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

Correlation does not imply causality, yet sport scientist are often interested in causal relationships without being able to conduct randomized controlled trials, the gold standard of causal analysis. Instead, they rely on observational or quasi-experimental data, where causal analysis is more complicated and requires specialized techniques. In this thesis, I introduce these methods of causal inference to the field of sport science. I present essential ideas of causal inference using directed acyclic graphs and demonstrate the utility of causal reasoning by revisiting two published observation studies from sport science. Introducing methods of causal inference in sport science enhances the field by limiting biases and providing new opportunities for research. This approach requires a thoughtful consideration of the causal network underlying a research phenomenon, necessitating a combination of both methodological and domain expertise. Applying causal inference methods is part of a larger five-step analytical framework proposed herein, which includes clearly stating a research goal, defining an empirical target quantity for analysis, using appropriate estimation procedures, investigating the robustness of results, and transparently communicating research findings and limitations.

Zusammenfassung (German Abstract)

Korrelation impliziert keine Kausalität, dennoch sind Sportwissenschaftler*innen oft an kausalen Zusammenhängen interessiert, ohne den Goldstandard einer randomisierten kontrollierten Studie nutzen zu können. Stattdessen müssen sie sich auf Beobachtungs- oder quasi-experimentelle Daten verlassen, bei denen die kausale Analysen komplizierter sind und spezielle Methoden erfordern. In dieser Arbeit stelle ich diese Methoden aus dem Bereich des “causal inference” in Bezug auf die Sportwissenschaft vor. Ich präsentiere wesentliche Ideen des “causal inference” unter Verwendung gerichteter azyklischer Graphen und demonstriere den Nutzen kausalen Denkens durch die Neubetrachtung zweier veröffentlichter Beobachtungsstudien aus der Sportwissenschaft. Die Einführung von “causal inference” Methoden in der Sportwissenschaft reduziert systematische Fehler in Untersuchungen und bietet neue Forschungsmöglichkeiten. Der Ansatz erfordert eine sorgfältige Berücksichtigung des kausalen Netzwerks, das einem Forschungsphänomen zugrunde liegt, und erfordert daher eine Kombination aus methodischem und fachspezifischem Wissen. Die Anwendung von “causal inference” Methoden ist Teil eines größeren, hier vorgeschlagenen fünfstufigen analytischen Rahmens. Dieser umfasst die klare Formulierung eines Forschungsziels, die Definition einer empirischen Zielgröße für die Analyse, die Verwendung geeigneter Analyseverfahren, die Untersuchung der Robustheit der Ergebnisse und die transparente Kommunikation der Forschungsergebnisse und ihrer Limitationen.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Relevance

Empirical research involves 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 a causal nature. We want to understand how sports works with the ultimate goal of intervention: If we comprehend why certain people or teams win competitions, we can use that knowledge to adjust training and tactics. Likewise, in health contexts, we seek 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 differently or if the individual had undergone 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 specific 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 various 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[1]. 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 for estimating causal effects from observational data. They can also aid in designing and analyzing experiments, and even provide benefits for description and prediction analyses.

[1] There is another line of research around causal modeling that focuses not on inference but on discovery (Glymour et al., 2019). Causal discovery can be useful when the causal structure of a system is largely unknown, such as the investigation of gene expression networks. However, this thesis will focus on causal inference.

As with all statistical analyses, causal modeling is not free of assumptions. These assumptions pertain to the underlying data and the data-generative process (the world in which the data were created). Causal modeling requires to consider these assumptions before conducting an analysis and is in general more transparent in communicating them (Grosz et al., 2020). In a way, this approach is more honest than relying on non-causal language when the actual research goal is to infer causality (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 largely 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 a different intervention. The intervention itself does not need to be practically possible, it can be purely hypothetical. For example, if we define the causal effect of biological sex on endurance performance, we are essentially asking: If we could intervene on an individual's sex (by changing it), what difference in endurance performance would we expect? We can pose this question without actually being able to change biological sex (when defined by chromosomes[2]).

It can be easy to define causal effects, but difficult to estimate them. For estimation, we can only use real data and not hypothetical data. 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 randomized controlled research design[3], but often this is impractical or even impossible.

Fisher (1925) was the first to suggest randomization as the basis for inferring causal effects in experiments. Randomized controlled designs quickly became the gold standard of experimental research (Cochran & Cox, 1957). Until the 1970s, it remained the common view that causal effects could only validly studied in randomized experiments, not in observational

[2]This "defined by 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 ambiguity makes the statement ill-defined, thus it cannot serve as a causal statement. For this thesis I follow a less strict approach by allowing causal statements that rely on (hypothetical) intervention, even if the intervention is not clearly decisive from the statement alone.

[3]See Section A.1 for the mathematical rationale behind this.

studies. But based on the earlier invention of potential outcome notation by Neyman (1923), Rubin (1974) provided a framework for estimating causal effects from both experimental and observational data. This framework, later termed the "Rubin Causal Model" (Holland, 1986), remains one of the predominant approaches to causal inference from observational data (see Section A.1 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 in estimating causal effects (Shrier & Platt, 2008). The graph-based approach has been criticized for being unnecessary (Rubin, 2022) or requiring a large number 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 (Didelez, 2022; Markus, 2021; Weinberger, 2023). In this thesis I will often follow Pearl's graph-based approach (Pearl, 2009), because it is in my view an intuitive and accessible way of learning causal inference[4], but I will also consider ideas and specific methods from the potential outcome framework (Angrist & Pischke, 2009).

Causal inference, whether within the framework of potential outcomes or through graphical representations, is considered one of the most influential 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 rely on observational data for inference. Despite its potential value, the use of causal inference in sport science remains limited.

It is not surprising that the most active research areas of causal inference in sport science are at the interface with the field of epidemiology (Lynch et al., 2020), mostly in the area of injury research. Calls to use causal modeling for researching the prevention of injuries are frequent (Kalkhoven, 2024; Nielsen, Simonsen, 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 the form of a slideshow by Hopkins (2008), the narrow scope of injuries by Shrier (2007), and the lack of an accessible, focused reasoning by both may have limited the impact of their ideas. Recently, Steele et al. (2020) undertook a new effort to highlight the need for causal thinking and modeling in sport science. Embedded in a general model of sport research (Bishop, 2008), they used an

[4] Naturally, the proponents of other frameworks will disagree on this. For example, Rubin argues that teaching 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 speaks for the accessibility of graph-based approaches for causal inference.

example of strength training to introduce key elements of causal inference such as potential outcomes and causal graphs. 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 on the causal effect of muscle hypertrophy on strength gains, all author groups agreed on the difficulties of distinguishing associations from 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), which again exemplifies the potential usefulness, but currently limited 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. He concludes the text with an appeal to all sport scientists to engage with the field of causal inference. This thesis will provide sport-scientists 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 overarching goal is to demonstrate the necessity and utility of causal inference methods for data analysis in sport science. I begin by demonstrating key concepts of causal models using directed acyclic graphs by introducing confounders, colliders, and conditioning rules. I then revisit two published observational studies from the field of endurance running from a causal inference perspective. Finally, I will discuss the 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 mathematical formulations are included in the appendix (Section A.1). My objective is to ensure that the thesis is understandable for any sport scientist with some basic statistical education. Instead of simply critiquing current statistical practices in sport science, the goal 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[5]. Conversely, an association only 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[6]. 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 distinguishes 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 relationships. In the following, we will typically denote the exposure[7] as X, the outcome as Y, and covariates with other letters. Variables in a 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 result 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 the other way around — is based purely on theoretical knowledge and understanding of the

[5] For the mathematical notation of (conditional) independence, see the Section A.1.

[6] There are of cause examples where causality can be bidirectional. For instance, in feedback loops, such as the price and demand models in economics, 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 occur 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.

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

world. There are no controlled interventions possible (because you cannot easily intervene on a person's biological sex). Ultimately, the direction of causality is an assumption by the researcher.

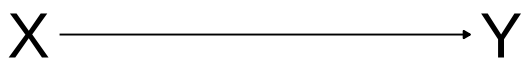

Figure 1: A simple graphical causal model with two variables. The variable X (exposure) is assumed to cause the variable Y (outcome). No other variables are assumed to influence this process.

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.

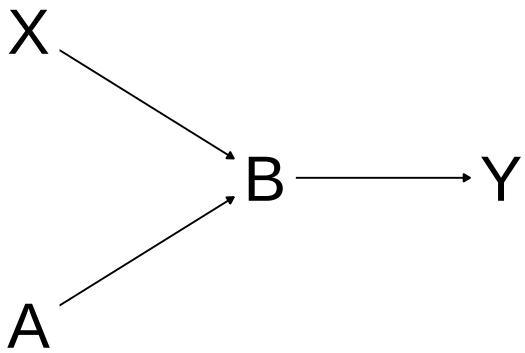

Figure 2: A more complex graphical causal model that includes four variables. X and A both cause B, which in turn causes Y.

The graph in Figure 2 is called a directed acyclic graph (DAG). It is directed, because all paths have arrows, which establish 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 the arrows a DAG contains is which 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[8]. 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 biological sex, A the nutritional 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 directly affects performance; rather, their effect are indirectly 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 one 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 precisely, as a linear structural equation model[9]). For instance, the simplest DAG in the form $X \rightarrow Y$ can be analyzed as the linear regression model $Y \sim X + \epsilon$. This assumes that Y is an additive linear combination of other variables. In this thesis, we will view 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 plays the error term ϵ.

If we knew the true causal model and could measure all variables perfectly, we could perfectly 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. These errors can include factors like random measurement error or biological variability. Furthermore, since we can 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 should have a small variance, or in other words, that repeated measurements will yield similar estimates. Random error terms add imprecision, but not bias, to our model. We will later encounter scenarios that introduce bias.

[8] In other words, if two variables are connected they may or may 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, it forms the foundation of causal inference.

[9] 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 questions from the potential outcome framework of causal inference) can be rewritten as a linear SEM, assuming the additional constraints of linearity and additive components, although SEMs can theoretically also be generalized to a non-linear setting (Bollen & Pearl, 2013). The analysis of DAGs via linear SEM can bring insights into causal systems (e.g., Ding & Miratrix, 2015).

Precision in causal effect estimates is higher in simpler models. This is primarily because simpler models have fewer random error terms. 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. Chains, therefore, introduce uncertainty into causal effect estimates but do not induce bias.

For an example from sport science, consider 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 do not influence mitochondrial density, such as motivation, pacing, or day-to-day variability.

Examining the causal model in Figure 3, we have to reconsider that the arrows drawn in a DAG are just as noteworthy as the arrows not drawn. In this example, both unobserved error terms are parent nodes, meaning that they are not influenced by any other relevant variable, including one 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 explicitly model them[10].

Figure 3: A simple causal path with random error. (a) X causes Y, but both variables are influenced by unobserved variables (random error). This adds imprecision to our model estimates, but on average, the true effect will be estimated (i.e., the model is unbiased).

2.4 Conditioning

Causal paths can be blocked by conditioning on intermediate variables. Take the causal path $X \rightarrow A \rightarrow Y$ as an example. Let X be the stroke volume of the heart, A the maximum

[10]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).

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 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[11]. 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 diminish. 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 4)[12].

Figure 4: A causal path blocked by conditioning. The causal path is blocked, because the analysis conditions on A. Since all effects of X on Y pass through A, conditioning on A means that no causal effect remains.

One of the main goals of causal inference using graph-based methods is identification — to identify which variables should be conditioned on to obtain estimates of causal effects. This process is crucial for providing unbiased and precise effect estimates. Depending on the model's structure, certain variables can introduce bias if not conditioned on, while others introduce bias if conditioned on. The following section will further elucidate these concepts by introducing confounders and colliders.

2.5 Confounders and Colliders

Confounders are variables that causally influence both the exposure and the outcome (see Figure 5 a). The confounder creates a spurious (non-causal) association between the exposure and the outcome. Conceptually, a confounder provides a set of similar knowledge to

[11] The mathematical notation of conditioning is straightforward (see Section A.1). The exact methods for conditioning are diverse and include methods that can be applied during experimental design or data analysis.

[12] Another popular way of conditioning is covariate-balancing (Stuart, 2010), which is discussed later in Section 5.2.1.

both exposure and outcome. This leads to both sharing common information, regardless of their true causal relationship, resulting in 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 from running. 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 quickly, 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 includes a confounder A, representing biological sex (see Figure 5 a). Based on expert knowledge, we understand that sex causally influences both sprinting and endurance performance, mainly via anthropometry and physiology. As a result, sex 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, of course, is not true in this simplified example here — controlling for sex gives us the true (unbiased) causal relationship between endurance and sprinting performance.

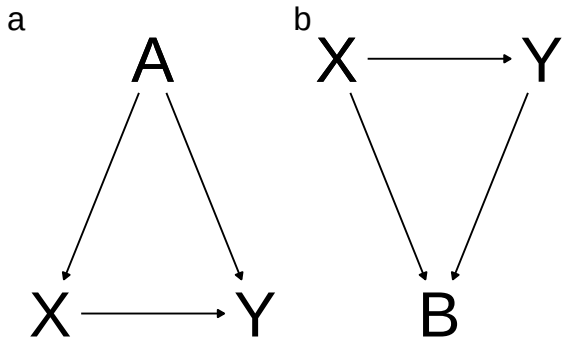

Figure 5: Confounders and colliders in a DAG. (a) A graphical example of confounding. Both X and Y share a common cause A. The confounder introduces bias in determining the causal effect of X on Y. Conditioning on A removes this bias in the analysis. (b) A graphical example of collider bias. Both X and Y directly affect the collider B. As long as B is not conditioned on, the causal effect of X on Y remains unbiased. However, conditioning on B will introduce bias into the analysis.

Colliders pose a more subtle form of bias. A collider is a variable causally influenced by both the exposure and the outcome (see Figure 5 b). Colliders themselves do not inherently cause harm. But conditioning on them introduces bias into a model[13]. This collider bias arises because a collider integrates information from both its source, the exposure and the outcome, and thus also of their causal relationship. If we condition on a collider we remove

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

some of this integrated information, which can obscure the true causal relationship between the exposure and the outcome. If the exposure X represents an experimental treatment that causes the collider B, this means that B must be a post-treatment variable. Conditioning on this collider can introduce bias into causal effect estimate, not only in observational data but also in experimental research. This is why researchers should generally avoid conditioning on post-treatment variables in their analyses (Montgomery et al., 2018).

As an example of collider bias, consider the causal relationship between X as the post-test lactate concentration in a ramp test and Y as the maximum oxygen uptake in the same ramp test. Our question is whether a higher lactate concentration causes a different (higher or lower) maximum oxygen uptake. In our model, both lactate concentration and maximum oxygen uptake influence the maximum speed achieved in the ramp test. This is reasonable because individuals with superior glycolytic or oxidatative energy metabolism are likely to outperform their counterparts that have neither in term of the maximum velocity. The maximum velocity attained thus acts as the collider B in this scenario (see Figure 5 b). Conditioning on it will introduce bias into our model.

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 (e.g., Pearl, 2009). The backdoor criterion works by two steps: 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 contains no collider or no variable conditioned on within it. It can be blocked (closed) by conditioning on a non-collider. For example, in Figure 5 a, $X \rightarrow Y$ is a causal path, while $X \leftarrow A \rightarrow Y$ is a non-causal path. The non-causal path can be blocked by conditioning on A, thus fulfilling the backdoor criterion and providing an unbiased estimate of the causal effect of X on Y.

Non-causal paths are blocked by default if they contain a collider. For example, in Figure 6, the non-causal path$X \leftarrow A \rightarrow B \leftarrow Y$ is blocked by default because B is a collider. Consequently, the backdoor criterion is satisfied and no conditioning is required. However, if one were to condition on B in this scenario (for example if A were unobserved, and we decided to condition on all observed covariates), this would reopen the backdoor-path and introduce bias into the estimate.

The backdoor criterion helps to determine which variables need to be conditioned on in graphical causal models of various complexities to obtain an unbiased estimate. These variables form the so-called minimal sufficient conditioning set. For example, in Figure 6, no conditioning is needed, and thus the minimal sufficient conditioning set is empty. Conditioning on more variables than necessary can increase precision in some cases but can also introduce

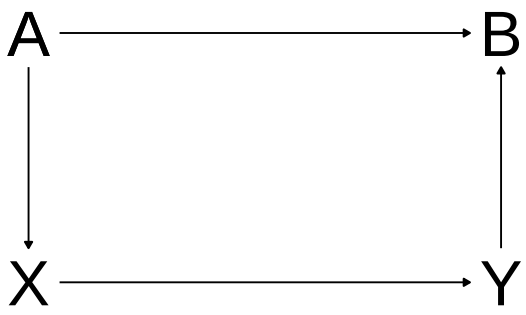

Figure 6: A graphical example of a backdoor path closed by default. The non-causal path via A and B contains a collider and is therefore closed. Conditioning on B would reopen the backdoor path.

the risk of new bias or reduced precision. When certain variables in a DAG are unobserved, they cannot be conditioned on. In such cases it may be impossible to find a minimal sufficient conditioning set that satisfies the backdoor criterion. Consequently, unbiased estimation of the causal effect, given the assumed causal model, becomes impossible.

3 Methods

For this thesis, I reviewed two exemplary research articles from sport science (Kruseman et al., 2005; Malisoux et al., 2015) from a causal inference point of view. I chose them based on a non-systematic search, implicitly following these criteria: (1) observational study research design; (2) from a subfield (endurance running) in which I have sufficient background knowledge to discuss potential causal models; (3) causal aim; (4) representative of current relevant sport science (i.e., published in respected journals and having received scientific attention in the form of citations). I also considered the potential to discuss different causal aspects in the studies and chose two studies differing in their aims and methods used. Due to the unsystematic search and screening of studies, the final choice of studies was subjective. However, I believe they not only serve to illustrate my point but also represent valid examples of contemporary sport science research.

This thesis was registered before its start using a free-form preregistration published on the Open Science Framework (Foster & Deardorff, 2017). However, as this thesis is mainly conceptual, writing a preregistration proved to be challenging. It was not possible to make any concrete statements about the planned research. The published preregistration captured my ideas for this thesis with the causal inference knowledge I had before starting to write it. Since then, both my knowledge and my ideas for the thesis have developed to the extent that ultimately this final thesis shows major deviations from its preregistration.

Originally, I planned to demonstrate techniques and principles of causal inference on a previously published data set from endurance running. However, during the process of writing, I realized that the data set is not adequate to perform causal analyses, and that it is rather impossible (and not useful) to demonstrate several general principles of causal inference on a single data set. Instead, it is more helpful to work on a concrete research example and discuss the appropriate causal inference techniques. Therefore, I decided to revisit two published studies from a causal inference view point, something I had not considered when writing the preregistration. Additionally, I originally planned to demonstrate the process of developing a causal model by creating a causal model of endurance performance. During the writing of the thesis, I realized that general causal models rarely exist (especially if the concept investigated is complex), but that appropriate causal models always depend on the exact research question and context (see Section 5.3.2 for a discussion). Thus, I did not develop a causal model for endurance performance, but developed potential causal models for the example articles I discuss. Finally, the preregistration focused mainly on the graphical model approach to causal inference. This was primarily because that was my first entry into causal inference, and at the point of writing the preregistration, I was only partly aware of other approaches. While the final thesis still focuses on graphical models in the theoretical background section, I also thoroughly discuss methods more often attributed to other causal inference frameworks (e.g., the methods in Section 5.2 that are often seen as related to the potential outcome framework). In a way, this thesis also demonstrates the process of

individual acquaintance with a research field and the difficulties in pre-planning conceptual work.

This thesis was written with Quarto version 1.3.450 (Allaire et al., 2023) in the RStudio IDE version 2024.4.0.735 (Posit team, 2024). The default settings and attached packages are documented in Appendix Section A.2. The DAGs in this thesis were drawn in R version 4.4.0 (R Core Team, 2024) using the ggdag package (Barrett, 2024), which is based on the software dagitty (Textor et al., 2016). All source code of this project is available on 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 in running practice is that the 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 changing running shoes at least twice between training sessions over the observation period. The authors fit several Cox proportional hazard regressions to the data[14]. 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. In this study, the conditions of single or multiple shoe use are not randomly assigned to a runner, but are chosen by the runners themselves (though implicitly, as the research goal was not communicated in advance). When treatment conditions are chosen instead of randomly assigned, this can introduce bias in estimating the causal effects of the treatment. Essentially, in this case the assignment to a condition is likely not independent of the (projected) outcomes. Confounding variables 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.

[14] 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).

While some baseline imbalances are natural even with randomization, this pattern indicates that some variables potentially had an influence on the 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 in causal analyses is investigating propensity score overlap. The propensity score is the probability to be assigned to one of the treatment conditions conditional on the observed covariates, and is usually estimated with a logistic regression (Rosenbaum & Rubin, 1983). Comparison of propensity score distributions between the treatment groups can help 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[15]. 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 who share similar values for covariates. Methods for covariate balancing involve matching, reweighing, and subclassification procedures (Stuart, 2010); these are discussed later in Section 5.2.1.

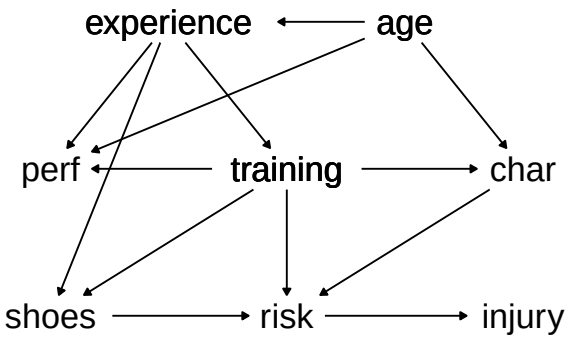

Figure 7: A potential graphical model for the effect of shoe usage on running injuries. “perf” represents race performance. “char” stand for characteristics of the biological structures (e.g., bone mineral density).

A potential DAG of the study is depicted in Figure 7. The occurrence of injuries depends on the volume of exposure and the injury risk; since exposure is controlled analytically by modeling injuries per hour of workload, it is omitted from this DAG for simplicity. In the causal model, injury risk has three potential causes: shoe usage, training characteristics unrelated

[15] In other words, we assume that we can create independence between group assignment and outcome given the observed covariates (sometimes called the “ignorability assumption”). See Section A.1 for the mathematical notation. Another assumption is that of common support. Despite covariate imbalances, some overlap between the covariate distributions of the treatment condition must exist. If this is not given (e.g., if 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 that 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.

to duration (e.g., intensity, elevation gain, running surface), and biological factors (e.g., bone mineral density). Notably, performance level is not listed as a direct cause of (increased or decreased) injury risk because there is no reason to believe that performance level itself impacts injuries directly (though it may indirectly as a proxy for experience, age, and training characteristics). While the DAG in Figure 7 is far from perfect and can be debated, it provides a potential causal model with considerable complexity and several potential confounders. Assuming this DAG adequately represents the underlying causal network, a minimal sufficient conditioning set to satisfy the backdoor criterion would contain age and training characteristics. This set is challenging to condition on, as confounding training characteristics are difficult to measure comprehensively.

Malisoux et al. (2015) acknowledge the potential impact of confounders. Consequently, 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 coefficients for other variables in the final model are interpreted causally (e.g., the effect of participation in sports other than running). Direct causal interpretation of multiple coefficients from multivariate models has been criticized as the "Table 2 fallacy" and is generally regarded as poor statistical practice (Keele et al., 2020; Westreich & Greenland, 2013). Moreover, in the absence of preregistration, this practice can lead researchers to present post-hoc hypothesis as though they were a priori (Kerr, 1998). Even when focusing solely on the primary causal effect of interest, the final model in Malisoux et al. (2015) is likely to yield a biased estimate. The "adjusted" model was selected by first performing a bivariate screening of all available variables followed by an automatic selection procedure applied to a subset of these variables (with two variables manually included). Methodological literature generally advises against bivariate screening and finds automated variable selection highly debatable (G.-W. Sun et al., 1996). Current best practices suggest using background knowledge for variable selection (Heinze et al., 2018) and, if this is not sufficient, at least applying regularization methods (e.g., Fan & Li, 2002).

Another potential source of bias lies within the definition of the treatment variable in Malisoux et al. (2015). Multiple shoe users are classified by having changed shoes at least twice during the observation period. This criterion might directly influence the outcome variable (non-injury over the observation period): On one hand, athletes who get injured subsequently drop out of the study, thus having less time to accumulate shoe changes and be categorized as multiple-shoe users. These athletes might have been classified as multiple-shoe users had they trained for a longer period instead of getting injured. On the other hand, those who voluntarily dropped out of the study were categorized as non-injured after a check. These athletes had less time to accumulate shoe changes and may therefore have been more likely to be characterized as single shoe users. Both scenarios create a non-causal relationship between the particular definition of shoe usage and injury risk. To address this potential source of bias and improve causal interpretation, providing information on observation duration could

be beneficial, such as through a survival curve (Kaplan & Meier, 1958). Additionally, directly modeling drop-out or testing the robustness of the model by using momentary rather than retrospective group assignment might offer statistical ways to handle these potential biases.

Taken together, from a causal inference perspective, the results by Malisoux et al. (2015) are subject to scrutiny. The study would benefit from a discussion of an underlying causal model (e.g., in the form of a DAG) and the application of statistical methods to address non-randomized group assignment (e.g., propensity score-based weighting). At a minimum, the definition of multiple shoe use should be reassessed, and survival curves should be included into the analysis. Additionally, the relatively small sample size (with a low absolute number of injuries) leads to imprecise estimates, even if the results were unbiased. Therefore, I concur with Malisoux et al. (2015) that either RCTs or larger observational studies should be conducted if the research question is deemed relevant. Moreover, employing appropriate causal inference methods could enhance the analysis in each case.

4.2 Example 2: Nutrient Intake and Mountain Marathon Performance

In ultra-endurance races, adequate nutrient intake is crucial for both performance and health reasons (Costa et al., 2019; Nikolaidis et al., 2018; Williamson, 2016). Athletes are advised to maintain appropriate fluid and carbohydrate intake troughout races (Thomas et al., 2016), which should in theory benefit performance. However, the effect of nutritional intake on ultra-endurance performance has been rarely investigated in field settings.

Kruseman et al. (2005) documented nutrient intake for 46 runners during an ultra-marathon mountain race. Additionally, they measured anthropometrics before and after the race and recorded 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 investigate "the association between nutrient intake and performance". To test their secondary aim, the authors devided the group into performance tertiles and tested bivariate relationships with anthropometric, running experience, and nutrient intake variables using analysis of variance and χ^2-tests. They then selected the statistically significant variables from these bivariate analyses and applied a multivariate regression model with backward stepwise selection. Kruseman et al. (2005) found that most athletes failed to meet the nutrient recommendations for the race, but no significant association with performance was observed.

The secondary aim by Kruseman et al. (2005) is causal; it is based on the hypothesis that inadequate nutrient intake impairs performance. However, they acknowledge that "[b]eing a cross-sectional, observational study, no causal relationship can be drawn between [nutrient] intake and performance", which appears counterintuitive to their research goal. Kruseman et al. (2005) found significant associations between nutrient intake and performance in their bivariate analysis, but not in the multivariate model, which adjusted (among other factors) for

previous race experience. The authors adopt a critical stance towards their own results and recommend further experimental studies[16]. In the following section, I will revisit the study by Kruseman et al. (2005) from a causal inference perspective.

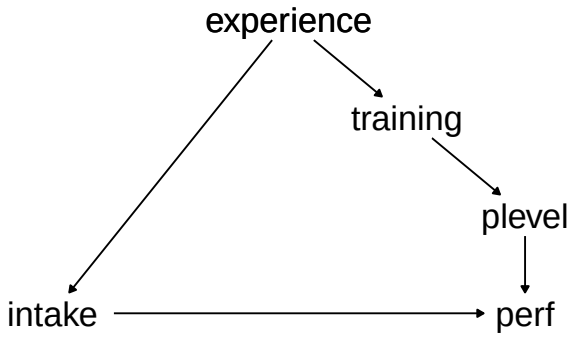

Figure 8: A potential graphical model for nutrition intake in an ultra-marathon race. "plevel" represents performance level (the physiological capability to perform the endurance task). "perf" stands for race performance.

A potential DAG for nutrient intake and ultra-endurance performance is depicted in Figure 8. Nutrient intake directly affects performance; for instance, low carbohydrate availability and dehydration can cause fatigue and thus decrease performance. Nutrient intake is largely influenced by an athlete's experience (e.g., knowledge of nutritional strategies, prior race experience). Experience also affects an athlete's training, both qualitative (e.g., experienced athletes may better tailor their training) and quantitative (e.g., more experienced athletes have had more time in their life to accumulate running training). In turn, training influences the performance level, the physiological capacity to perform the given endurance task before the race begins. This performance level, combined with nutrient intake during the race, ultimately determines the final race performance.

If the DAG in Figure 8 adequately represents the causal model underlying Kruseman et al. (2005), then the effect of nutrient intake on performance is biased due to an open backdoor path. This backdoor path could be closed by conditioning on any intermediate variable. Since experience is the only of the three intermediate variables of Figure 8 that was measured by Kruseman et al. (2005), it seams reasonable to condition the analysis on this variable[17]. Kruseman et al. (2005) acknowledge that experience is a confounder in the causal relationship between nutrient intake and performance, noting: "Because experienced runners are well trained, fitter, and know their personal needs better during such a race, it is impossible

[16] The authors' critical discussion of the causality of their findings is surprising for sport science. Considering that they correctly identified the limitations of their data analysis, I question why the authors chose to conduct such an analysis in the first place. It would have been interesting to see how critical the authors would have been if their multivariate model had indeed revealed the predicted significant effect of nutrient intake on performance instead of a null result.

[17] Conditioning on experience, rather than on training, offers additional benefits if we modify the model to allow a direct effect of experience on performance that is not mediated by training. For example, psychological readiness and pacing strategy could both influence performance and may be influenced more by experience than by training.

to precisely separate the associations we found, especially in a cross-sectional design." But given the DAG in Figure 8, 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 this was likely done rather inadvertently as their variable selection procedure was driven by automated rules rather than background knowledge. The resulting conditional effect of nutrient intake on performance is non-significant[18]. According to the DAG in Figure 8, this should be a less biased estimate compared to the bivariate associations between nutrient intake and performance that Kruseman et al. (2005) also report, but caution as potentially spurious. Interestingly, Kruseman et al. (2005) disregards both the unadjusted and adjusted estimates as biased, 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 given the futility of chasing "definitive causal relationships", it should be the aim of any researcher to minimize bias in causal effect estimates to ensure that their analysis remains meaningful at all.

A modified DAG can address some of the skepticism expressed by Kruseman et al. (2005) regarding their effect estimate adjusted for experience. If we introduce a further causal relationships from training to nutrient intake in Figure 8 , this creates an additional background path that cannot be closed by merely conditioning on experience. Kruseman et al. (2005) hint at such a relation by noting that "in addition, training increases the benefits of adequate nutrition, and favors the accumulation of muscle glycogen after exercise". This suggests that training might act as a moderator of the relationship between nutrient intake and performance. Since previous training was not documented in their study, it could not be controlled for, leaving this backdoor path open and potentially introducing confounder bias into the causal effect estimate. Again – though not in the language of causal inference – Kruseman et al. (2005) acknowledge this by stating: "It would have been interesting to record the training 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 can serve as an indirect marker of training given the DAG in Figure 8, but assuming a direct effect of training on nutrient intake the statement does not hold true. Presenting and discussing potential causal models would have reasonably benefited 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 does not include this causal knowledge, relying solely on automated variable selection procedures. Both bivariate screening and backward stepwise selection are generally inadvisable for causal inference (G.-W. Sun et

[18] Technically we do not know from Kruseman et al. (2005) 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. But again, we do not actually know this because of the automated variable selection model used.

al., 1996). Additionally, with a sample size of 46 athletes and substantial heterogeneity in covariates, obtaining precise effect estimates may be challenging. An alternative experimental approach could involve analyzing performance across different sections of the ultra-marathon race. Nutrient intake likely has varying importance depending on the race segment; for instance, it may not affect performance in the initial hour of the race. Examining the causal effect of nutrient intake by segmenting race performance could yield a better causal effect estimate and provide an additional plausibility check for the model.

5 Discussion

The aim of this thesis was to introduce methods of causal inference to sport science. After an introduction to graphical causal models, I demonstrated the utility of adopting a causal viewpoint in planning research and analyzing observational data through two examples. First, I will discuss the general application of causal inference in sport science and present special causal inference methods and their potential use in sport science. Then, I will address the limitations of causal modeling to ultimately derive a workflow for implementing causal modeling into the sport research process.

5.1 General Applications of Causal Inference in Sport Science

Causal questions in sport science require causal answers. Because of the infeasibility of randomized controlled trials in many instances (Bullock et al., 2023), much of the research relies on observational data (Abt et al., 2022). However, causal inference structures and tools to aid these analyses are largely absent from the sport science literature (Kalkhoven, 2024). This can lead to unreliable and overly speculative findings (Kalkhoven, 2024). Without these support tools, causal analyses of observational data in sport science are prone to bias, as shown in real examples (e.g., Smoliga & Zavorsky, 2017). Implementing causal inference methods in sports research can help identify true causal effects (Nielsen, Simonsen, et al., 2020; Shrier, 2007; Stovitz et al., 2019).

The current absence of causal inference methodology in sport science can be possibly attributed to a lack of knowledge in the field. Research articles dealing with causal inference in sport science are rare and are mainly limited to a handful of editorials. While these few articles have been published in high-profile sport science journals (e.g., Stovitz et al., 2019), they are often authored by individuals from outside sport science (e.g., epidemiology) and the sport science community is far from adopting the suggested methods.

Correctly applying causal inference techniques requires both a deep understanding of the domain of interest and a solid grasp of fundamentals of causal inference. As demonstrated in this thesis, causal inference is a complex field with various (and sometimes competing) frameworks and a variety of methods to consider. Given that deficits in basic statistical knowledge sometimes appear in published sports research (Sainani et al., 2021), it seems excessive to demand the teaching of causal inference in sport science curricula. However, a fundamental understanding of key concepts of causal inference (e.g., the idea of potential outcomes, confounder bias, and colliders) appear to be not just helpful but necessary for conducting any research addressing causal questions (Kalkhoven, 2024). These methods can be taught even at the undergraduate level and in field-specific educational articles, as seminal texts from psychology (Rohrer, 2018) and epidemiology (Hernán, 2004) demonstrate. Such texts are currently missing from the sport science literature, as existing articles focus on the subfield of injury research (e.g., Kalkhoven, 2024; Shrier, 2007) or barely scratch the surface

of causal inference (Nielsen, Bertelsen, et al., 2020; Nielsen, Simonsen, et al., 2020). The most compelling text so far is probably by Stovitz et al. (2019), but its editorial format limits the amount of content presented. An accessible introduction to causal inference for sport science has yet to be written and published.

As with any discussion on research methods, the application of causal inference in sport science must be viewed within the broader picture of scientific practice. Using causal inference does not eliminate other problems in research (see Section 5.3), nor is it a (partial) solution (Briggs, 2023) to the replication crisis currently affecting science[19]. Instead, its adoption should complement other measures to increase research quality in sport science, such as data sharing and preregistration (Caldwell et al., 2020). Individual sport researcher who do not adopt causal inference methods for addressing causal questions cannot be blamed, as long as external incentives for doing so are lacking. For example, collaborating with statistical experts seems ideal (Sainani et al., 2021), but is often impossible for small to medium research projects with limited resources. Learning appropriate causal inference methods by oneself requires large amounts of time, which researchers might prefer to allocate elsewhere, especially if causal inference is not yet standard in sport science publications. However, introducing causal inference in sport science by those intrinsically motivated to improve their research quality may lead to more early adopters and ultimately change the norms of the scientific field.

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 methods have become 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 tools "identification methods", because they help to identify causal effect estimates in certain common situations. These well-researched analysis tools 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 these five methods.

[19] Regarding replication issues, sport science is more of an illustrative example than an exception (Mesquida et al., 2022), as primarily results from Murphy et al. (2024) demonstrate.

[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 from observational data is to mimic the characteristics of a RCT. In an 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 previously examined study of multiple running shoe use and injury risk (Malisoux et al., 2015, Section 4.1), runners may decide if they use different shoes based on their training characteristics. If training characteristic also influence the outcome parameter of injury risk, this is an classical example of confounder bias (see Figure 5 a). To mimic a RCT of multiple running shoe use, we could decide to only compare individuals with similar training characteristic (and other covariates). This is the basic idea of covariate balancing.

Covariate balancing can broadly be classified into three categories: subclassification, matching, and reweighting (Stuart, 2010). Subclassification involves grouping individuals with similar covariates into subclasses, comparing different treatments only within these subclasses, and then calculating 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 assigns 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, meaning it occurs before the actual causal data analysis and does not include information on the outcome values. In some sense, it aims to address the same problem of observed confounders as simple conditioning in a regression does address. Current advice suggests using covariate balancing not as a replacement for regression adjustments but as a complementary approach, such as in the so called “doubly-robust” methods (Bang & Robins, 2005). A benefit of covariate balancing over regression adjustments is that it facilitates the checking of overlap in covariate distributions. If this overlap is not present, linear adjustments may perform poorly as they must rely on extrapolation, a factor often not checked in standard linear regression workflows. Ultimately, whether to use regression adjustment or covariate balancing methods to estimate a causal effect depends on the individual’s assessment of the contextual advantages and drawbacks of each strategy[21].

[21] We live in fascinating times for research. Answer posts on the online discussion forum Stack Exchange can be more detailed, informative, and engaging than possibly any journal article could ever be, as it is the case for Noah Greifer’s answer on whether to prefer regression-based methods or matching for causal inference

A critical question in covariate balancing is defining "closeness" or similarity of covariates. Exact equality on all covariates is feasible only for large samples with few discrete covariates. In most cases, researchers need 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 one's covariates (Dehejia & Wahba, 2002; Rosenbaum & Rubin, 1983). The propensity score reduces the multidimensional covariates to a single value, which can be used to form subclasses, match units, or weight observations. The exact choice of balancing method and distance measure should be context-specific, and the scientific debate about which procedure works best is vivid[22]. Stuart (2010) offers general recommendations regarding the procedure selection. Usually, it is beneficial to compare different methods (and their parameters) within the given data set.

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

In general, covariate balancing can be beneficial in sport science scenarios where there is a limited numver of interventions, a set of observed pre-treatment variable, and relatively large sample sizes. While covariate balancing methods can help reduce bias, matching effectively discards observation units and is typically unstable with small sample sizes. Therefore, these methods are particularly suited for contexts with a sufficiently large control group. For example, this could involve comparing physiological markers between injured and uninjured athletes or analyzing genetic profiles of elite athletes versus a non-elite population. Covariate balancing can also be applied in non-randomized studies investigating the health benefits of recreational sport participation. Additionally, it can aid in understanding the causal mechanisms behind team sport performance (e.g., Gibbs et al., 2022).

(Greifer, 2021).

[22] To provide a short overview of the debate: Frölich (2004) argues that weighting is always inferior to matching, a claim contested by Busso et al. (2014). Iacus et al. (2011) heavily criticize the widely used propensity-score matching (Dehejia & Wahba, 2002), proposing instead their own method of coarsened exact matching (Iacus et al., 2011, 2012), which has in turn faced opposition from Black et al. (2020).

5.2.2 Instrumental Variables

Confounders between the received treatment and outcome are often unobserved or even unknown, making it challenging to estimate an unbiased causal treatment effect directly. One approach to reduce bias is to use an instrumental variable. Instrumental variables are those that influence the outcome only mediated by the treatment. For a graphical representation see Figure 9. When a set of unobserved variables U affects both the treatment X and the outcome Y, their causal relationship is biased because we cannot control for the unobserved U. If we instead use the instrumental variable Z, which affects X directly and Y only indirectly via X, we can isolate the part of the effect of X on Y that is not influenced by U.

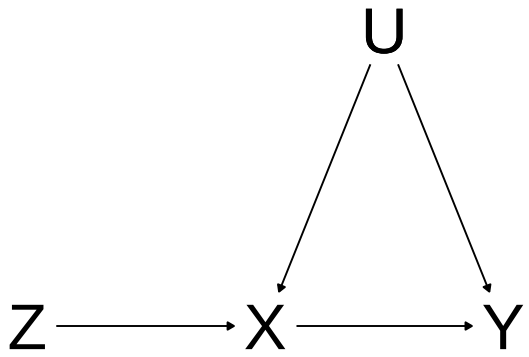

Figure 9: A graphical example of an instrumental variable. The relationship between exposure X and outcome Y is confounded by a set of unobserved variables U. The instrumental variable Z, which is unconfounded and affects Y only via X, can be used to provide an unbiased estimate of the causal effect of X on Y.

A classic example of an instrument is random treatment assignment (Greenland, 2000). People may actively decide whether to receive an assigned treatment, and these decisions could be influenced by unobserved confounders that also affect the outcome variable. While the assignment to a treatment does not directly influence the outcome, the actual receipt of the treatment does. Because individuals assigned to the treatment group are much more likely to follow the assignment and thus receive the treatment, the treatment assignment can serve as an instrument. By using an appropriate procedure (typically a two-stage least square estimator), researchers can estimate the causal effect of receiving a treatment on the outcome[23].

[23] Depending on whether we assume heterogeneity in the treatment effect, the causal effect estimated by using an instrumental variable is somewhat limited in its definition. Strictly speaking, we are estimating the causal effect of treatment only in those who adhere to the treatment assignment. This is known as the locale average treatment effect for compliers. If the focus is on the mechanistic consequences of receiving a treatment, this estimate can be quite relevant. However, if the goal is to evaluate the impact of implementing a treatment

Instrumental variables can be used to adjust for unobserved confounders and are particularly useful in addressing issues of measurement error and non-compliance. For instance, in sport science, Ruseski et al. (2014) investigated the causal effect of sport participation on happiness, a relationship that is likely confounded by unobserved variables such as 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 a positive causal effect of sport participation on happiness. Edouard et al. (2021) suggested using instrumental variables to analyze sport injury prevention programs. But Shrier et al. (2020) demonstrated that they made several severe mistakes in both their theoretical presentation and example data analysis. Their critique does not negate the potential value of instrumental variables in sport science but underscores the need for caution when adopting new methodological approaches.

The search for good instrumental variables is challenging. Aside from adhering to the assumptions depicted in the DAG structure in Figure 9, instruments should be strongly related to the treatment variable. Instruments that only moderately influence the treatment, are termed "weak instruments" (Bound et al., 1995). Weak instruments often fail to remove bias, even in large samples. Good instruments frequently involve an element of randomness. As an example in sport science, we might want to investigate the effect of being part of a youth national squad on adult sport success. Both variables are likely confounded by various unobserved factors, such as, social, psychological, and biological influences. An potential instrument in this context could be 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 affects being part of a youth national squad, because squad selection norms are based on birth years, with later-born being less mature and thus less likely to be selected. The month of birth, however, does not directly affect later adult success. Hence, the DAG in Figure 9 remains applicable, though we have to verify that the month of birth is not a weak instrument. This example highlights the potential of using instrumental variable approaches in sport science.

5.2.3 Regression Discontinuity

Regression discontinuity is one of the most widespread special methods of causal inferences. First introduced in psychology by Thistlethwaite & Campbell (1960), it gained popularity in the late 1990s (Cook, 2008) and has become one of the most common designs in causal analysis,

at a population level (e.g., for policy research), it is more appropriate to account for non-compliance in the estimation. In such cases, using instrumental variables may not be suitable.

particularly in economics (Lee & Lemieux, 2010). The basic idea of regression discontinuity is that there is a running variable determining treatment based on a fixed cut point. Individuals slightly above and below the cut point are likely to be similar in terms of observed and unobserved covariates but 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 for the estimation of a (local) causal effect of the treatment.

A key assumption of regression discontinuity is continuity. It means that if no treatment were applied (i.e., if the variable exceeding a cut point did not lead to any consequences), we would expect no discontinuities in the outcome variable. Regression discontinuity designs are typically analyzed 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 in their models, although this approach can be misleading (Gelman & Vehtari, 2021). Additionally, people aware of the relevance of the threshold may manipulate their running variable values to game the treatment assignment[24]. Therefore, analysts should check the distribution density and covariate characteristics of individuals near the threshold to detect such manipulation (Barreca et al., 2016; McCrary, 2008).

As an example from sports, consider the system of promotion and relegation in sport leagues. Estimating the causal effect of promotion and relegation on performance and financial outcomes is difficult, as teams being relegated are generally weaker and less financially equipped than teams remaining in their league. However, by only comparing teams that were just slightly above or below the cutoff for relegation, we can compare teams with likely similar background. Speer (2023) used this approach to estimate the financial effect of relegation and promotion in sport leagues using a regression discontinuity design.

Other application of regression discontinuity in sports primarily come from sport economics. For instance, Keefer (2016) and Branson et al. (2019) investigated the effects of the draft

[24]This manipulation would bias our causal estimate, as treatment assignment would no longer be essentially random around the threshold. A famous example is provided by Barreca et al. (2011), who showed an unusual high number of babies slightly just below the body weight threshold that made them eligible for intense medical care. This suggests that doctors and nurses may have deliberately manipulated weight measurement if they believed intense care to be beneficial. Not correcting for this bias yields unreliable results of the regression discontinuity analysis.

system in American basketball. Hon & Parinduri (2016) researched the impact 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 a health-related application, Fredslund & Leppin (2019) estimated the influence of a holiday break on fitness routines in the general population. One of the most popular applications of regression discontinuity designs in sport, although published in an economic journal, was by Berger & Pope (2011). They found that being slightly behind at half-time disproportionately increases the chances of winning a basketball game. They speculated that being behind by a slight margin increases motivation, leading to a performance benefit. But Klein Teeselink et al. (2023) showed that the findings by Berger & Pope (2011) did not hold in a larger sample covering more seasons and different sports.

Regression discontinuity is suited for any sport context where slight differences in an indicator cause large potentially effects in outcomes. The aforementioned concepts of team relegation and winning a game are prime use cases for regression discontinuity analyses in sport. Additional applications could include squad nomination based on fixed performance thresholds, tournament seeding based on ranking systems, or the effect of health interventions implemented based on biological measures (e.g., body mass index thresholds). Since the field of sports and exercise science is rich with indicator variables that cause treatment using somewhat arbitrary break points, a considerable 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, 1855). Over the past decades, difference-in-difference 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 to use a natural experiment in the absence of a RCT. Two groups of individuals are observed over a period of time, 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 factors are constant, the second group can act as the control group

for the treatment. The between-group difference between the two within-group differences provides 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 similarly. 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 (Section 5.2.1) or synthetic control groups (Section 5.2.5). 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, great care must be taken to calculate appropriate measures of uncertainty for estimates (Bertrand et al., 2004).

Difference-in-difference designs can be found in sport science, but they are rarely termed as such and often not systematically analysed from a causal perspective. Essentially, most analysis of covariance and time*group interaction analyses can be understood as some forms of difference-in-difference. Research in sport science that explicitly uses observational data to estimate causal effects with well-researched difference-in-differences methods is rare, likely because the spread of this design has been mainly limited to economics and social science rather than medical science. Consequently, the few articles published stem from sport economics, where authors are likely inspired by research from their primary discipline (Böheim et al., 2022; Budzinski & Kunz-Kaltenhäuser, 2020; Weimar & Breuer, 2022). However the difference-in-difference design has potential for sport science in many situations where randomized experiments are not feasible, but quasi-experimental research of observational data is possible. This includes examining the the effect of rule changes or abrupt technological advancements in sports. For example, the effects of video assistant referees in soccer or new shoe developments in running could be studied using difference-in-difference methods.

5.2.5 Synthetic Control

Synthetic control is arguably the most novel and rapidly developing special causal inference method in recent years (Athey & Imbens, 2017). It was first introduced by Abadie & Gardeazabal (2003) and first systematically presented by 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 is typically done by

weighting the covariates of the different groups from the control pool to match the covariates of the treatment group as closely as possible. For the pre-treatment period, the intervention and the synthetic control group should display similar trends in outcomes. Any differences in time trends after the treatment can then be causally attributed to the treatment.

Despite its promising and innovative nature, synthetic control methods have yet to be adopted in sport science. The primary reason may be that this research technique is still relatively new, even within its original fields of economics and political science. It may take several more years before researchers in sport science begin to apply these methods. Nevertheless, sport science presents numerous opportunities for employing synthetic control analyses. For instance, one could investigate the effect of head coach changes on team performance. Another example is evaluating the effectiveness of talent development programs in certain countries on the later success of athletes.

5.3 Challenges and Limitations

5.3.1 Need for Theoretical Models

Causal modeling is futile without causal knowledge. While general knowledge of causal modeling is necessary to perform any causal analysis, applying it to a specific research problem requires deep institutional knowledge of the underlying phenomena. Understanding of the causal structure of a research question is something that can hardly be provided by a data analyst alone; it requires expert knowledge of the research area accumulated over years. While the final estimation process may or even should be performed by a trained statistician (Sainani et al., 2021), constructing an estimation strategy should be the joint work of individuals familiar with statistical methods and those with expertise in the specific research field.

Theoretical causal models are inherently subjective. They rely on personal understanding of a research phenomena, which is not factual per se. But it is a researcher's obligation to justify their understanding of the field and the conclusion they draw for their work. Essentially, most of science can be seen as subjective, so the question should not be whether a particular scientific approach is subjective, but whether it is reasonably justified in its inevitable subjectively. In addition, scientific knowledge is not set in stone; it is a constantly evolving construct that heavily relies on its social component and is ultimately determined by consensus rather than universal truth. Therefore, the subjectivity of causal models, which form the foundation of causal inference, is not a limitation but an inherent feature of most scientific research.

Non-causal inference also relies on theoretical models, but these models are often less transparently communicated. Essentially all data analyses embed some form of institutional and statistical assumptions. Researchers without appropriate background may not realize these assumptions, as they are often made implicitly and rarely presented. Causal inference, by contrast, explicitly demands a discussion of these assumptions during the process of developing a causal model. This transparency makes it easier to discuss and understand the implications of assumptions for the research. The need for theoretical models, which may initially seem like a needless extra effort, ultimately proves to be more of a strength than a limitation of causal inference.

5.3.2 Complex Systems

When I first planned the structure of this thesis, my understanding of causal inference as a research field was fairly limited. One of my early ideas (as written in the preregistration) was to showcase the process of developing a causal model in a research field I consider myself experienced in (endurance running). However, during the writing process, I struggled with this idea and eventually abandoned my initial plan. My naive self had made two mistakes from beginning: First, it is extremely difficult to draw a general causal model of a research field, and causal models are rather specific to the exact research question. Second, it is nearly impossible to develop any exhaustive model, as causal models are always simplifications of reality.

Any causal model depends heavily on the exact research context and question at hand. For instance, the causal determinants of running performance in the laboratory might differ from those in a simulated field-performance, and both might again differ from those in an actual race. Without a specific causal research question that clearly defines the estimate of interest within its context, the causal model can vary. Thus it is nearly impossible to define a general causal model that could answer all questions regarding the causes of endurance running performance. While the first articles introducing graphical causal models in sport science present such general models for injury research (Shrier, 2007) and strength training responses (Steele et al., 2020), these models function more as illustrative examples of how causal modeling might work rather than as blueprints for future studies using causal inference in these fields. The causal models required for causal inference do not only rely on the structure of the underlying research problem but are also influenced by the exact research question of interest.

Most systems analyzed in sport embody an extreme complexity – ultimately we are investigating human behavior, which is determined by a set of biological, psychological, social, and environmental factors. It may be a tempting yet impossible task to include all potential causes of sport performance in the structure of a causal model. Therefore, any model is "wrong" in the sense that it will never capture the full reality. But this should not be the primary aim of any model (whether causal or not). Models always simplify reality, and these simplification often bring their own benefits by allowing a focus on the question of interest.

Causal inference does not require exhaustive DAGs, but rather sufficient DAGs that capture those causal relationships relevant to bias for the causal estimand. Even if a researcher determines that the current knowledge is not sufficient to draw a complete DAG for a research question, other causal information or incomplete DAGs can still provide valuable insights. Essentially, many of the causal inference methods discussed in Section 5.2 work without a complete understanding of the causal model of the researched topic. For example, determining which variables to control for in a statistical analyses to reduce bias in estimating causal effects can be answered fairly well without a full DAG (VanderWeele, 2019; VanderWeele & Shpitser, 2011). It will remain impossible to completely understand the complex systems present in sport science. However, causal inference provides a methodical approach to simplify these systems in causal models, ultimately enabling researcher to find causal relationships within these systems.

5.3.3 Small Samples

Causal inference cannot overcome difficulties inherent to small sample sizes. Given that a large portion of the sports science literature uses small samples (Abt et al., 2020), this is particularly relevant for the field. Small sample sizes generally yield imprecise estimates, or, if using a hypothesis testing framework, they have low statistical power. One of the main goals of causal inference is to provide unbiased estimates, but this only scratches the surface of the problem of precision. In other words, even an unbiased estimate may vary greatly in a small sample based on sampling variation, rendering it practically useless. However in larger observational samples, the uncertainty created by bias is generally much larger than that caused by sampling variation, highlighting the necessity of causal inference methods. Some of the causal inference methods discussed in Section 5.2 require rather large data sets (e.g., instrumental variables or some matching procedures), while others theoretically work on relatively small samples (e.g., difference-in-difference, synthetic control). The problem of

small samples is something that causal inference can neither solve, nor is this a designated goal of it. To deal with small sample sizes in sport, other approaches must be taken. This could include reconsidering study design and data analysis choices (e.g., including outside-the-trial knowledge in the data analysis with Bayesian methods) (Hecksteden et al., 2022). Small sample sizes, which are common in sport science, limit the utility of causal inference just as they do for any other statistical method.

5.3.4 Data Quality

Even an adequate model cannot answer causal questions if it has the wrong data. Data quality is a critical issue, strongly related to small sample sizes, as both decrease the precision of estimates. Sometimes, issues of data quality may be much more important than the often-discussed issues of sample size (Meng, 2018). In sport science, many commonly measured variables (e.g., physiological markers such as the maximum oxygen uptake or biological markers of training effects such as genome transcription activity) are noisy because of both biological and technical variation. In exercise interventions studies, participants may drop out for various reasons. Missing data and measurement error are two key points that hinder a causal interpretation of research findings.

Causal inference cannot solve the problem of data quality, but it offers tools to explicitly deal with measurement issues. While many popular tools to address measurement uncertainty are not uniquely causal (e.g., multiple imputation for missing data analysis, Schafer, 1999), others stem from causal inference (e.g., instrumental variables for measurement errors, Hu & Schennach, 2008), or can be directly embedded into a causal framework (Edwards et al., 2015; Smeden et al., 2021). Moreover, considering a causal model during the study planing process can help increase data quality by discussing upfront what variables to measure and how to measure them. Causal inference cannot compensate bad experimental design and low data quality. But it helps to handle these issues during research design and analysis.

5.4 Causal Modeling Workflows in Sport Science Practice

Data analysis is more than just running a single model; it constitutes a whole workflow of choosing, running, evaluating, and interpreting models (e.g., Gelman et al., 2020). Ideally, the statistical workflow should be tightly integrated with the scientific workflow. Throughout this thesis, we have seen that causal inference plays a role at different stages of the scientific

workflow. To place the causal inference methods discussed here into a broader context, I propose a research workflow for sport science that implements causal modeling practices. This workflow aims to provide sport scientists with an overview on how to integrate principles of causal inference into their research processes. Table 2 provides an overview of the different steps, which I will explain in detail in the following sections.

Table 2: A research workflow for integrating elements of causal inference into sport science.

Stage	Description
1. State a research question and a theoretical estimand	State a precise research question, specify the type of research goal (description, prediction, causal inference), and identify the theoretical (typically non-measureable) quantity of interest.
2. Identify your empirical estimand	Identify the target measure for your analysis. Clearly specify the assumptions under which your research design and/or the data analysis allow your empirical estimand to answer the research question. Consider which variables to measure and to analyze (e.g., by a causal graph)
3. Estimate the estimand	Use the actual data to perform the pre-defined analysis. The final analysis may depend on features of the data or intermediate modeling results.
4. Perform robustness checks	Check the robustness of the modeling results. This should include checks on technical aspects (e.g., model convergence), statistical aspects (e.g., sensitivity), and content aspects (e.g., plausibility).
5. Interpret the results	Interpret the model results according to your research goal. Clearly communicate the assumptions made and the robustness of the results.

5.4.1 Research Goal Formulation

All non-exploratory research projects should begin with a clear research goal. Research objectives typically fall into one of three categories: descriptive, predictive, or causal inference (Hernán et al., 2019). While edge cases exist, categorizing the research aim into one of these categories helps guide subsequent steps in the research process, as each goal entails priorities and methods of analysis (Tredennick et al., 2021). In sport science, consider an example from injury research to illustrate these distinctions (Nielsen, Simonsen, et al., 2020): Researchers might aim to *describe* the incidence of injuries in a specific athlete population, to *predict* future injuries based on training load and performance diagnostics, or estimate the *causal effect* an injury prevention program has on injury occurrence. Although these three questions could theoretically be addressed with the same (ideal) data set, the methods required to answer each question vary greatly based on the type of nature of the data analysis task.

Defining the explicit research goal before designing a study is crucial for implement goal-specific elements not only into the data analysis, but also into the research design. This does not mean that exploratory analyses are invalid or unreliable. Exploratory analyses are valuable for generating hypothesis and have historically led to many scientific discoveries. But exploratory analyses should be treated as what they are: These should be clearly communicated as exploratory rather than as confirmations of pre-existing hypothesis. Results from exploratory analyses should be interpreted with caution and ideally validated using other research methods[25]. Exploratory analyses often arise from studies with non-exploratory aims, serving as supplementary findings. Nonetheless, establishing a clear research goal at the start of the research process remains essential. Any deviations from this goal should be clearly and transparently communicated.

The research question should translate into an target quantity of interest, known as the theoretical estimand (Kahan et al., 2024; Lundberg et al., 2021). In causal research questions, the theoretical estimand often represents the outcome of a hypothetical intervention and is typi-

[25] Unfortunately, merely stating the limitations of exploratory results when presenting them does not eliminate their inherent issues. The reader is left to navigate how to correctly interpret these results. For instance, many hypothesis-testing studies in sport science have low-statistical power (Mesquida et al., 2023). It seems that underpowered research often gets published under the guise of a “pilot study”, accompanied by a standard sentence in the limitations section noting the pilot nature of the trial and the need for further studies. As long as such studies are treated as evidence by the community, simply labeling analyses as “exploratory” may not substantially change their impact or interpretation.

cally defined using potential outcomes[26]. It is a hypothetical construct that cannot be directly measured. For example, if a researcher investigates the causal effect of a strength training intervention on running performance in a group of trained runners, the theoretical estimand is the difference in performance for an individual runner if they participated in the intervention versus if they did not. Since each runner can either participate or not, this measure remains unobservable. But it still serves as the theoretical target quantity of the research. Defining and stating the theoretical estimand helps formalize the research question and clarifies how and under which assumptions the subsequent analyses of actual measured quantities can answer the question.

5.4.2 Identifying an Empirical Estimand

The actual measurable target quantity in research is the empirical estimand (Lundberg et al., 2021). The quantity is derived from the data collected given the methods used. Whether an empirical estimand actually allows to identify the theoretical estimand – the ultimate, though unmeasurable, goal of the research – depends on several assumptions.

To identify the necessary assumptions for causal research graphical causal models or other structures of causal knowledge are particularly useful. Following the previous example, in research examining the effect of strength training on running performance, the empirical estimand might be the mean group difference in running performance between an intervention and a control group. This differs from the theoretical estimand, which was the mean difference in performance for each individual. But under some assumptions, the mean group difference can be informative of the theoretical estimand we want to investigate. In this example, we might assume that the mean performance of the control group represents what the intervention group's performance would have been without the intervention, and vice versa. Randomization of treatment typically supports this assumption, but non-random treatment assignment can introduce bias into the empirical estimand. Issues such as measurement error and non-compliance (which might be addressed, for example, through an instrumental variable approach) or working with observational data, can further complicate matters. The main focus of this thesis, and causal inference research in general, is to provide methods for determining under which conditions empirical estimands can offer unbiased answers to causal questions.

[26] Following Lundberg et al. (2021), the theoretical estimand is formally defined as the unit-specific target quantity aggregated over a target population. Thus, the theoretical estimand consists of two components: the target quantity for each individual and the target group from which these quantities should be collected.

5.4.3 Estimation

Once an empirical estimand has been determined and justified, the actual data analysis process begins with the goal of estimating the desired quantity and its uncertainty based on the collected data. The estimation procedure itself is primarily a mathematical problem. Although the estimation should align with the identification method established earlier, it often does not require direct reference to the theoretical model underlying the research. For example, in traditional linear models, parameters can be estimated analytically. But other types of models might only offer numeric solution, meaning that different algorithms can yield different approximations (e.g., estimation procedures for non-linear models). This is particularly relevant for newer machine learning approaches or prediction problems, where model parameters are optimized to maximize goodness-of-fit rather than being derived from a theoretical model. Conceptually, it is useful to distinguish between this “estimation” model building and the previous “identification” model building, as both serve different roles in achieving the research goal.

5.4.4 Supplementary Analyses as Robustness Checks

In many fields, including sport science, the results from a single chosen model are often accepted at the face value. This overlooks the reality that no single true model typically exists (Box, 1976). Instead, a range of plausible models may be available, and researchers may select among them based on their outcomes (Ho et al., 2007). Thus, researchers should not only present their one primary model for answering the research question but also provide empirical evidence supporting the credibility of this model through supplementary analyses (Athey & Imbens, 2017).

Supplementary analyses can address any of the three previous stages of the modeling process (research question, empirical estimand, estimation). The robustness of research questions and theoretical estimand is rarely examined in a single research article, as it usually requires taking a totally different experimental approach. This is more commonly addressed through cumulative research across multiple projects. The robustness of empirical estimands, however, is what typical supplementary analyses focus on, which I will discuss in more detail in the following paragraph. Assessing the robustness of the estimation procedure, meanwhile, is a rather mathematical issue, focusing on whether the estimation process has been correctly executed. This is generally no issue for simple models (e.g., linear frequentist models), but

becomes crucial for more complex models where checks for convergence and stability are necessary[27].

Traditional robustness analyses focus on the identification of the empirical estimand by examining the influence of plausible deviations in model specification. This involves assessing how model results change when different sets of variables are included (Patel et al., 2015), alternative preprocessing steps are applied (Steegen et al., 2016), or various functional forms of the model are used (Young & Holsteen, 2017). A systematic approach is to develop a set of reasonable model specifications and analyze them collectively through a multimodel analysis (Steegen et al., 2016; Young & Holsteen, 2017). Other methods include explicitly modeling unobserved confounders (e.g., Ding & VanderWeele, 2016). In predictive settings, cross-validation serves as a form of robustness checks. A completely different approach to assessing robustness in a supplementary analysis involves using data where the true parameters of interest are known. This can be done by either simulating data with known parameters (Jordon et al., 2022) and then running the model to verify correct parameter estimation. Alternatively, placebo analyses (Athey & Imbens, 2017) can be conducted on data where from a theoretical view no meaningful effect is expected.

Performing supplementary analyses is an approach that highly depends on the exact research context. While individual assessments of research robustness have long been part of the scientific process, systematic methods such as multimodel analyses have been introduced only recently. Altough systematic supplementary analyses are still largely absent from sport science, they hold potential for enhancing the credibility of primary research findings.

5.4.5 Interpreting and Communicating Results

The interpretation of research findings should align with the original research goals (Kezios, 2021; Nielsen, Simonsen, et al., 2020). This means that descriptive research should be interpreted as descriptive, predictive research as predictive, and causal research as causal. Accurate causal interpretation is particularly critical. Often, current research falls into the trap of indirect causal interpretation: researcher may recognize that their methodology does not support causal conclusion but use non-causal wording while still implicating causal meanings (Hernán, 2018). For instance, studies might refer to “associations” to avoid the more causal

[27] This is an important issue, for example, in machine learning, but also in Bayesian statistics. In Bayesian models fitted using Markov Chain Monte Carlo algorithms, appropriate diagnostics should be conducted to ensure that the resulting sample accurately depicts the target posterior distribution (Roy, 2020).

terms “effects” or “relationship”, even while implying causal effects in their discussions and summaries. This practice creates ambiguity[28] (Haber et al., 2022), which can impede scientific progress (Grosz et al., 2020). Sport science research is likely affected by similar issues of misaligned reporting and ambiguous causal language, though this needs to be confirmed by empirical investigations. Nonetheless, researchers should strive to clearly communicate their findings in sports science, with explicit attention to the scope and limitations of their research (Stovitz et al., 2019).

[28] The use of ambiguous causal language is rather not an attempt to mislead readers, but more the consequence of teaching and the current scientific system. Studies that are explicit about their causal intend (and its limitation) may have poorer chances to get through peer review, leading to even more studies being published with ambiguous causal language. This may lead to the observed paradox, that researchers view studies that are causally ambiguous as of higher quality and more practically relevant than studies with a clear causal language (Alvarez-Vargas et al., 2023).

6 Conclusion

Sport science aims to understand sports performance and health, along with their underlying causes and mechanisms. Such causal research requires both appropriate research designs and analysis methods. However, randomized controlled trials, the gold standard for causal research designs, are often unfeasible in sport science, and causal analysis methods are rarely used. Leveraging causal analysis methods from fields like applied statistics, epidemiology, and economics offers valuable possibilities. These methods help distinguish mere statistical associations from true causal relationships.

As demonstrated through two examples of published research, sport science benefits from discussing the causal networks underlying research questions and selecting analysis methods accordingly. New analytical tools popular in other domains provide innovative methods for analyzing observational data in sport science, yielding estimates with less bias than traditional methods (though they cannot solve existing problems with small samples and low-quality data). Furthermore, systematic causal data analysis typically offers greater transparency in communicating assumptions and interpreting research findings.

Causal data analysis requires collaboration between domain experts in data analysis and those in the investigated research field, as the analysis is never blind to the context in which the data was collected. Causal inference methods play a crucial role in a broader reform of the scientific workflow, which includes clearly stating research questions and aims and testing the robustness of findings. Embedding causal thinking and analysis methods can ultimately enhance sport science's pursuit for knowledge.

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

Abt, G., Boreham, C., Davison, G., Jackson, R., Nevill, A., Wallace, E., & Williams, M. (2020). Power, precision, and sample size estimation in sport and exercise science research. *Journal of Sports Sciences*, *38*(17), 1933–1935. https://doi.org/10.1080/02640414.2020.1776002

Abt, G., Jobson, S., Morin, J.-B., Passfield, L., Sampaio, J., Sunderland, C., & Twist, C. (2022). Raising the bar in sports performance research. *Journal of Sports Sciences*, *40*(2), 125–129. https://doi.org/10.1080/02640414.2021.2024334

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

Alvarez-Vargas, D., Braithwaite, D., Lortie-Forgues, H., Moore, M., Wan, S., Martin, E., & Bailey, D. H. (2023). Hedges, mottes, and baileys: Causally ambiguous statistical language can increase perceived study quality and policy relevance. *PLOS ONE*, *18*(10), e0286403. https://doi.org/10.1371/journal.pone.0286403

Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (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 University 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://doi.org/10.1287/mnsc.1110.1328

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. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (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

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.2307/2286841

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

Briggs, W. M. (2023). A partial solution for the replication crisis in economics. *Asian Journal of Economics and Banking*, *7*(2), 180–190. https://doi.org/10.1108/AJEB-03-2023-0027

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

Caldwell, A. R., Vigotsky, A. D., Tenan, M. S., Radel, R., Mellor, D. T., Kreutzer, A., Lahart, I. M., Mills, J. P., Boisgontier, M. P., Boardley, I., Bouza, B., Cheval, B., Chow, Z. R., Contreras, B., Dieter, B., Halperin, I., Haun, C., Knudson, D., Lahti, J., … Consortium for Transparency in Exercise Science (COTES) Collaborators. (2020). Moving Sport and Exercise Science Forward: A Call for the Adoption of More Transparent Research Practices. *Sports Medicine*, *50*(3), 449–459. https://doi.org/10.1007/s40279-019-01227-1

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://doi.org/10.3386/w4509

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. https://doi.org/10.12987/9780300255881

Dankel, S. J., Buckner, S. L., Jessee, M. B., Mouser, J. G., 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

Didelez, V. (2022). Perspective on interviews with Heckman, Pearl, Robins and Rubin. *Observational Studies*, *8*(2), 95–104. https://doi.org/10.1353/obs.2022.0010

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

Ding, P., & VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, *27*(3), 368. https://doi.org/10.1097/EDE.0000000000000457

Edouard, P., Steffen, K., Navarro, L., Mansournia, M. A., & Nielsen, R. Ø. (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

Edwards, J. K., Cole, S. R., & Westreich, D. (2015). All your data are always missing: Incorpo-

rating bias due to measurement error into the potential outcomes framework. *International Journal of Epidemiology*, *44*(4), 1452–1459. https://doi.org/10.1093/ije/dyu272

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.

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association*, *105*(2), 203–206. https://doi.org/10.5195/jmla.2017.88

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

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). *Bayesian workflow*. arXiv. https://doi.org/10.48550/arXiv.2011.01808

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

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00524

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.

Greifer, N. (2021). *Why do we do matching for causal inference vs regressing on confounders?* Cross Validated. https://stats.stackexchange.com/q/544958

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

Haber, N. A., Wieten, S. E., Rohrer, J. M., Arah, O. A., Tennant, P. W. G., Stuart, E. A., Murray, E. J., Pilleron, S., Lam, S. T., Riederer, E., Howcutt, S. J., Simmons, A. E., Leyrat, C., Schoenegger, P., Booman, A., Dufour, M.-S. K., O'Donoghue, A. L., Baglini, R., Do, S., … Fox, M. P. (2022). Causal and associational language in observational health research: A systematic evaluation. *American Journal of Epidemiology*, *191*(12), 2084–2097. https://doi.org/10.1093/aje/kwac137

Hecksteden, A., Kellner, R., & Donath, L. (2022). Dealing with small samples in football research. *Science and Medicine in Football*, *6*(3), 389397. https://doi.org/10.1080/24733938.2021.1978106

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. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, *58*(4), 265–271. https://doi.org/10.1136/jech.2002.006361

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

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing

for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(3), 199–236. https://doi.org/10.1093/pan/mpl013

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.

Hu, Y., & Schennach, S. M. (2008). Instrumental Variable Treatment of Nonclassical Measurement Error Models. *Econometrica*, *76*(1), 195–216. https://doi.org/10.1111/j.0012-9682.2008.00823.x

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

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). *Synthetic data – what, why and how?* arXiv. https://doi.org/10.48550/a

rXiv.2205.03257

Kahan, B. C., Hindley, J., Edwards, M., Cro, S., & Morris, T. P. (2024). The estimands framework: a primer on the ICH E9(R1) addendum. *BMJ*, *384*, e076316. https://doi.org/10.1136/bmj-2023-076316

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

Keele, L., Stevenson, R. T., & Elwert, F. (2020). The causal interpretation of estimated associations in regression models. *Political Science Research and Methods*, *8*(1), 1–13. https://doi.org/10.1017/psrm.2019.31

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

Kezios, K. L. (2021). Is the way forward to step back? Documenting the frequency with which study goals are misaligned with study methods and interpretations in the epidemiologic literature. *Epidemiologic Reviews*, *43*(1), 4–18. https://doi.org/10.1093/epirev/mxab008

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/10.1561/0800000014

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

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, *86*(3), 532–565. https://doi.org/10.1177/00031224211004187

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

Markus, K. A. (2021). Causal effects and counterfactual conditionals: contrasting Rubin, Lewis and Pearl. *Economics & Philosophy*, *37*(3), 441–461. https://doi.org/10.1017/S0266267120000437

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*, *14*(5), 320–335. https://doi.org/10.2165/00007256-199214050-00004

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, *12*(2), 685–726. https://doi.org/10.1214/18-AOAS1161SF

Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2022). Replication concerns in sports and exercise science: A narrative review of selected methodological issues in the field. *Royal Society Open Science*, *9*(12), 220946. https://doi.org/10.1098/rsos.220946

Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2023). Publication bias, statistical power and reporting practices in the journal of sports sciences: Potential barriers to replicability. *Journal of Sports Sciences*, *41*(16), 15071517. https://doi.org/10.1080/02640414.2023.2269357

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It. *American Journal of Political Science*, *62*(3), 760–775. https://doi.org/10.1111/ajps.12357

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

Murphy, J., Caldwell, A. R., Mesquida, C., & Warne, J. P. (2024). My research got published, so what's the big deal? The need for formal replication in sports science. *European Congress of Sport Science 2024*. https://osf.io/4pu8v

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. Ø., Bertelsen, M. L., Møller, M., Hulme, A., Mansournia, M. A., Casals, M., & Parner, E. T. (2020). Methods matter: exploring the 'too much, too soon' theory, part 1: causal questions in sports injury research. *British Journal of Sports Medicine*, *54*(18), 1119–1122. https://doi.org/10.1136/bjsports-2018-100245

Nielsen, R. Ø., 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

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*(9), 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

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

Roy, V. (2020). Convergence Diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application*, *7*(1), 387–412. https://doi.org/10.1146/annurev-statistics-031219-041300

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/jp

ah.2012-0001

Sainani, K. L., Borg, D. N., Caldwell, A. R., Butson, M. L., Tenan, M. S., Vickers, A. J., Vigotsky, A. D., Warmenhoven, J., Nguyen, R., Lohse, K. R., Knight, E. J., & Bargary, N. (2021). Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. *British Journal of Sports Medicine*, *55*(2), 118–122. https://doi.org/10.1136/bjsports-2020-102607

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, *8*(1), 3–15. https://doi.org/10.1177/096228029900800102

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

Smeden, M. van, Penning de Vries, B. B. L., Nab, L., & Groenwold, R. H. H. (2021). Approaches to addressing missing values, measurement error, and confounding in epidemiologic studies. *Journal of Clinical Epidemiology*, *131*, 89–100. https://doi.org/10.1016/j.jclinepi.2020.11.006

Smoliga, J. M., & Zavorsky, G. S. (2017). Team Logo Predicts Concussion Risk: Lessons in Protecting a Vulnerable Sports Community from Misconceived, but Highly Publicized Epidemiologic Research. *Epidemiology*, *28*(5), 753–757. https://doi.org/10.1097/EDE.0000000000000694

Snow, J. (1855). *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

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

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

Stovitz, S. D., Verhagen, E., & Shrier, I. (2019). Distinguishing between causal and non-

causal associations: implications for sports medicine clinicians. *British Journal of Sports Medicine*, *53*(7), 398–399. https://doi.org/10.1136/bjsports-2017-098520

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

Tredennick, A. T., Hooker, G., Ellner, S. P., & Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, *102*(6), e03336. https://doi.org/10.1002/ecy.3336

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

VanderWeele, T. J., & Shpitser, I. (2011). A New Criterion for Confounder Selection. *Biometrics*, *67*(4), 1406–1413. https://doi.org/10.1111/j.1541-0420.2011.01619.x

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

Weinberger, N. (2023). Comparing Rubin and Pearl's causal modelling frameworks: a commentary on Markus (2021). *Economics and Philosophy*, *39*(3), 485–493. https://doi.org/10.1017/S0266267121000353

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

Young, C., & Holsteen, K. (2017). Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis. *Sociological Methods & Research*, *46*(1), 3–40. https://doi.org/10.1177/0049124115610347

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 δ_i of the treatment 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 and 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 obtain 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 is sometimes called the exchangeability assumption. It 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 (set) 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 is observed, we can use appropriate strategies such as sub-classification, matching, reweighting, or conditioning, to get an unbiased estimate of δ_i given the conditional independence assumption.

A.2 Technical Details

A.2.1 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] glue_1.7.0              httpcode_0.3.0          xml2_1.3.6
[79] tweenr_2.0.3            jsonlite_1.8.8          R6_2.5.1
[82] systemfonts_1.1.0
```

A.2.2 Packages

```
p_used <- suppressMessages(unique(renv::dependencies(path = "../")$Package))
```

```
Finding R package dependencies ... Done!
```

```
p_inst <- as.data.frame(installed.packages())
out <- p_inst[p_inst$Package %in% p_used, c("Package", "Version")]
rownames(out) <- NULL
out
```

```
    Package Version
1   dagitty   0.3-4
2 flextable   0.9.6
3   ftExtra   0.6.4
4     ggdag  0.2.12
5   ggplot2   3.5.1
6 patchwork   1.2.0
7      renv   1.0.5
8 rmarkdown    2.26
9     shiny 1.8.1.1
```