More Than Just Associations: An Introduction to Causal Inference for Sport Science

Master thesis

From

Simon Nolte

German Sport University Cologne

Cologne 2024

Thesis supervisor:

Dr. Oliver Jan Quittmann

Institute of Movement and Neurosciences

Affirmation in lieu of an oath

Herewith I affirm in lieu of an oath that I have authored this Bachelor thesis independently and did not use any other sources and tools than indicated. All citations, either direct quotations or passages which were reproduced verbatim or nearby-verbatim from publications, are indicated and the respective references are named. The same is true for tables and figures. I did not submit this piece of work in the same or similar way or in extracts in another assignment.

_____

Personally signed

Abstract

Zusammenfassung (German Abstract)

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Relevance

Empirical research is acquiring knowledge through systematic observations by analyzing data. Data analysis typically encompasses three primary tasks: description, prediction, and causal inference (Carlin & Moreno-Betancur, 2023; Hernán et al., 2019). Description means characterizing features in a subset of a population. Prediction means forecasting outcomes based on available data. Causal inference means making claims about causality — what would have happened under different circumstances.

Most research in sport science is of causal nature. We want to understand how sports works with the ultimate goal to intervene: If we understand why certain people or teams are winning a competition, we can use that knowledge to adjust training and tactics. Likewise, in health contexts, we seek for sport intervention that change an individual's fitness to ultimately increase well-being compared to if no intervention were undertaken. Ultimately, we are interested in potential outcomes — what would have happened if the team had played different or if the individual had undergone a different training. This exactly is causal thinking.

Research has devised a framework for conducting studies that can infer causality without knowledge of the exact underlying causal mechanisms: the randomized controlled trial (RCT). But in sport science, RCTs are often not feasible, because of the difficulty or undesirability of implementing randomized interventions, particularly in the context of elite sports (Bullock et al., 2023). Consequently, causality must often be inferred through alternative designs, such as observational studies. The field of causal inference offers tools for this particular task.

An association on its own does not inherently indicate causality, echoing the famous adage: "correlation does not imply causation." Associations observed in data may indeed stem from causality, but they can also arise from different types of bias, resulting in spurious associations. Conversely, causation does not necessarily imply correlation. Genuine causal relationships might remain obscured within the data. Distinguishing between associations and causal relationships necessitates looking beyond the data itself.

Causal data analysis requires something that is not relevant to most description and prediction tasks: A scientific model informed by expert-domain knowledge, that depicts the causal nature of the phenomena under investigation. This causal model serves as the foundation for all causal inference. By adhering to the rules implied by the causal model, we can analyze our data in a manner that allows for the estimation of causal effects. Methods of causal inference are vital to estimate causal effects from observational data. But they can also aid in designing and analyzing experiments, and even provide benefits for description and prediction analyses.

As all statistical analyses, causal modeling is not free of assumptions. Those are assumptions about the underlying data, but also about the underlying data generative process (the

world in which the data have been created). Causal modeling requires to think more clearly about these assumptions before conducting an analysis, and is in general more transparent in communicating them (Grosz et al., 2020). In a way, this is a more honest way of doing inference than relying on non-causal language when inferring causality was the actual research goal (Hernán, 2018).

I will start by establishing a working definition of causality and by providing an overview of causal inference as a research field, with its history and popular frameworks. Following this, I will outline recent applications of causal inference across various disciplines with a focus on the (sparse) literature of causal inference in sport science.

1.2 Previous Research

What causality actually means is a merely philosophical question (Illari & Russo, 2014). For the sake of this thesis, we use the framework of potential outcomes to define causality (Rubin, 1974). If we intervene on a variable and this leads to changes compared to if we had not intervened, we can define the intervention as causing the outcome. A causal effect is therefore defined by the comparison between two states, what has actually happened, and what would have potentially happened under different intervention. The intervention itself does not need to be actually possible to conduct, it can be purely hypothetical. For example, e.g., if we define the causal effect of biological sex on endurance performance we are actually asking: If we could intervene on an individual's sex (by changing it), what difference in endurance performance would we expect. We can state this without actually being able to change biological sex (when defined via chromosome[1]).

It can be easy to define causal effects, but difficult to estimate them. For estimation, we can only use real data and not hypothetical. We still want to estimate the difference between potential outcomes, with the caveat that for each unit of observation we only have one actual outcome available. Essentially, causal inference can be viewed as a missing data problem (Ding & Li, 2018). The most straightforward way to deal with this problem is using a randomization controlled research design[2], but often this is impossible or impractical.

Fisher (1925) was the first to suggest randomization as the basis to inference of causal effects in experiments. Randomized controlled designs quickly became the gold standard of experimental research (Cochran & Cox, 1957). Possibly until the 1970s it remained the common view that causal effects can only validly studied in randomized experiments, and not in observational studies. But based on the earlier invention of potential outcome notation by Neyman

[1] This "defined via X" is exactly the reason why Imbens & Rubin (2015) would oppose the effect of sex on endurance performance as being a causal statement. Their argument is that this example does not clarify what intervening on sex would actually mean. It could be (hypothetically) intervening on chromosomes, on genitalia, or on hormones. According to their view, this ambiguousness makes the statement ill-defined so that it cannot serve as a causal statement. For this thesis I follow a less strict approach in allowing causal statements that rely on (hypothetical) intervention, even if the intervention is not clearly decisive from the statement alone.

[2] See APPENDIX for the mathematical rationale behind this.

(1923), Rubin (1974) provided a framework for estimating causal effect from both experimental and observation data. This framework later termed the 'Rubin Causal Model' (Holland, 1986) remains one of the predominant approaches to causal inference from observational data (see Appendix for the mathematical notation of this framework).

Another approach to causal inference is the use of graphical models. Pioneered by Pearl (1993, 1995), directed acyclic graphs (DAGs) have become a popular tool to assist estimating causal effects. They serve as an easy tool to aid estimating causal effect (Shrier & Platt, 2008). The graph-based approach has been criticized for being unnecessary (Rubin, 2022) or requiring a vast of (often not considered) assumptions (Dawid, 2010), yet it is popular in many fields (Morgan & Winship, 2014). Other approaches to causal inference aim to bring the potential outcome framework into a graph form (Richardson & Robins, 2013), or are less structural in that they neither require potential outcomes nor graphs (Dawid, 2000). Discussion about the different frameworks of causal inference can be found elsewhere. In this thesis I will often follow Pearl's graph-based approach (Pearl, 2009), because it is in my view the most intuitive and accessible way of learning causal inference[3], but I will also consider ideas and specific methods from the potential outcome framework (Angrist & Pischke, 2009).

Causal inference, whether in the framework of potential outcomes or graphical representations is considered one of the most influental statistical ideas of the past decades (Gelman & Vehtari, 2021). While the potential outcome framework dominates contemporary economic research (Imbens, 2020), graph-based causal inference has gained wide popularity in other fields, such as epidemiology (Greenland et al., 1999; Tennant et al., 2021), psychology (Rohrer, 2018), and sociology (Morgan & Winship, 2014). These fields share similar challenges with sport science: They study complex systems (i.e., humans) and often have to rely on observational data for inference. Despite its potential value, the use of causal inference in sport science is so far limited.

It is unsurprisingly that the most active research areas of causal inference in sport science are at the intersection to the field of epidemiology (Lynch et al., 2020), mostly in the area of injury research. For researching the prevention of injuries, calls to use causal modeling are frequent (Kalkhoven, 2024; Nielsen et al., 2020; Shrier, 2007), but its actual use is rare (Rommers et al., 2021). Shrier (2007) and Hopkins (2008) were the first to propose graphical causal models for sport science. The unusual presentation in form of a slideshow by Hopkins (2008), the narrow scope on injuries by Shrier (2007), and the lack of an accessible and focused reasoning by both may have limited the impact of their ideas. Recently, Steele et al. (2020) undertook a new try to highlight the need of causal thinking and modeling in sport science. Embedded in a general model of sport research (Bishop, 2008), they used an example of strength training to introduce key elements of causal inference such as potential outcomes and causal graphs.

[3] Naturally, the proponents of other frameworks will disagree on this. E.g., Rubin argues that teaching in the potential outcome framework is the easiest way to introduce researchers to causal inference (see his comment in Dawid, 2000). But the vast majority of newer applied introductory texts are built on the graph based approach (e.g., Cunningham, 2021; Rohrer, 2018; Shrier & Platt, 2008). I think the success of these texts speak for the accessibility of graph-based approaches for causal inference.

But they rather focused on the process of answering a specific research question (in part utilizing causal inference tools) rather than explicitly introducing causal inference to sport science.

In a recent extensive debate revolving around the causal effect of muscle hypertrophy on strength, all author groups agreed on the difficulties of distinguishing associations and causal relations, and the challenge of adequately controlling experiments or using observational data for causal statements (Balshaw et al., 2017; Buckner et al., 2017; Dankel et al., 2018; Loenneke et al., 2019; Taber et al., 2019). Yet none of them mentioned causal inference as a potential way to deal with these problems until a later publication by Nuzzo et al. (2019), again exemplifying the potential usefulness, but currently low dissemination of causal inference methods in sport science. In a recent article, Kalkhoven (2024) calls for the use of graphical causal models in sports injury research. Kalkhoven (2024) concludes his text with an appeal to all sport scientist to engage with the field of causal inference. This thesis will provide sport-scientist with an accessible, field-specific introduction to causal inference.

1.3 Aim

The aim of this thesis is to bring the methods of causal inference to sport science. The overachieving goal is to demonstrate the utility and necessity of causal inference methods for data analysis in sport science. I start with demonstrate key concepts of causal models using directed acyclic graphs by introducing confounders, colliders, and conditioning rules. I then revisit two published observation studies from the field of endurance running from a causal inference perspective. Finally, I will discuss opportunities that causal inference brings to sport science as well as challenges and limitations of adopting such approaches.

I aim to make the thesis as accessible as possible to readers who are new to causal inference. Detailed methodologies of modeling and mathematical formulations will be included in the appendices. My objective is to ensure that the thesis is understandable for any sport scientist with some basic statistical education. Instead of critiquing current statistical practices in sport science, the objective of this work is to showcase the effectiveness of methods that extend beyond these practices.

2 Theoretical Background

2.1 Causality, Associations, and (In)dependence

In the preceding section we defined causality as a concept involving hypothetical interventions. When intervening on a variable X results in changes in another variable Y we assert that X causes Y. From a statistical standpoint, X and Y become dependent[4]. Conversely, an association, implies that X and Y share information; Knowledge about one variable implies knowledge about the other variable, and *vice versa*. Crucially, associations lack directionality, whereas causality is typically understood as directional[5]. Causality can be one reason for associations to arise, but other reasons for associations exist, for example a shared common cause. Consequently, both causal relations and spurious relations can produce associations and render variables dependent. It is the underlying causal model that can distinguish between mere associations and causal relationships.

2.2 Graphical Causal Models

Graphical models provide a straightforward framework for conceptualizing causal systems. Pioneered by Pearl (1995), they offer a visual representation of causal relationships, which eases development and comprehension of causal models. A graphical causal model visualizes the exposure, outcome, covariates, and their (assumed) causal relationship. In the following, we will typically denote the exposure[6] as X, the outcome as Y, and covariates with other letters. Variables in graphical causal model are linked by arrows. An arrow between X and Y means that a direct causal relationship between the two is possible (see Figure 1) . The direction of the arrow indicates the direction of causality. As depicted in Figure 1, $X \rightarrow Y$ means that X causes Y (and not the other way around). In accordance with our definition of causality, this implies that intervening on X should results in a change in Y.

The direction of causality has to be determined by theoretical knowledge; it cannot be found in the data alone. Suppose that in our first example in Figure 1, X represents biological sex and Y denotes endurance performance. It seems apparent that a causal relationship exists between them (though it is undoubtedly much more complicated than that depicted in this simple model). However, the fact that it is sex that causes performance — and not

[4] For the mathematical notation of (conditional) independence, see the appendix.

[5] There are of cause examples, where causality can be bidirectional. For example in feedback loops, such as the price and demand models in economy, changes in price cause changes in demand and the other way around. But even in this case one can argue that these are essentially two different paths of causality, that happen sequentially if observed with enough precision. For this thesis we will not deal with feedback systems, but stick with simpler models that assume purely directional causality.

[6] Exposure here is the medical term for what is often named the “independent variable” in a statistical model. It is the variable that we image our intervention on, so it does not need to be an actual *exposure* in the strict sense of the word.

the other way around — is based purely on theoretical knowledge and understanding of the world. There are neither randomized trials available (because you cannot randomly assign sex), nor are controlled interventions feasible (because you cannot easily intervene on sex) possible. Ultimately, the direction of causality is an assumption by the researcher.

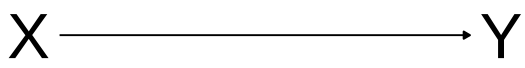

Figure 1: A simple graphical causal model with two variables. The variable X (exposure) is assumed to cause the variable Y (outcome). No other variables are believed to influence this process.

Causal systems in the world are typically more complex than consisting of only exposure and outcome, and thus the graphical causal models depicting them are more complex as well. A slightly more complex graph is displayed in Figure 2. X and Y are not directly linked anymore, but are connected indirectly via B. This sequence $X \rightarrow B \rightarrow Y$ is called a *causal path*. We will later see, that some models also have non-causal paths.

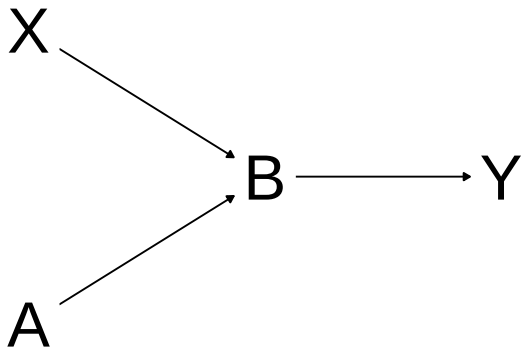

Figure 2: A more complex graphical causal model with four variables. X and A both cause B, which in turn causes Y.

The graph in Figure 2 is called a directed acyclic graph (DAG). It is directed, because all paths have arrows (establishing the direction of causality). It is acyclic, because there are no circular paths in it. Finally, it is a graph. All graphs in this thesis will be DAGs, as many of the concepts presented herein require this, and most research problems can be adequately formulated using them. More important than which arrows a DAG contains is which it arrows are absent. A DAG should depict all *potential* causal relations relevant to the research question. If two variables are not connected, we explicitly assume that they do not causally relate to each

other[7]. For example, in Figure 2, there is no direct link between X and A, or between X and Y.

DAGs tell a story. For example, we can assign the variables in Figure 2 to a simple model of endurance performance. Let X be the biological sex, A the nutrition status, B the physiological capacity to perform endurance tasks, and Y the endurance performance in a competition. Our model assumes that sex and nutrition both directly affect the physiological capacity, which subsequently affects performance. Conversely, it assumes that sex and nutrition are not causally related, and that neither direct affects performance; rather their effect are indirect mediated through physiological capacity.

2.3 Modeling Causal Systems & Error Terms

DAGs serve as an abstract concept to describe research problems. This level of abstraction allows to plan a study and its data analysis on a conceptual level. However, for the actual data analysis or demonstration purposes, a DAG has to be filled with data and functions. One way to fill a DAG, is to think of it as a linear regression model (or more precise, as a linear structural equation model[8]). For instance, the easiest DAG in the form $X \rightarrow Y$ can be analyzed as the linear regression $Y \sim X + \epsilon$. This assumes, that Y is an additive linear combination of other variables. In this thesis, we will analyze all DAGs as linear models, keeping in mind, that other types of models (e.g., non-linear relationships, interactions) are possible. A special role in these linear models has the error term ϵ.

If we knew the true causal model and could measure all variables perfectly, we could exactly determine all causal effects. In reality, this is impossible. One of the main reasons for this is the presence of unobserved factors (errors), that influence our relevant variables in the model. This could also be factors like random measurement error or biological variability. Furthermore, since we can always only investigate causal effects in a sample of the population, our research will only result in an estimate of the true causal effect we seek to determine (the estimand).

Just like in any statistical analysis, we aim to obtain unbiased and precise estimates. Unbiasedness means that, on average, our estimate will correspond to the true value of the estimand. Precision means, that the estimate will have a small variance, or in other words,

[7] In other words, if two variables are connected they might or might not have a causal relation. If two variables are not connected we assume that they definitely have no causal relation. This is a strong assumption in many scenarios, but when reasoned properly the foundation of causal inference.

[8] A linear structural equation model (SEM) is essentially a linear regression model with additional causal assumptions (Bollen & Pearl, 2013). All DAGs (and many of the research question from the potential outcome framework of causal inference) can be rewritten as a linaer SEM, assuming the additional constraints of linearity and additive components, though SEMs can in theory also be generalized to a non-linear setting (Bollen & Pearl, 2013). The analysis of DAGs via linear SEM profs to bring insights into causal systems, both in theory (e.g., Ding & Miratrix, 2015) and in practice [e.g.,].

that repeated measurement will yield similar estimates. Random error terms add imprecision, but not bias to our model[9]. We will later encounter scenarios that introduce bias.

To illustrate the concepts of precision and bias of causal effect estimates, I will use toy data simulations in the following. These simulations generate samples of data (with n = 100) corresponding to the simulated linear model including random error terms. Each simulated sample (k = 1000) is modeled to yield a single causal effect estimate. I then visualize the distribution of simulated effect estimates. Further details of the simulation procedure can be found in Section A.2. Figure 3 demonstrates how unobserved factors (random error terms) add uncertainty to a causal effect estimate. In the absence of random error terms, each sample would give the exact true causal effect. However, in the presence of random error terms, some samples will give estimates that deviate from the true causal effect.

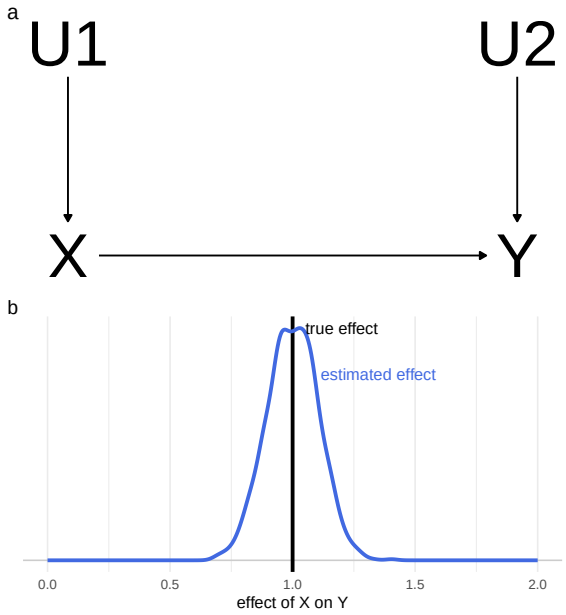

Figure 3: A simple causal path, with random error. (a) X causes Y, but both variables are influenced by other unobserved variables (random error). (b) A simulation of the model. The density plot shows the distribution of k = 1000 simulations of the model with random error terms. The random error adds uncertainty to the estimate of the causal effect, but no bias (i.e., on average, the true causal effect can be correctly estimated).

Precision in causal effect estimates is higher in simpler models. This is primarily because simpler models have fewer random error terms. This can be demonstrated by comparing a simple causal relation with a causal path (a chain). Along a causal path, information is typically lost, even if the causal effects remain unchanged. This loss of information is caused by the additional error terms of intermediate variables (see Figure 4). Chains therefore introduce uncertainty into causal effect estimates, but do not induce bias[10].

[9] At least this is an extremely common assumption. See the appendix for the mathematical notation.
[10] View the appendix for a mathematical proof.

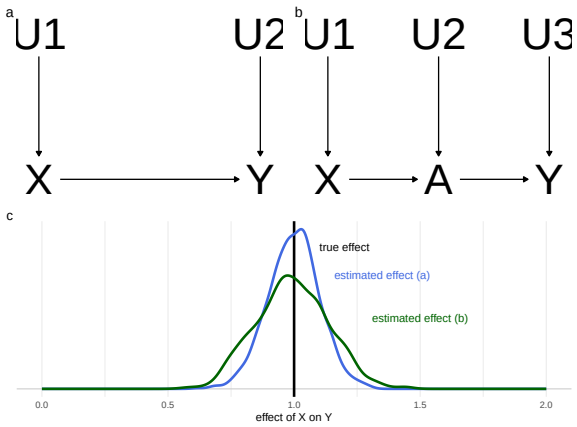

Figure 4: Random errors in a causal path. (a) X causes Y directly. Both variables are influenced by random errors. (b) X causes Y via A. All three variables are influenced by random errors. (b) A simulation of the effect of X on Y in both models. The chain introduces additional uncertainty in the effect estimate, but no bias.

For an example from sport science think of two different causal effects. First, the effect of a running intervention on mitochondrial density. Second, the effect of a running intervention on endurance performance. Even if we assume, that the effect in the second case is entirely mediated trough mitochondrial density (i.e., $intervention \rightarrow density \rightarrow performance$), the effect on endurance performance is harder to estimate. The primary reason is, that endurance performance will be influenced by additional unobserved factors, that will not influence mitochondrial density, for example motivation, pacing, or day-to-day variability.

Examining the causal model in Figure 4, we have to reconsider that the arrows drawn in a DAG are just as noteworthy as the arrows not drawn. In this example, all unobserved error terms are parent nodes, meaning that they are not influenced by any other relevant variable, including each other. This is a general assumption regarding unobserved error terms: We assume random errors to be uncorrelated. As soon as errors influence each other (directly or via other variables), we should model them explicitly[11].

2.4 Conditioning

Causal paths can be blocked by conditioning on intermediate variables. Take for example the causal path $X \rightarrow A \rightarrow Y$. Let X be the stroke volume of the heart, A the maximum oxygen uptake, and Y the endurance performance in a competition. We assume that all of the causal effect of stroke volume on endurance performance is mediated via maximum oxygen uptake. However, if we condition on maximum oxygen uptake, no relationship between stroke volume

[11] The assumption of uncorrelated error terms is also common in applied statistics outside of causal inference. If error terms are correlated this complicates the estimation of effects. We can model correlated error terms in a DAG by creating a node for an unobserved variable. Another way to investigate the consequences of correlated error terms in linear SEMs is by drawing them from a multivariate normal distribution with an appropriate covariance matrix (e.g. in Ding & Miratrix, 2015).

and endurance performance remains. Conditioning on the intermediate variable A effectively blocks the causal path between X and Y, rendering the causal effect of stroke volume on endurance performance non-existing.

Several ways to condition on variables exist[12]. An experimental approach is to stratify the sample by the variable. For instance, if we would only investigate athletes with a similar maximum oxygen uptake, we would anticipate that the relationship between stroke volume and endurance performance would deminish. A modeling approach of conditioning on a variable is to include it in the statistical model. For example, modeling $Y \sim A + X + \epsilon$ would effectively block the causal effect of X on Y (see Figure 5)[13].

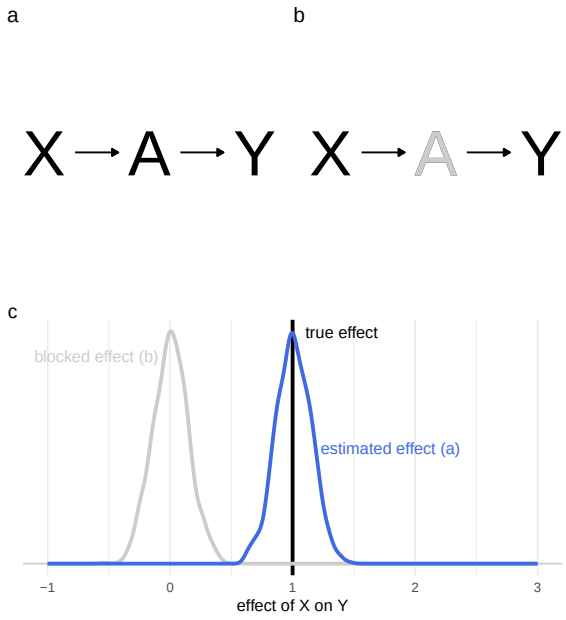

Figure 5: A causal path blocked by conditioning. (a) X causes Y via A. (b) The causal path is blocked, because the analysis conditions on A. As all affects of X on Y trail through A, no causal effect remains. (c) A simulation of the effect of X on Y in both models. Blocking removes the true causal effect entirely.

One of the main goals of causal inference using graph-based methods is identification — to identify which variables should be conditioned on. This process is crucial to provide unbiased and accurate effect estimates. Depending on the structure of the model, certain variables can introduce bias if left unconditioned, while others bias the estimate if conditioned on. The following section will further elucidate these concepts by introducing confounders and colliders.

[12]The mathematical notation of conditioning is straightforward (see Appendix). The experimental ways to condition are diverse and include methods that can be used during experimental design or data analysis.
[13]Other popular ways of conditioning include matching, …

2.5 Confounders and Colliders

Confounders are variables that causally influence both the exposure and the outcome (see Figure 6 a). The confounder creates a spurious (non-causal) association between the exposure and the outcome. Conceptually, a confounder gives a set of similar information (knowledge) to both exposure and outcome. This leads to both sharing common information, irrespective of their true causal relationship. This leads to bias in the causal effect estimate.

Confounders can be controlled for by conditioning on them in the model. This removes the entire bias and preserves the true causal relationship. Let's take an example illustrated in Figure 6. We are interested in the relationship between the (average) 5000-m time trial speed and the (average) 100-m sprint speed. We assume, that being fast in an endurance task reduces the ability to sprint fast, and thus decreases the 100-m speed. Therefore, we are interested in the causal relationship between X (endurance speed) and Y (sprinting speed). Note that this is a very simplistic causal model, as we could also model the unobserved ability to sprint and ability to perform endurance tasks, as well as their potential causes.

Our model has a collider A, representing biological sex. Based on expert knowledge, we understand that sex causally influences both sprinting and endurance performance, mainly via anthropometry and physiology. Sex thus biases the causal relationship between sprinting and endurance performance. To remove this bias, the analysis must control for sex. For a discrete variable like sex is typically documented as, controlling for means in practice stratifying the analysis by it. Assuming our causal model is correct — which holds of course not true in our toy example here — controlling for sex gives us the true (unbiased) causal relationship between endurance and sprinting performance.

Colliders pose a more subtle form of bias. A collider is a variable that is causally influenced by both the exposure and the outcome (see Figure 7 a). Per se, colliders do not yield harm. But when they are conditioning on they introduce bias into a model[14]. This collider bias can be understood by the following: A collider combines knowledge from both its source, the exposure and the outcome, and thus also of their causal relationship. If this combined knowledge is being removed from a model by conditioning on the collider, then some of the actual causal relationship between exposure and outcome is also removed.

Consider the causal relationship between X as the post-lactate in a ramp test and Y as the maximum oxygen uptake in the same ramp test. Essentially, our question is if more lactate causes a different (higher or lower) maximum oxygen uptake. In our model, both lactate and VO2max influence the maximum velocity achieved in the ramp test. This appears reasonable, as individuals with a more capable glycolytic or oxidatative energy metabolism are likely to outperform their counterparts that have neither in term of the maximum velocity. The maximum velocity attained is thus the collider B. Conditioning on it will bias our model.

[14] Equally, conditioning on a descendant of a collider introduces bias (though generally not that large as when conditioning on the collider itself).

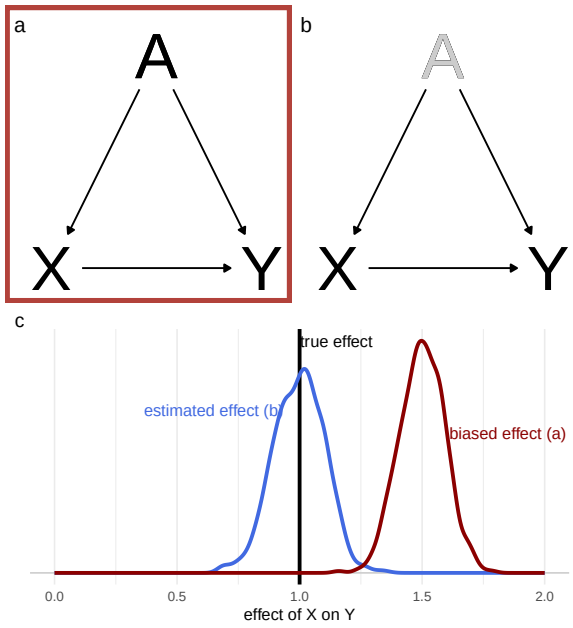

Figure 6: A graphical example of confounding. Both X and Y share a common cause A. (a) This confounder biases determining the causal effect of X on Y. (b) Conditioning on A removes the bias in the analysis. (c) A simulation of the effect of X on Y in both models.

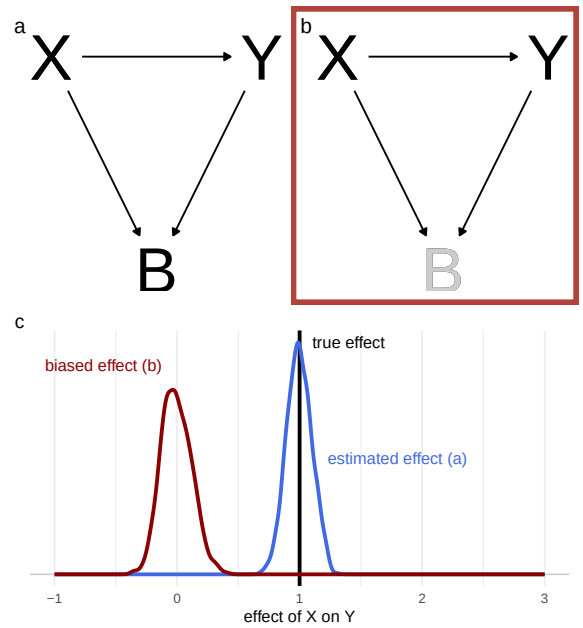

Figure 7: A graphical example of collider bias. Both X and Y directly affect the collider B. (a) As long as B is not conditioned on, the causal effect of X on Y is unbiased. (b) Conditioning on B will introduce bias in the model. (c) A simulation of the effect of X on Y in both models.

2.6 Conditioning Rules: The Backdoor Criterion

Building on the concepts of confounders and colliders, we can derive more general rules for determining the optimal conditioning set for a given causal model. The most famous of these conditioning rules is the backdoor criterion (CITE!!!). It works by first identifying all non-causal paths (backdoor paths), and second blocking all of them.

A non-causal path is any path between X and Y that starts with an arrow pointing into X. A non-causal path is open, if it has no collider or no variable conditioned on in it. It can be blocked (closed) by conditioning on a non-collider. For example, in Figure 6 $X \rightarrow Y$ is a causal path, whereas $X \leftarrow A \rightarrow Y$ is a non-causal path. The non-causal path can be blocked by conditioning on A, fulfilling the backdoor criterion, and thus providing an unbiased estimate of the causal effect of X on Y.

On the contrary, non-causal paths are blocked by default if they contain a collider. For example, in Figure 8 one non-causal path$X \leftarrow A \rightarrow B \leftarrow Y$ exists, but is is blocked by default because B is a collider. Therefore the backdoor criterion is satisfied and no conditioning is required (i.e., the minimal sufficient conditioning set is empty). If one would decide to condition on B in this scenario (for example if A were unobserved, and we chose to condition on all observed covariates), this would reopen the backdoor-path, biasing the estimate.

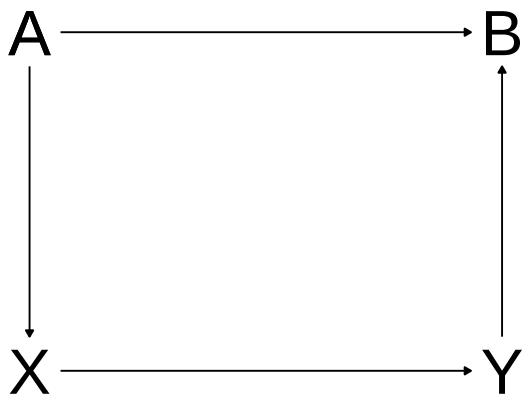

Figure 8: A graphical example of a backdoor path closed by default. The non-causal path via A and B contains a collider and is therefore closed. Conditioning on B would reopen the backdoor path.

The backdoor criterion helps to determine the variables that need to be conditioned on in graphical causal models of various complexity to obtain an unbiased estimate. These variables form the so-called minimal sufficient conditioning set. Conditioning on more variables than sufficient can increase precision in certain cases, but often brings the risk of introducing new bias or reducing precision. When certain variables in a DAG are unobserved, these can not be conditioned on. In this case it is possible that no minimal sufficient conditioning set exists that fulfills the backdoor criterion. Therefore; an unbiased estimation of the causal

effect given the assumed causal model is impossible. We will come back to the question of selecting conditioning variables in the DISCUSSION.

3 Methods

I conducted all analyses in this thesis using R version 4.4.0 (R Core Team, 2024) in the RStudio IDE version 2024.4.0.735 (Posit team, 2024). The thesis was written in Quarto version 1.3.450 (Allaire et al., 2023). The default settings and attached packages are documented in Appendix Section A.3. The DAGs in this thesis were drawn using the ggdag R package (Barrett, 2024), which is based on the software daggity (Textor et al., 2016). All source code of this project is available at GitHub.

4 Results

4.1 Example 1: Use of Different Running Shoes and Injury Risk

In running, overuse injuries occur frequently (Lopes et al., 2012). As the feet transmit all ground forces, running shoes have been the focus of many injury prevention strategies (X. Sun et al., 2020). A common belief coming from running practice is that parallel use of different shoes increases movement variability and decreases injury risk, but the scientific evidence for this is limited (Mechelen, 1992).

Malisoux et al. (2015) tested the claim that concomitant use of running shoes decreases injury risks in an observational study. Using a prospective cohort design, they followed a group of 264 runners training for a marathon and documented their anthropometrics, training characteristics, shoe use, and injury occurrence. Malisoux et al. (2015) categorized runners into multi-shoe and single-shoe users, where multi-shoe users were those who reported to have changed running shoes at least twice between training sessions over the observation period. The authors fit several Cox proportional hazard regressions to the data[15]. Using a semi-automated parameter selection, they finally arrived at a multivariate ("adjusted") model, with the coefficients indicating that multiple-shoe users had indeed a lower injury risk.

The study by Malisoux et al. (2015) was clearly causal in its aim. The title 'Can parallel use of different running shoes decrease running-related injury risk?' poses a causal question, the hypothesis is of causal nature, the authors discuss 'protective factors', speculate about potential causal mechanism of their findings, and state that multiple shoe use "could be advised to recreational runners to prevent running-related injuries" (Malisoux et al., 2015). However they acknowledge the low statistical power of their study and suggest that larger and longer observational studies, or randomized controlled trials should be conducted to confirm their findings. I will here revisit the study by Malisoux et al. (2015) from a causal inference perspective.

The inherent drawback of the observational study by Malisoux et al. (2015) is the lack of randomization. The conditions of single or multiple shoe use are not randomly assigned to a runner, but are chosen by them (though implicit, as the research goal was not communicated in advance). When treatment conditions are chosen instead of randomly assigned this can introduce bias when estimating causal effects of the treatment. Essentially, in this case the assignment to a condition is likely not independent of the (projected) outcomes. Confounders may bias the causal relationship between treatment and outcome. An indicator of this might be the group imbalances in the pre-treatment variables seen in Table 1 of Malisoux et al. (2015). Multiple shoe users were on average older and had more regular training and racing in the year before the study. While some baseline imbalances are natural even in the case of randomization, this pattern indicates that some variables potentially had an influence in the

[15] Cox regression is a popular regression tool for survival analysis. In short, the survival rate over time depending on one or more covariates is modeled. In this case, being non-injured over a given amount of training volume was compared between the group of multiple shoe and single shoe users (Malisoux et al., 2015).

choice of using multiple running shoes, and these variables are confounders if they also have an effect on injury risk as the outcome variable.

A useful way to check for baseline imbalances would be investigating propensity score overlap. The propensity score is the probability to be assigned to one of the treatment conditions based conditional on the observed covariates, and is usually estimated with a logistic regression (Rosenbaum & Rubin, 1983). Comparison of propensity score distributions between the treatment group can diagnose covariate imbalances in a multivariate setting. In a completely randomized design, the propensity score distribution of treatment groups should be fairly similar. Differences in the propensity score distributions indicate a dependency between treatment assignment and covariates, potentially biasing the causal effect estimate. Assuming that the variables leading to this bias are all observed, we can use methods to correct for covariate imbalances[16]. In this example, if we assume that certain variables influence the outcome of injury and the treatment variable of shoe use, we may compare only those single and multiple shoes users that share similar values for covariates. Methods for covariate balancing involve matching, reweighing, and subclassification procedures (Stuart, 2010), these are discussed later (LINK).

A potential DAG is

Malisoux et al. (2015) are aware of the potential impact of confounders. This is why they do not directly interpret bivariate analyses of any variable with injury risk, but provide an "adjusted" multivariate model. This model is used to estimate the effect of parallel running shoe usage on injury risk, while controlling for confounders. However, not only the coefficient of running shoe use, but also the other coefficients of the final model are interpreted in a causal way (e.g., the participation in sports other than running). The direct causal interpretation of multiple coefficients from multivariate models has been called the "Table 2 fallacy" and it is generally regarded as bad statistical practice (Westreich & Greenland, 2013). Moreover when no preregistration was done, this practice may invite researchers to present post-hoc hypothesis as a priori (Kerr, 1998). Even if only the primary causal effect of interest is considered, the final model by Malisoux et al. (2015) is likely to provide a biased estimate. The "adjusted" model is chosen by first performing a bivariate screening of all available variables and then an automatic selection procedure on a subset of these (with two variables manually included). In the methodological literature, bivariate screening is generally advised against, while automated variable selection is highly debatable (G.-W. Sun et al., 1996). Current best advice is to use background knowledge when selecting appropriate variables (Heinze et al., 2018) and if this is not sufficient at least use regularization methods (e.g., Fan & Li, 2002).

[16] In other words, we assume that we can create independence between group assignment and outcome conditional on the observed covariates (sometimes called the "ignorability assumption". See the APPENDIX for the mathematical notation. Another assumption is that of common support. Despite covariate imbalances, some overlap between the covariate distributions of the treatment condition needs to exist. If this is not given (e.g., the propensity score distributions between groups not only differ, but are completely separated), we cannot reasonably balance the sample and estimate the causal effect because it would heavily rely on extrapolation. For the given example this means, when multiple and single shoe users are almost totally different in their characteristics, we cannot adequately adjust the data to identify the true causal effect of shoe use.

Another potential source of bias is hidden in the definition of the treatment variable in Malisoux et al. (2015). Multiple shoe users are defined by a minimum number of two shoe changes over the observation period. There may be a direct dependence between this grouping criterion and the outcome variable (non-injuredness over the observation period): On the one hand, athletes who receive an injury subsequently drop out of the study and thus have less time to accumulate shoe changes for being categorized as multiple-shoe users. Potentially, athletes would be considered multiple-shoe users if they had trained for a longer time instead of receiving an injury. On the other hand, people who dropped out of the study were, after a check, considered as non-injured. These athletes had less time to accumulate shoe changes and may have been more likely to be characterized as single shoe users. In both ways a non-causal relationship between the particular definition of shoe use and injury may exist. A way to check this source of bias and aid the causal interpretation of the study would be to give information on the observation duration, possible in form of a survival curve (Kaplan & Meier, 1958). Directly modeling drop-out or testing the robustness of the model by using the momentarily instead of the retrospective group assignment may be a statistical way to deal with these potential biases.

Taken together, from a causal inference viewpoint the results by Malisoux et al. (2015) should be questioned. The study could benefit from the discussion of an underlying causal model (e.g., in form of a DAG) and the use statistical methods to deal with non-randomized group assignment (e.g., propensity score-based weighting). At a minimum, the definition of multiple shoe use should be rechecked and survival curves should be included in the analysis. Finally, the rather small sample size (low absolute number of injuries occurred) will lead to imprecise estimates even if unbiasedness can be assumed. Therefore I agree with Malisoux et al. (2015), that either RCTs or larger observational studies should be conducted, if the research question is of enough relevance. I would just add that appropriate causal inference methods could help in all of these cases.

4.2 Example 2: Nutrient intake and marathon mountain performance

In ultra-endurance races, appropriate nutrient intake is essential both for performance and health reasons (Costa et al., 2019; Nikolaidis et al., 2018; Williamson, 2016). Athletes are encouraged to maintain proper fluid and carbohydrate intake during races (Thomas et al., 2016), which should in theory benefit performance. The influence of nutritional intake on performance in actual ultra-endurance races has however rarely been investigated.

Kruseman et al. (2005) documented nutrient intake during a ultra-marathon mountain race in 46 runners. Additionally, they measured anthropometrics before and after the race and registered the race performance. The observational cohort study had the primary aim of providing descriptions of actual nutrition strategies during an ultra-endurance competition and compare them with published guidelines. The secondary aim by Kruseman et al. (2005) was to study “the association between nutrient intake and performance”. To test their secondary

aim, the authors split the group into performance tertiles and tested bivariate relationships to anthropometric, running experience, and nutrient intake variables using analysis of variance and χ^2-tests. They then took the statistically significant variables from the bivariate analyses and ran a multivariat regression model with backward stepwise selection. Kruseman et al. (2005) showed that most athletes failed to meet the nutrient recommendations for the race, but there was no significant association with performance.

The secondary aim by Kruseman et al. (2005) is causal; it is build on the hypothesis that inadequate nutrient intake hinders performance. Yet they acknowledge, that "[b]eing a cross-sectional, observational study, no causal relationship can be drawn between [nutrient] intake and performance", which seems counterintuitive to the research goal. Kruseman et al. (2005) found significant associations between nutrient intake and performance in their bivariate analysis, but not in the multivariate model, that adjusted (among others) for previous race experience. The authors take a quite critical stance towards their own results and advise for further experimental studies[17]. In the following section I will revisit the study by Kruseman et al. (2005) from a causal inference perspective.

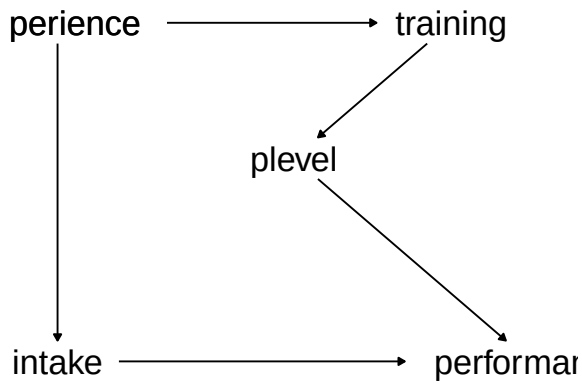

Figure 9: A potential graphical model for nutrition intake in an ultra-marathon. plevel stands for performance level (the physiological capability to perform the endurance task).

A potential DAG of nutrient intake and ultra-endurance performance is shown in Figure 9. Nutrient intake has a direct causal effect on performance, as low carbohydrate availability and dehydration induce fatigue and thus reduce performance. Nutrient intake is mainly determined by an athlete's experience (e.g., knowledge about nutritional strategies, prior race experience). Experience determines the training of an athlete, both qualitative (e.g., experienced athletes may know better which training is suited for them) and quantitative (e.g., athletes with more running experience have had more time in their life to accumulate running training). Training in term influences the performance level, the physiological ability to perform the given endurance task pre-start. This performance level together with the nutrient

[17]The way the authors critically discuss the causility of their findings is bracing for sport science. Given the correctly identified limitations of the data analysis, I wonder why the authors chose to perform such an analysis in first place. It would have been interesting to see how critical the authors would have been if their multivariate model had indeed included the predicted significant effect of nutrient intake on performance.

intake during the race determines the final race performance.

Given that the DAG in Figure 9 is an appropriate representation of the causal model underlying Kruseman et al. (2005), the effect of nutrient intake on performance is biased by an open backdoor path. This backdoor path could be closed by conditioning on any intermediate variable. As experience is the only of the three intermediate variables of Figure 9 measured by Kruseman et al. (2005), it seams reasonable to condition the analysis on it[18]. Kruseman et al. (2005) recognize that experience is a confounder of the causal relationship between nutrient intake and performance, as they write: "Because experienced runners are well trained, fitter, and know their personal needs better during such a race, it is impossible to precisely separate the associations we found, especially in a cross-sectional design." But given the DAG in Figure 9, conditioning on experience in the statistical model does indeed allow to "separate" the causal effect of nutrient intake on performance.

Kruseman et al. (2005) close the backdoor path by conditioning on experience in their multivariate model (though rather inadvertently as their variable selection procedure is not determined by background knowledge but by automated rules). The resulting conditional effect of nutrient intake on performance is non-significant[19]. Based on the DAG in Figure 9, this should be a less biased estimate than the bivariate associations between nutrient intake and performance, that Kruseman et al. (2005) also report, but flag as potentially spurious. Interestingly, Kruseman et al. (2005) disregards both estimates (unadjusted and adjusted) as biases, stating that neither "does allow us to conclude any definitive causal relationships". This is of course true for any effect estimate, particularly in the context of observational studies. But it should be aim of any researcher to reduce bias in causal effect estimates, or otherwise the analysis would be of no value at all.

A modified DAG can elucidate some of the skepticism by Kruseman et al. (2005) regarding their effect estimate adjusted for experience. If we assume in Figure 9 a further causal relationships from training to nutrient intake, this would create another background path that cannot be closed by only conditioning on experience. Kruseman et al. (2005) hint to such a relation by writing that "in addition, training increases the benefits of adequate nutrition, and favors the accumulation of muscle glycogen after exercise". This suggests, that training acts as a moderator of the relationship between nutrient intake and performance. As previous training was not documented in the study, it cannot be controlled for in the model, thus this open backdoor path would remain open and the causal effect would be biased by a confounder. Again — though not in the language of causal inference — this is recognized by Kruseman et al. (2005), as they write: "It would have been interesting to record the training

[18] Conditioning on experience instead of conditioning on training has additional benefits if me modify the model by allowing direct effect of experience on performance, that are not mediated via training. Two possible examples would include psychological readiness and pacing strategy. Both influence performance and are possible more caused by experience than by training.

[19] Technically we do not know from Kruseman et al. (2005) not know if the effect estimate is non-significant. We just know that the variables related to nutrient intake were removed from a model by backward stepwise selection. As the selection criteria was probably statistical significance, it is likely that nutrient intake variables would also have been non-significant in a separate model only conditioned on experience.

level of the participants and study the potential confounding effect of training level on nutritional intake during a race. However, race experience seems an adequate indirect marker of training, as is body fat mass." They are right that experience is an adequate marker of training given the DAG in Figure 9, but assuming a direct cause of training on nutrient intake this is false. Presenting and discussing potential causal models would have provided a reasonable benefit to the analysis of Kruseman et al. (2005).

Even without a DAG Kruseman et al. (2005) discuss their results in light of potential causal relationships between variables. But their data analysis is blind to this causal knowledge, as it only uses automated procedures for variable selection in the models. Both the bivariate screening and the backward stepwise selection are methods that should in general not be used for causal inference (G.-W. Sun et al., 1996). But even with an model selection informed by background knowledge, the study sample size of 46 athletes with great heterogeneity in their covariates may be too small to get precise effect estimates. A different experimental approach would to include the performance over different sections of the ultra-marathon race in the analysis. In general we can assume, that in-race nutrient intake becomes more important later during the race, as it should not have any influence on performance in for example the first hour of racing. Investigating the causal effect of nutrient intake by dissecting in-race performance may provide both a better causal effect estimate and an additional plausibility check for the model.

5 Discussion

5.1 General Applications of Causal Inference in Sport Science

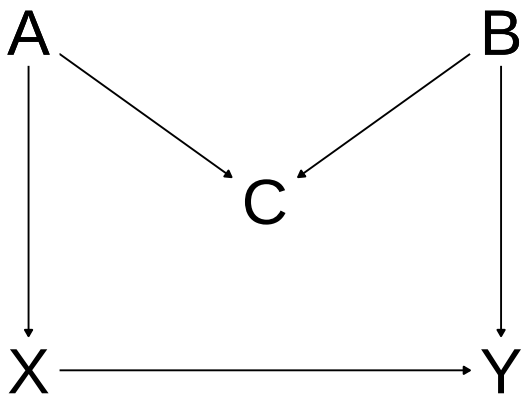

Figure 10: A graphical example of M bias. C is caused by both A and B, which also effect X and Y, respectively. In this scenario, a non-causal path exists, but it is closed, because C is a collider. When C is conditioned on, this would open the backdoor path, because the conditiong creates a spurious relationship between X and Y, so that both act together as a confounder.

5.2 Applicability of Special Causal Inference Methods in Sports

Apart from general principles of causal thinking and modeling based on graphical representations, a set of special causal inference methods has gained popularity in the past decades. Based on the potential outcome framework (Rubin, 1974), these became standard tools in the analysis of observational data especially in the field of economics (Athey & Imbens, 2017), but also beyond. Angrist & Krueger (1999) called these set of causal inference tools "identification methods", as they help to identify causal effect estimates in certain common situations. The identification methods are well-researched analysis tools, that may also prove helpful in many applications in sport science. I will here introduce five common identification methods and discuss their potential application in sport science[20]. Table 1 provides a summary of the five methods.

[20] I chose the set of five methods based on Angrist & Krueger (1999) while making some modifications. I replaced the general "conditioning in a regression model" strategy, which has been discussed earlier, with covariate balancing methods, and added the newer method of synthetic control (both to some extent inspired by Athey & Imbens, 2017; Cunningham, 2021).

Table 1: A summary of causal identification methods and their application to sport science.

Method	Basic Idea	Reference	Applications in Sport Science
Covariate Balancing	Creating groups balanced on observed covariates when group assignment was not random	Stuart (2010)	Genetic profiling, injury research, team sport analytics
Instrumental Variables	Control for unobserved confounders by using a variable that only relates to the outcome via the treatment.	Greenland (2000)	Non-compliance and measurement errors in sport interventions, talent development
Regression Discontinuity	Finding a treatment that is assigned based on a certain threshold, to compare individuals slightly above and below this threshold	Imbens & Lemieux (2008)	Effects of winning and relegation, draft systems, squad nominations
Difference-in-Difference	Observing a quasi-experimental treatment and control group over time to estimate treatment effects	Lechner (2011)	Rule changes, technological developments
Synthetic Control	Comparing a single time-series to a synthetic control time-series based on imperfect control groups	Abadie (2021)	Coach changes, talent development programs

5.2.1 Covariate Balancing

A common approach to causal inference of observational data is to mimic the characteristics of a RCT. In a RCT, treatment assignment is random, and thus treatment groups only differ in their covariates by chance. Conversely, in observational studies covariates may influence treatment assignment. For example in the study of multiple running shoe use and injury risk from the previous chapter [LINK] (Malisoux et al., 2015), runners may decide if they use different shoes based on weekly running volume. If weekly running volume also influences the outcome parameter of injury risk, this is an classical example of confounder bias [LINK DAG]. To mimic a RCT of multiple running shoe use, we could decide to only compare individuals with similar running volume (and other covariates). This is the basic idea of covariate balancing.

Covariate balancing can broadly be defined in three categories: subclassification, matching, and reweighting (Stuart, 2010). Subclassification groups individuals with similar covariates into subclasses, then compares different treatments only within the subclasses, and finally calculates a (weighted) average of these comparisons (Cochran, 1968). Matching aims to find individuals with equal or similar covariates and compare only them, often on a 1:1 basis, before pooling all comparisons (Rubin, 1973). This often involves discarding data for which no (sufficient) matches could be found. Reweighting keeps all observation, but gives them new weights based on how representative they are for their group.

Regardless of the method used, covariate balancing is typically a pre-analysis routine, i.e., it happens in a step before the actual causal data analysis and without including information on the outcome values. In some sense, it aims to solve the same problem of observed confounders that simple conditioning in a regression does address. Current advise is to use covariate balancing not instead of regression adjustments, but complementary, for example in the so called “doubly-robust” methods (Bang & Robins, 2005). A benefit of covariate balancing over regression adjustments is that it eases the checking of overlap in covariate distributions. If this overlap is not given, linear adjustments tend to perform bad as they have to rely on extrapolation, but typical modelling workflows of linear regression do not involve simple checks of this.

A crucial question in covariate balancing is what defines “closeness” or similarity of covariates. Exact equality on all covariates will only work for large samples with few discrete covariates. In most cases researchers have to compute distance measures. One of the most common measures is the propensity score, the conditional probability that an individual was assigned to the treatment group given its covariates (Dehejia & Wahba, 2002; Rosenbaum & Rubin,

1983). The propensity score thus reduces the multidimensional covariates to a single value. This value can be used to form subclasses, match units, or weight observations. The exact choice of a balancing method and a distance measure should be context-specific and the scientific debate which procedure works best is vital[21]. Stuart (2010) provides general recommendations regarding the choice of procedures. In general, it can be helpful to compare different methods (and method parameters) in a given data set.

The first researchers have begin to adopt covariate balancing methods into sport science. In the field of injury rehabilitation, propensity score matching was used to find adequate control groups for athletes undergoing recovery (Farinelli et al., 2023; Fenn et al., 2023; Jimenez et al., 2022; Owens et al., 2022). Others compared fitness levels in soccer players over different decades (Gonaus et al., 2023) and estimated the effect of playing venue of match outcome (Kneafsey & Müller, 2018) using different matching methods. Nakahara et al. (2023) used propensity-score based subclassification to evaluate pitching strategy in baseball. In an innovative article, Gibbs et al. (2022) used matching to investigate the effect of timeouts on stopping point runs in basketball games.

In general, covariate balancing can help in any situation in sport science, where we have a limited set of intervention, a set of observed pre-treatment variable, and rather large sample sizes. While covariate balancing methods can reduce bias, matching effectively discards observation units, and matching estimators are unstable in small sample sizes. Therefore, these methods are rather suited when there is at least a rather large control group of individuals. This could, for example, be comparing physiological markers of injured vs. uninjured athletes, or comparing genetic profiles of elite athletes against a non-elite population. An additional field of application could be non-randomized studies of health benefits from recreational sport participation. Covariate balancing can also help in understanding causal mechanisms of team sport performance (e.g., Gibbs et al., 2022).

5.2.2 Instrumental Variables

Confounders between received treatment and outcome are often not observed, or even known. In situation it is not directly possible to estimate an unbiased causal treatment effect. A way to reduce bias is by using an instrumental variable. Instrumental variables are variables, that

[21] Just as a short glimpse into the debate: Frölich (2004) argues that weighting is always worse than matching, a finding that is challenged by Busso et al. (2014). Iacus et al. (2011) heavily criticize the widely used propensity-score matching (Dehejia & Wahba, 2002). They instead propose their own method of coarsed exact matching (Iacus et al., 2011, 2012), which has in turn received opposition by Black et al. (2020).

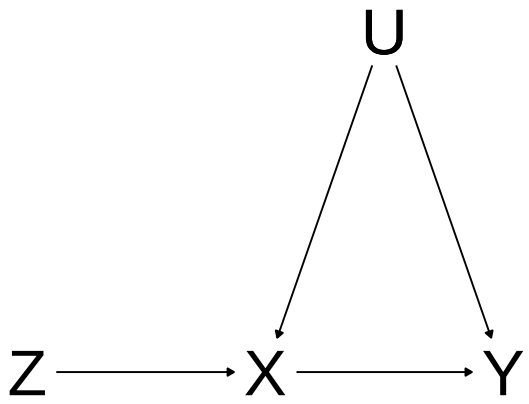

Figure 11: A graphical example of an instrumental variable. The relationship between exposure X and outcome Y is confounded by a set of unobserved variables U. The instrumental variable Z, which is unconfounded and only affects Y via X, can be used to provide an unbiased estimate of the causal effect of X on Y.

cause the outcome only mediated by the treatment. For a graphical representation see Figure 11. A set of unobserved variables U influences both treatment X and outcome Y, thus their causal relationship ist biased, as we cannot control for U because it is unobserved. If we instead use the instrument Z, which only causes X directly and Y indirectly via X, we can isolate a part of the effect of X on Y that is not influenced by U.

A classic example of an instrument is random treatment assignment (Greenland, 2000). People may actively decide if they want to receive a treatment, and these decisions may be driven by unobserved confounders that also cause the outcome variable. The assignment to a treatment does not directly influence the outcome, but only receiving a treatment does. As people assigned to the treatment group are much more likely to follow the assignment and thus receiving treatment, treatment assignment works as an instrument. Using an appropriate procedure (typically a two-stage least square estimator), researchers can estimate the causal effect of receiving a treatment on the outcome[22].

Instrumental variables can be used to adjust for unobserved confounders and is helpful in problems of measurement error and non-compliance. As an example from sport science, Ruseski

[22] Depending on whether we assume heterogeneity in the treatment effect, the causal effect estimated is somewhat limited in its definition. Strictly speaking we only estimate the causal effect of treatment in those that adhere to treatment assignment. This is often called the locale average treatment effect of compliers. If we are interested in the mechanistic causes of receiving a treatment this should be what interests us. If the goal is to evaluate the causes of implementing a treatment on a population level (e.g., for policy research) it is more appropriate to include non-compliance in the estimation, and therefore not use instrumental variables.

et al. (2014) investigated the causal effect of sport participation on happiness, a relationship that is likely confounded by unobserved variables (e.g., biography and socio-economic background). They used physical distance to the nearest sporting facility and personal belief in the benefits of exercising as instrumental variables and found that there was indeed a positive causal effect of sport participation on happiness. Edouard et al. (2021) suggested to use instrumental variables for the analysis of sport injury prevention treatments. But Shrier et al. (2020) demonstrated that they made several mistakes in both their theoretical presentation and the example data analysis. This does not invalidate the potential use of instrumental variables in sport science, but highlights the caution that has to be taken when implementing new approaches.

The search for good instrumental variables is a challenging one. Aside from the assumptions of the DAG structure in Figure 11, instruments should be strongly related to the treatment variable. If they only moderately influence the treatment, they are called “weak instruments” (Bound et al., 1995). Weak instruments are often unsuccessful in removing bias, even in large samples. Good instruments often involve an element of randomness. As an example of sport science, we may be interested in the effect of being part of youth national squad on adult sport success. Both are very likely confounded by a variety of unobserved factors (e.g., social, psychological, and biological). An example for an instrument is in this case the month of birth. The month of birth does not share any unobserved confounders with the other variables, as it can be seen as essentially random. But it effects being part of a youth national squad (as the norms are defined by birth years and later born are less matured and thus less likely to be part of the squad). It does however have no direct influence on later adult success. Therefore the DAG in Figure 11 holds, though we have to test if birth month is not a too weak instrument. This example demonstrates the potential use of instrumental variable approaches in sport science.

5.2.3 Regression Discontinuity

Regression discontinuity is one of the most widespread dedicated methods of causal inferences. First used in psychology by Thistlethwaite & Campbell (1960), it took until the late 1990s to finally gain popularity (Cook, 2008), to becoming one of the most common designs in causal analysis, particularly in economics (Lee & Lemieux, 2010). The basic idea of regression discontinuity is that there is a running variable that decides treatment based on a fixed cut point. Individuals slightly above and below the cut point will likely be similar in terms of

observed and unobserved covariates, but they will differ in the treatment they receive. Exceeding the threshold will either always lead to the application of treatment (sharp regression discontinuity design) or disproportionately increase the probability of receiving a treatment (fuzzy regression design) (Imbens & Lemieux, 2008). The discontinuity in the outcome at the cut point allows to estimate a (local) causal effect of the treatment.

A key assumption of regression discontinuity is the continuity. It means that we expect no discontinuities in outcome if no treatment had been applied (if the variable exceeding a cut point did not lead to any consequences). Regression discontinuity design are usually analysed by running regressions of the running variable on the outcome for each side of the cut point (Imbens & Lemieux, 2008). The difference in expected outcomes at the cut point of the running variable corresponds to the local causal effect of the treatment. To account for non-linear relationships researchers often use polynomials, though this procedure can be misleading (Gelman & Vehtari, 2021). Individuals who are aware of the relevance of the threshold may game the treatment assignment by manipulating their running variable value[23]. Therefore analysts should check the density and characteristics of individuals near the threshold (Barreca et al., 2016; McCrary, 2008).

As an example from sports take the system of promotion and relegation in sport leagues. The causal effect of promotion and relegation on performance and financial is difficult to estimate, as teams being relegated are likely less strong and financially equipped than teams holding their league. But when only comparing teams that were slightly above or below the cut down for relegation, we can compare teams that probably share a similar background. Speer (2023) used this idea to estimate the financial effect of relegation and promotion in sport leagues using a regression discontinuity design.

Other sport application of regression discontinuity come mainly from sport economics. Keefer (2016) and Branson et al. (2019) investigated the effects of the draft system in American basketball. Hon & Parinduri (2016) researched the causes of introducing the three-point rule in the German soccer league. Engist et al. (2021) focused on the effect of the seeding system on team performance in soccer tournaments. In an application more related to health care, Fredslund & Leppin (2019) estimated the influence of a holiday break on fitness routines in the

[23]This would bias our causal estimate, as treatment assignment would not be essentially random around the threshold anymore. A famous example is Barreca et al. (2011), who showed that an unusual high number of babies slightly fell below a body weight threshold that made them eligible for intense medical care. This suggests that doctors and nurses may have deliberately manipulated the children weight measurement if they believed intense care to be helpful. Not correcting for this bias will result in unreliable results of the regression discontinuity analysis.

general population. In one of the most popular applications of regression discontinuity designs in sport (though published in economic journal), Berger & Pope (2011) found that being slightly behind in a game at half-time disproportionately increases the chances of winning a basketball game. They speculated that being behind by a slight margin increases motivation and thus causes a performance benefit. But Klein Teeselink et al. (2023) showed that the findings by Berger & Pope (2011) did not hold in a larger sample of more seasons and different sports.

Regression discontinuity is suited for any sport context that involves slight differences in an indicator causing large potentially effects in outcomes. The aforementioned concepts of team relegation, and in general winning a game are thus potential use cases for regression discontinuity analyses in sport. Further applications could be squad nomination based on fixed performance thresholds, tournament seeding based on ranking systems, or the effect on health interventions implemented based on biological measures (e.g., body mass index thresholds). As the field of sports and exercise science is rich of indicator variables that cause treatment by using rather arbitrary break points, a promising potential for using regression discontinuity designs exists.

5.2.4 Difference-in-difference

Difference-in-difference is the oldest and most common quasi-experimental research design to estimate causal effects from observational data. Its first use can be traced down to John Snow's investigation of the London cholera pandemic in the mid 1850s (Coleman, 2019; Snow, 1955). In the past decades it has evolved into the most popular design in economics and social-science, with one of the most influential application being Card & Krueger (1994) investigating the influence of a minimum wage rise on the labor market. The basic idea of difference-in-difference design is — in the absence of a RCT — to use a natural experiment. Two groups of individuals are observed over a time period, in which one group receives an intervention and the other does not. Both groups may be influenced by unobserved time-varying factors, but if these are constant, the second group can act as the control group for the treatment. Finally, the between-group difference between the two within-group differences is an unbiased estimate of the causal treatment effect.

The key assumption of difference-in-difference designs is the parallel time trend. Without any intervention, we expect both groups to develop in a similar way. To ensure this parallel time trend the choice of the control group is crucial. This is why difference-in-difference is sometimes combined with matching [SEE PREV] or synthetic control groups [SEE NEXT]. The

treatment and the control group do not necessarily have to be observed over the same time period (Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021), and the treatment can also be continuous instead of binary (Callaway et al., 2024). Because difference-in-differences works with time series data, much care has to be taken into calculating appropriate measures of certainty of estimates (Bertrand et al., 2004).

Difference-in-difference designs can be found in sport science, but are rarely termed that way and often not systematically analysed from a causal viewpoint. Essentially most analysis of covariance and time*group interaction analyses can be understood as some form of difference-in-difference. Research in sport science that explicitly use observational data to estimate causal effects using the well-researched difference-in-differences methods are rare, possible because the spread of this design has been mainly limited to economics and social science, but not medical science. Consequently the few articles published stem from the field of sport economics, where we can expect that authors were inspired by research from their mother discipline (Böheim et al., 2022; Budzinski & Kunz-Kaltenhäuser, 2020; Weimar & Breuer, 2022). However the difference-in-difference design has potential for sport science in far more cases when randomized experiments are not feasible, but quasi-experimental research of observational data is possible. This could for example be the effect of rule changes or abrupt technological advancements in sports.

5.2.5 Synthetic Control

Synthetic control is arguably the most novel and thus most developing special causal inference method of the past years (Athey & Imbens, 2017). It was first used by Abadie & Gardeazabal (2003), and first systematically presented in Abadie et al. (2010). In synthetic control studies, researchers observe a single time-series of a group-level intervention, for which no adequate control group exists. Therefore a larger set of non-fitting (i.e., differing in covariates) control groups is combines to create a single "synthetic" control group. This combining usually works by weighting covariates of the different groups from the control pool in such a way, that they match the covariates of the treatment group. For the pre-treatment period, the intervention and the synthetic control group should have similar trends in outcomes, and any differences in time trends after the treatment can be causally attributed to the treatment.

Despite its promising and innovative nature, synthetic control methods have not yet been used in sport science. The main reason may be that the research field is quite new even in its home discipline of economics and political science, so that it may take a few more years

until first researchers will begin to adopt these methods to sport science. That said, sport science offers many instances, where single-unit time series data could be compares against synthetic control groups. An example is the influence of head coach changes on performance in team sports. Another example is the influence of talent development programs in certain countries on later athlete success.

5.3 Challenges and Limitations

5.3.1 Need for Theoretical Models

5.3.2 Complex Systems

5.3.3 Small Samples

5.3.4 Data Quality

5.4 Perspectives and Further Possibilities

5.4.1 Modeling Missing Data and Measurement Error

5.4.2 Heterogenous Effects

5.4.3 Longitudinal Data

5.4.4 Understanding Big Data

5.4.5 Communicating Causality

5.5 Causal Modeling Workflows in Sport Science Practice

6 Conclusion

References

Abadie, A. (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, *59*(2), 391–425. https://doi.org/10.1257/jel.20191450

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*. https://doi.org/10.1198/jasa.2009.ap08746

Abadie, A., & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, *93*(1), 113–132. https://doi.org/10.1257/000282803321455188

Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y., & Dervieux, C. (2023). *Quarto*. https://doi.org/10.5281/zenodo.5960048

Angrist, J. D., & Krueger, A. B. (1999). *Empirical strategies in labor economics* (O. C. Ashenfelter & D. Card, Eds.; Vol. 3, pp. 1277–1366). Elsevier. https://doi.org/10.1016/S1573-4463(99)03004-7

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton Univers. Press. https://doi.org/10.1515/9781400829828

Athey, S., & Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32. https://doi.org/10.1257/jep.31.2.3

Balshaw, T. G., Massey, G. J., Maden-Wilkinson, T. M., & Folland, J. P. (2017). Muscle size and strength: debunking the "completely separate phenomena" suggestion. *European Journal of Applied Physiology*, *117*(6), 1275–1276. https://doi.org/10.1007/s00421-017-3616-y

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Barreca, A. I., Guldi, M., Lindo, J. M., & Waddell, G. R. (2011). Saving babies? Revisiting the effect of very low birth weight classification*. *The Quarterly Journal of Economics*, *126*(4), 2117–2123. https://doi.org/10.1093/qje/qjr042

Barreca, A. I., Lindo, J. M., & Waddell, G. R. (2016). Heaping-Induced Bias in Regression-Discontinuity Designs. *Economic Inquiry*, *54*(1), 268–293. https://doi.org/10.1111/ecin.12225

Barrett, M. (2024). *ggdag: Analyze and create elegant directed acyclic graphs*. https://github.com/r-causal/ggdag

Berger, J., & Pope, D. (2011). Can losing lead to winning? *Management Science*, *57*(5), 817–827. https://www.jstor.org/stable/25835742

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?*. *The Quarterly Journal of Economics*, *119*(1), 249–275. https://doi.org/10.1162/003355304772839588

Bishop, D. (2008). An Applied Research Model for the Sport Sciences. *Sports Medicine*, *38*(3), 253–263. https://doi.org/10.2165/00007256-200838030-00005

Black, B. S., Lalkiya, P., & Lerner, J. Y. (2020). The Trouble with Coarsened Exact Matching. *Northwestern Law & Econ Research Paper Forthcoming*. https://doi.org/10.2139/ssrn.3694749

Böheim, R., Lackner, M., & Wagner, W. (2022). Raising the Bar: Causal Evidence on Gender Differences in Risk-Taking From a Natural Experiment. *Journal of Sports Economics*, *23*(4), 460–478. https://doi.org/10.1177/15270025211059533

Bollen, K. A., & Pearl, J. (2013). *Eight Myths About Causality and Structural Equation Models* (S. L. Morgan, Ed.; pp. 301–328). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_15

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443–450. https://doi.org/10.2307/2291055

Branson, Z., Rischard, M., Bornn, L., & Miratrix, L. W. (2019). A nonparametric bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*, *202*, 14–30. https://doi.org/10.1016/j.jspi.2019.01.003

Buckner, S. L., Dankel, S. J., Mattocks, K. T., Jessee, M. B., Grant Mouser, J., & Loenneke, J. P. (2017). Muscle size and strength: another study not designed to answer the question. *European Journal of Applied Physiology*, *117*(6), 1273–1274. https://doi.org/10.1007/s00421-017-3615-z

Budzinski, O., & Kunz-Kaltenhäuser, P. (2020). *Promoting or Restricting Competition? – The 50plus1-Rule in German Football* (No. 141). Ilmenau Economics Discussion Papers. https://doi.org/10.2139/ssrn.3623779

Bullock, G. S., Ward, P., Hughes, T., Thigpen, C. A., Cook, C. E., & Shanley, E. (2023). Using Randomized Controlled Trials in the Sports Medicine and Performance Environment: Is It Time to Reconsider and Think Outside the Methodological Box? *Journal of Orthopaedic & Sports Physical Therapy*. https://doi.org/10.2519/jospt.2023.11824

Busso, M., DiNardo, J., & McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *The Review of Economics and*

Statistics, *96*(5), 885–897. https://doi.org/10.1162/REST_a_00431

Callaway, B., Goodman-Bacon, A., & Sant'Anna, P. H. C. (2024). *Difference-in-differences with a continuous treatment*. National Bureau of Economic Research. https://doi.org/10.3386/w32117

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Card, D., & Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review*, *84*(4), 772–793. https://ideas.repec.org//a/aea/aecrev/v84y1994i4p772-93.html

Carlin, J. B., & Moreno-Betancur, M. (2023). *On the uses and abuses of regression models: A call for reform of statistical practice and teaching*. arXiv. https://doi.org/10.48550/arXiv.2309.06668

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*(2), 295–313. https://doi.org/10.2307/2528036

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs* (2nd ed). John Wiley & Sons.

Coleman, T. (2019). *Causality in the Time of Cholera: John Snow As a Prototype for Causal Inference*. SRRN. https://doi.org/10.2139/ssrn.3262234

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, *142*(2), 636–654. https://doi.org/10.1016/j.jeconom.2007.05.002

Costa, R. J. S., Hoffman, M. D., & Stellingwerff, T. (2019). Considerations for ultra-endurance activities: Part 1- nutrition. *Research in Sports Medicine*, *27*(2), 166181. https://doi.org/10.1080/15438627.2018.1502188

Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press.

Dankel, S. J., Buckner, S. L., Jessee, M. B., Grant Mouser, J., Mattocks, K. T., Abe, T., & Loenneke, J. P. (2018). Correlations Do Not Show Cause and Effect: Not Even for Changes in Muscle Size and Strength. *Sports Medicine*, *48*(1), 1–6. https://doi.org/10.1007/s40279-017-0774-3

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, *95*(450), 407–424. https://doi.org/10.2307/2669377

Dawid, A. P. (2010). Beware of the DAG! *Causality: Objectives and Assessment*, 59–86.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, *84*(1), 151–161. https://doi.org/10.1162/003465302317331982

Ding, P., & Li, F. (2018). Causal inference: A missing data perspective. *Statistical Science*,

33(2), 214–237. https://doi.org/10.1214/18-STS645

Ding, P., & Miratrix, L. W. (2015). To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias. *Journal of Causal Inference*, *3*(1), 41–57. https://doi.org/10.1515/jci-2013-0021

Edouard, P., Steffen, K., Navarro, L., Mansournia, M. A., & Nielsen, R. O. (2021). Methods matter: instrumental variable analysis may be a complementary approach to intention-to-treat analysis and as treated analysis when analysing data from sports injury trials. *British Journal of Sports Medicine*, *55*(18), 1009–1011. https://doi.org/10.1136/bjsports-2020-102155

Engist, O., Merkus, E., & Schafmeister, F. (2021). The Effect of Seeding on Tournament Outcomes: Evidence From a Regression-Discontinuity Design. *Journal of Sports Economics*, *22*(1), 115–136. https://doi.org/10.1177/1527002520955212

Fan, J., & Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, *30*(1), 74–99. https://doi.org/10.1214/aos/1015362185

Farinelli, L., Csapo, R., Meena, A., Abermann, E., Hoser, C., & Fink, C. (2023). Concomitant Injuries Associated With ACL Rupture in Elite Professional Alpine Ski Racers and Soccer Players: A Comparative Study With Propensity Score Matching Analysis. *Orthopaedic Journal of Sports Medicine*, *11*(8), 23259671231192127. https://doi.org/10.1177/23259671231192127

Fenn, T. W., Horner, N. S., Ingawa, H. S., Hevesi, M., Beals, C., & Nho, S. J. (2023). High-Level Competitive Athletes Who Undergo Hip Arthroscopy Demonstrate Durable 5-Year Outcomes and Lower Subjective Pain: A Propensity-Matched Analysis. *Sports Health*, 19417381231183658. https://doi.org/10.1177/19417381231183658

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver; Boyd.

Fredslund, E. K., & Leppin, A. (2019). Can the Easter break induce a long-term break of exercise routines? An analysis of Danish gym data using a regression discontinuity design. *BMJ Open*, *9*(2), e024043. https://doi.org/10.1136/bmjopen-2018-024043

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *The Review of Economics and Statistics*, *86*(1), 77–90. https://doi.org/10.1162/003465304323023697

Gelman, A., & Vehtari, A. (2021). What are the Most Important Statistical Ideas of the Past 50 Years? *Journal of the American Statistical Association*, *116*(536), 2087–2097. https://doi.org/10.1080/01621459.2021.1938081

Gibbs, C. P., Elmore, R., & Fosdick, B. K. (2022). The causal effect of a timeout at stopping an opposing run in the NBA. *The Annals of Applied Statistics*, *16*(3), 1359–1379. https://doi.org/10.1214/21-AOAS1545

Gonaus, C., Müller, E., Stöggl, T., & Birklbauer, J. (2023). Determining the effect of one decade on fitness of elite austrian youth soccer players using propensity score matching. *Frontiers in Sports and Active Living*, *5*. https://doi.org/10.3389/fspor.2023.1186199

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, *29*(4), 722–729. https://doi.org/10.1093/ije/29.4.722

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, *10*(1), 37.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. *Perspectives on Psychological Science*, *15*(5), 1243–1255. https://doi.org/10.1177/1745691620921521

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449. https://doi.org/10.1002/bimj.201700067

Hernán, M. A. (2018). The c-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health*, *108*(5), 616–619. https://doi.org/10.2105/AJPH.2018.304337

Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *CHANCE*, *32*(1), 4249. https://doi.org/10.1080/09332480.2019.1579578

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945960. https://doi.org/10.1080/01621459.1986.10478354

Hon, L. Y., & Parinduri, R. A. (2016). Does the Three-Point Rule Make Soccer More Exciting? Evidence From a Regression Discontinuity Design. *Journal of Sports Economics*, *17*(4), 377–395. https://doi.org/10.1177/1527002514531790

Hopkins, W. (2008). Research designs: Choosing and fine-tuning a design for your study. *Sportscience*, *12*(1), 13.

Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, *106*(493), 345361. https://doi.org/10.1198/jasa.2011.tm09599

Iacus, S. M., King, G., & Porro, G. (2012). Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, *20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

Illari, P., & Russo, F. (2014). *Causality: Philosophical theory meets scientific practice*. OUP Oxford.

Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature*, *58*(4), 1129–1179. https://doi.org/10.1257/jel.20191597

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Jimenez, A. E., Owens, J. S., Monahan, P. F., Maldonado, D. R., Saks, B. R., Sabetian, P. W., Ankem, H. K., Lall, A. C., & Domb, B. G. (2022). Return to Sports and Minimum 2-Year Outcomes of Hip Arthroscopy in Elite Athletes With and Without Coexisting Low Back Pain: A Propensity-Matched Comparison. *The American Journal of Sports Medicine*, *50*(1), 68–78. https://doi.org/10.1177/03635465211056964

Kalkhoven, J. T. (2024). Athletic Injury Research: Frameworks, Models and the Need for Causal Knowledge. *Sports Medicine*. https://doi.org/10.1007/s40279-024-02008-1

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1958.10501452

Keefer, Q. A. W. (2016). Rank-Based Groupings and Decision Making: A Regression Discontinuity Analysis of the NFL Draft Rounds and Rookie Compensation. *Journal of Sports Economics*, *17*(7), 748–762. https://doi.org/10.1177/1527002514541448

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Klein Teeselink, B., Assem, M. J. van den, & Dolder, D. van. (2023). Does losing lead to winning? An empirical analysis for four sports. *Management Science*, *69*(1), 513–532. https://doi.org/10.1287/mnsc.2022.4372

Kneafsey, L., & Müller, S. (2018). Assessing the influence of neutral grounds on match outcomes. *International Journal of Performance Analysis in Sport*, *18*(6), 892905. https://doi.org/10.1080/24748668.2018.1525678

Kruseman, M., Bucher, S., Bovard, M., Kayser, B., & Bovier, P. A. (2005). Nutrient intake and performance during a mountain marathon: an observational study. *European Journal of Applied Physiology*, *94*(1), 151–157. https://doi.org/10.1007/s00421-004-1234-y

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, *4*(3), 165–224. https://doi.org/10.1561/08

00000014

Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, *48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281

Loenneke, J. P., Buckner, S. L., Dankel, S. J., & Abe, T. (2019). Exercise-Induced Changes in Muscle Size do not Contribute to Exercise-Induced Changes in Muscle Strength. *Sports Medicine*, *49*(7), 987–991. https://doi.org/10.1007/s40279-019-01106-9

Lopes, A. D., Hespanhol, L. C., Yeung, S. S., & Costa, L. O. P. (2012). What are the Main Running-Related Musculoskeletal Injuries? *Sports Medicine*, *42*(10), 891–905. https://doi.org/10.1007/BF03262301

Lynch, B. M., Dixon-Suen, S. C., Ramirez Varela, A., Yang, Y., English, D. R., Ding, D., Gardiner, P. A., & Boyle, T. (2020). Approaches to Improve Causal Inference in Physical Activity Epidemiology. *Journal of Physical Activity & Health*, *17*(1), 80–84. https://doi.org/10.1123/jpah.2019-0515

Malisoux, L., Ramesh, J., Mann, R., Seil, R., Urhausen, A., & Theisen, D. (2015). Can parallel use of different running shoes decrease running-related injury risk? *Scandinavian Journal of Medicine & Science in Sports*, *25*(1), 110–115. https://doi.org/10.1111/sms.12154

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714. https://doi.org/10.1016/j.jeconom.2007.05.005

Mechelen, W. van. (1992). Running injuries. A review of the epidemiological literature. *Sports Medicine (Auckland, N.Z.)*, *14*(5), 320–335. https://doi.org/10.2165/00007256-199214050-00004

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781107587991

Nakahara, H., Takeda, K., & Fujii, K. (2023). Pitching strategy evaluation via stratified analysis using propensity score. *Journal of Quantitative Analysis in Sports*, *19*(2), 91–102. https://doi.org/10.1515/jqas-2021-0060

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. *Ann. Agricultural Sciences*, 151.

Nielsen, R. O., Simonsen, N. S., Casals, M., Stamatakis, E., & Mansournia, M. A. (2020). Methods matter and the 'too much, too soon' theory (part 2): what is the goal of your sports injury research? Are you describing, predicting or drawing a causal inference? *British Journal of Sports Medicine*, *54*(22), 1307–1309. https://doi.org/10.1136/bjsports-2020-102144

Nikolaidis, P. T., Veniamakis, E., Rosemann, T., & Knechtle, B. (2018). Nutrition in Ultra-

Endurance: State of the Art. *Nutrients*, *10*(12), 1995. https://doi.org/10.3390/nu10121995

Nuzzo, J. L., Finn, H. T., & Herbert, R. D. (2019). Causal Mediation Analysis Could Resolve Whether Training-Induced Increases in Muscle Strength are Mediated by Muscle Hypertrophy. *Sports Medicine*, *49*(9), 1309–1315. https://doi.org/10.1007/s40279-019-01131-8

Owens, J. S., Jimenez, A. E., Lee, M. S., Hawkins, G. C., Maldonado, D. R., & Domb, B. G. (2022). Basketball Players Undergoing Primary Hip Arthroscopy Exhibit Higher Grades of Acetabular Cartilage Damage but Achieve Favorable Midterm Outcomes and Return to Sports Rates Comparable With a Propensity-Matched Group of Other Cutting Sports Athletes. *The American Journal of Sports Medicine*, *50*(7), 1909–1918. https://doi.org/10.1177/03635465221092762

Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statistical Science*, *8*(3), 266269. https://doi.org/10.1214/ss/1177010894

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. https://doi.org/10.2307/2337329

Pearl, J. (2009). *Causality*. Cambridge university press. https://doi.org/10.1017/CBO9780511803161

Posit team. (2024). *RStudio: Integrated development environment for R*. Posit Software, PBC. http://www.posit.co/

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Richardson, T. S., & Robins, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *CSSS Technical Report*, *128*, 1–150.

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rommers, N., Rössler, R., Shrier, I., Lenoir, M., Witvrouw, E., D'Hondt, E., & Verhagen, E. (2021). Motor performance is not related to injury risk in growing elite-level male youth football players. A causal inference approach to injury risk assessment. *Journal of Science and Medicine in Sport*, *24*(9), 881–885. https://doi.org/10.1016/j.jsams.2021.03.004

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, *29*(1), 159–183. https://doi.org/10.2307/2529684

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (2022). Interview with Don Rubin. *Observational Studies*, *8*(2), 77–94. https://doi.org/10.1353/obs.2022.0009

Ruseski, J. E., Humphreys, B. R., Hallman, K., Wicker, P., & Breuer, C. (2014). Sport Participation and Subjective Well-Being: Instrumental Variable Results From German Survey Data. *Journal of Physical Activity and Health*, *11*(2), 396–403. https://doi.org/10.1123/jpah.2012-0001

Shrier, I. (2007). Understanding causal inference: The future direction in sports injury prevention. *Clinical Journal of Sport Medicine*, *17*(3), 220. https://doi.org/10.1097/JSM.0b013e3180385a8c

Shrier, I., & Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, *8*(1), 70. https://doi.org/10.1186/1471-2288-8-70

Shrier, I., Stokes, T., & Steele, R. (2020). *Methods matter: Some important challenges with instrumental variable methods*. SportRxiv. https://doi.org/10.31236/osf.io/ve4na

Snow, J. (1955). *On the mode of communication of cholera* (2nd Ed). John Churchill. http://archive.org/details/b28985266

Speer, J. D. (2023). The consequences of promotion and relegation in european soccer leagues: A regression discontinuity approach. *Sports Economics Review*, *1*, 100003. https://doi.org/10.1016/j.serev.2022.100003

Steele, J., Fisher, J., & Crawford, D. (2020). Does increasing an athletes' strength improve sports performance? A critical review with suggestions to help answer this, and other, causal questions in sport science. *Journal of Trainology*, *9*(1), 20. https://doi.org/10.17338/trainology.9.1_20

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Sun, G.-W., Shook, T. L., & Kay, G. L. (1996). Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology*, *49*(8), 907–916. https://doi.org/10.1016/0895-4356(96)00025-X

Sun, X., Lam, W.-K., Zhang, X., Wang, J., & Fu, W. (2020). Systematic Review of the Role of Footwear Constructions in Running Biomechanics: Implications for Running-Related Injury and Performance. *Journal of Sports Science & Medicine*, *19*(1), 20–37.

Taber, C. B., Vigotsky, A., Nuckols, G., & Haun, C. T. (2019). Exercise-Induced Myofibrillar Hypertrophy is a Contributory Cause of Gains in Muscle Strength. *Sports Medicine*, *49*(7), 993–997. https://doi.org/10.1007/s40279-019-01107-8

Tennant, P. W. G., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe, M. S., & Ellison, G. T. H. (2021). Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: Review and recommendations. *International Journal of Epidemiology*, *50*(2), 620–632. https://doi.org/10.1093/ije/dyaa213

Textor, J., Zander, B. van der, Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: The R package 'dagitty'. *International Journal of Epidemiology*, *45*(6), 1887–1894. https://doi.org/10.1093/ije/dyw341

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317. https://doi.org/10.1037/h0044319

Thomas, D. T., Erdman, K. A., & Burke, L. M. (2016). American College of Sports Medicine Joint Position Statement. Nutrition and Athletic Performance. *Medicine and Science in Sports and Exercise*, *48*(3), 543–568. https://doi.org/10.1249/MSS.0000000000000852

Weimar, D., & Breuer, C. (2022). Against the mainstream: Field evidence on a positive link between media consumption and the demand for sports among children. *Kyklos*, *75*(2), 317–336. https://doi.org/10.1111/kykl.12292

Westreich, D., & Greenland, S. (2013). The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, *177*(4), 292–298. https://doi.org/10.1093/aje/kws412

Williamson, E. (2016). Nutritional implications for ultra-endurance walking and running events. *Extreme Physiology & Medicine*, *5*(1), 13. https://doi.org/10.1186/s13728-016-0054-0

A Appendix

A.1 Mathematical Background

A.1.1 Probability Theory

A random variable is a property we cannot absolutely predict. The probability of the random variable X is given by $Pr(X)$. An event is the assignment of a value to a random variable. The probability of event A given that event B has occurred is the conditional probability of A given B and is denoted by $Pr(A|B)$. The events A and B are statistically independent if the observation of B does not alter the probability of A, or $Pr(A|B) = Pr(A)$. Another way to note independence is $A \perp\!\!\!\perp B$. Two events are conditionally independent if they are independent given a third event C, implying that $Pr(A|B,C) = Pr(A|C)$. This conditional independence can also be denoted as $A \perp\!\!\!\perp B|C$. The expected value of a random variable X is the weighted probability of the values it can take denoted by $E(X)$.

A.1.2 Potential Outcome Notation

For simplicity, we use a binary variable that takes on the value 0 if a unit i received no treatment and the value 1 if the unit i received treatment. Every unit i has two potential outcomes Y_i^0 and Y_i^1. These outcomes are hypothetical, as each unit only can or cannot receive a treatment and therefore only one of the two potential outcomes is realized. The observed Y_i can be defined as $Y_i = (D_i - 1)Y_i^0 + D_i Y_i^1$ with D_i as the unit-specific treatment indicator. The individual causal effect of the treatment δ_i is defined as a comparison of the two potential outcomes for each unit $\delta_i = Y_i^1 - Y_i^0$. This poses a problem, as we never observe both potential outcomes for a single unit simultaneously, as thus cannot calculate δ_i. The average treatment effect is defined by $E(\delta_i) = E(Y_i^1 - Y_i^0) = E(Y_i^1) - E(Y_i^0)$. Making the strong assumptions that $E(Y_i^1|D=0) = E(Y_i^1|D=1)$ and $E(Y_i^0|D=0) = E(Y_i^0|D=1)$ we get an unbiased estimate of the average treatment effect by calculating the simple differences in means $E(Y_i^1|D=1) - E(Y_i^0|D=0)$, which are both observed quantities. Or in other words, we obtained an unbiased estimate of the causal treatment effect by comparing the mean of the treatment group and the mean of the untreated group, if we assume that the mean of the treatment group equals the mean that the untreated group would have had if they had received the treatment (and vice versa). This implies, that the assignment of treatment was independent of the potential outcomes, or $(Y^0, Y^1) \perp\!\!\!\perp D$, something that could, for

example, be guaranteed by randomization. Often the strict independence of assignment and potential outcomes only holds when conditioning on another variable W that influenced the randomization process. The independence assumption then changes to an assumption of conditional independence $(Y^0, Y^1) \perp\!\!\!\perp D|W$. As long W, which can also be a set of covariates, is observed, we can use appropriate strategies such as sub-classification or matching to get an unbiased estimate of δ_i given the conditional independence assumption.

A.2 Simulations

For demonstrating the basic concepts of causal inference I use simulations of simple linear models. The exposure is normally distributed as $X \sim N(0,1)$. For the simplest causal inference path of $X \rightarrow Y$, Y is a linear combination of X and an (in reality unobserved) error term $U_1 \sim N(0,1)$. Therefore, the true causal effect of X on Y equals 1. More complex simulation models work in the same way, with each variable given by a linear combination of its ancestor variables and a random error term.

The causal effect in each simulation is estimated by a linear regression model. For the simplest model of $X \rightarrow Y$, this means estimating the regression coefficient b_1 of $Y = b_0 + b_1 * X + \epsilon$ via ordinary least square estimation with the R Code `lm(Y ~ X, data)`. For each simulation, the estimated regression coefficient is assumed to be the best unbiased estimate of the causal effect, creating a distribution of estimated causal effects.

A.3 Technical Details

A.3.1 Session Info

```
sessionInfo()
```

```
R version 4.4.0 (2024-04-24 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 22631)

Matrix products: default


locale:
[1] LC_COLLATE=German_Germany.utf8  LC_CTYPE=German_Germany.utf8
[3] LC_MONETARY=German_Germany.utf8 LC_NUMERIC=C
[5] LC_TIME=German_Germany.utf8

time zone: Europe/Berlin
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] patchwork_1.2.0 ggplot2_3.5.1   ggdag_0.2.12    dagitty_0.3-4

loaded via a namespace (and not attached):
 [1] gtable_0.3.5           xfun_0.43              ggrepel_0.9.5
 [4] vctrs_0.6.5            tools_4.4.0            generics_0.1.3
 [7] curl_5.2.1             tibble_3.2.1           fansi_1.0.6
[10] pkgconfig_2.0.3        data.table_1.15.4      uuid_1.2-0
[13] lifecycle_1.0.4        flextable_0.9.6        compiler_4.4.0
[16] farver_2.1.1           stringr_1.5.1          textshaping_0.4.0
[19] munsell_0.5.1          ggforce_0.4.2          graphlayouts_1.1.1
```

```
[22] httpuv_1.6.15          fontquiver_0.2.1        fontLiberation_0.1.0
[25] htmltools_0.5.8.1      yaml_2.3.8              later_1.3.2
[28] pillar_1.9.0           crayon_1.5.2            tidyr_1.3.1
[31] MASS_7.3-60.2          gfonts_0.2.0            openssl_2.2.0
[34] cachem_1.0.8           viridis_0.6.5           boot_1.3-30
[37] mime_0.12              fontBitstreamVera_0.1.1 zip_2.3.1
[40] tidyselect_1.2.1       digest_0.6.35           stringi_1.8.3
[43] dplyr_1.1.4            purrr_1.0.2             labeling_0.4.3
[46] polyclip_1.10-6        fastmap_1.1.1           grid_4.4.0
[49] colorspace_2.1-0       cli_3.6.2               ftExtra_0.6.4
[52] magrittr_2.0.3         ggraph_2.2.1            tidygraph_1.3.1
[55] crul_1.4.2             utf8_1.2.4              withr_3.0.0
[58] promises_1.3.0         gdtools_0.3.7           scales_1.3.0
[61] officer_0.6.6          rmarkdown_2.26          igraph_2.0.3
[64] gridExtra_2.3          ragg_1.3.2              askpass_1.2.0
[67] shiny_1.8.1.1          memoise_2.0.1           evaluate_0.23
[70] knitr_1.46             V8_4.4.2                viridisLite_0.4.2
[73] rlang_1.1.3            Rcpp_1.0.12             xtable_1.8-4
[76] httpcode_0.3.0         glue_1.7.0              xml2_1.3.6
[79] tweenr_2.0.3           rstudioapi_0.16.0       jsonlite_1.8.8
[82] R6_2.5.1               systemfonts_1.1.0
```

A.3.2 Packages

```
# p_used <- suppressMessages(unique(renv::dependencies(path = "../")$Package))
# p_inst <- as.data.frame(installed.packages())
# out <- p_inst[p_inst$Package %in% p_used, c("Package", "Version")]
# rownames(out) <- NULL
# out
```