalmus O., Rosing K., Trescher A.-L., Listl S. (2020). Implementation of altered provider incentives for a more individual-risk-based assignment of dental recall intervals: evidence from a health systems reform in Denmark. *Health Economics* 29, 475-488.

Publikation 1 basiert auf den Ergebnissen der in Kapitel 2.3 vorgestellten Fallstudie und wurde in enger Zusammenarbeit mit Prof. Neal Maskrey erstellt. Bei der Manuskripterstellung wurde ich von Prof. Maskrey assistiert, die statistischen Analysen wurden vollständig von mir durchgeführt. Der Eigenanteil an dieser Publikation lag bei 70 %.

Publikation 2 basiert auf den Ergebnissen der in Kapitel 2.1 vorgestellten

Fallstudie. Die Konzeption ging hierbei auf Prof. Stefan Listl zurück, die statistischen Analysen sowie die Manuskripterstellung wurden vollständig von mir durchgeführt. Der Eigenanteil an dieser Publikation lag bei 80 %.

Publikation 3 basiert auf den Ergebnissen der in Kapitel 2.2 vorgestellten Fallstudie. Die Konzeption ging hierbei auf Kasper Rosing und Stefan Listl zurück, die statistischen Analysen sowie die Manuskripterstellung wurden vollständig von mir durchgeführt. Der Eigenanteil an dieser Publikation lag bei 80 %.

Weitere eigene Veröffentlichungen:

1. **Gabel F.**, O'Hanlon K., Brankin P., Bryce R., Trescher, A.-L., Whelton, H., van der Heijden, G., Listl, S. (2016). Linkage of health care claims data and apps data: The ADVOCATE oral health care dashboard. *Population Data Science Conference, Swansea, Wales*.
2. Haux C., Kalmus O., Trescher A.-L., **Gabel F.**, Listl S., Knaup P. (2018). Detecting and resolving Data Conflicts. *Studies in Health Technology and Informatics* 247, 1-5.
3. Clovis F., Diaz K. T.., Aranda L., **Gabel F.**, Listl S., Alarcon M. A.. (2016). The risk of bias of animal experiments in implant dentistry: a methodological study. *Clinical Oral Implants Research* 38/7, 39-45.

DANKSAGUNG

An dieser Stelle möchte ich noch einmal die Gelegenheit nutzen, all jenen zu danken, die zum Gelingen dieser Arbeit beigetragen haben.
Besonderer Dank gilt zunächst meinem Betreuer Herr Prof. Dr. Dr. Stefan Listl für stets konstruktiven Rat, hilfreiche Diskussionen und die Sicherstellung höchster wissenschaftlicher Standards.
Den Kolleginnen und Kollegen aus dem Forschungsprojekt ADVOCATE danke ich für den vielseitigen inhaltlichen Anregungen und den stets angenehmen Austausch. Ebenso danke ich all meinen KoautorInnen für hilfreiche Diskussionen, Kommentare und die freundliche Zusammenarbeit.
Mein herzlicher Dank gilt auch meinem Kollegen Dr. Olivier Kalmus und meiner Kollegin Dr. Anna-Lena Trescher für interessante fachliche sowie nicht-fachliche Diskussionen im Büro 11.303. Diese freundschaftliche Arbeitsatmosphäre werde ich vermissen.
Zuletzt möchte ich meiner Familie danken: meinen Eltern, Schwester und Bruder. Danke, dass ihr mir immer liebevoll und unterstützend zur Seite steht - dieser Rückhalt macht alles leichter.

EIDESSTATTLICHE VERSICHERUNG

1. Bei der eingereichten Dissertation zu dem Thema **Empirical Methods for Causal Inference** handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort, Datum Unterschrift