The application of Evidence-Based Medicine methodologies in sports science: problems and solutions

William Levack-Payne
Department of Philosophy
University of Kent

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy
2021
Word count: 92,135

Declaration

I, William Levack-Payne, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

18/11/2021

Dedication

To Roy Levack, for the affirmations, and for making me pull my finger out and apply.

When I said,

> I'll get a PhD, all I have to do is keep going,

and you said,

> That's fine, Levacks don't quit,

I don't think you knew how important that was.

Acknowledgements

Firstly, I would like to extend deep gratitude to Professor Jon Williamson and Dr. Michael Wilde, my co-supervisors. Whilst I may not have enjoyed it at the time, their resolute willingness to critique and question even the smallest details in this thesis helped me refine my arguments and ideas in a way I could not have done alone. In the same vein, I express great thanks to post-graduate students and research staff at the sports science department at the University of Brighton, not just for putting up with me critiquing their field, but also for being valuable in their contribution to my understanding of key physiological mechanisms and discussions about the state of research methodology.

I am deeply grateful, also, to those in the Philosophy department at the University of Kent who, on a wintry walk by the Kent coast during my masters, told me I should consider a PhD to be a sensible next step. I would also like to thank Dr. Matthew Sinnicks who proofread the application to my masters degree after I had studied physics for three years and forgotten how to write, without whom I would not have a masters, let alone have applied for a PhD.

I offer, also, special thanks to fellow PhD candidates Daniel Auker-Howlett, and Joe Jones, my unofficial secondary supervisors, for the many impromptu common room seminars, but mostly for telling me when my reasoning was ‘stupid’, or my ideas were ‘half-baked’. Without you, Jon and Michael would have read some terrible things. I, also, thank Joe

and Daniel for their valuable contributions in proof-reading my work, and their affirming words about the final draft of my thesis.

Last, and by no means least, my friends, family, and partner, must be thanked from the bottom of my heart. My friends, Luke, Matt, Jav, Nick, and Joe, have not only dealt with me exclusively talking about my research for the last three years, but they have also at least pretended to listen almost the entire time. This must have been particularly hard when, for a year, our social time was entirely virtual. My partner Abbie too, has my everlasting gratitude, for always being there through the course of my PhD, for always getting excited for me when I can't, for being my biggest fan, and for finding the time to proofread my entire thesis whilst also navigating her first term at Cambridge. And, of course, my family, Andy, Annie, and Naiomi. Your contribution to my thesis has extended far beyond the academic. Having your love, having you to confide in, and having your constant support, has been one of the most integral parts of completing this thesis.

For small erections may be finished by their first architects;
grand ones, true ones, ever leave the copestone to posterity.
God keep me from ever completing anything.

Herman Melville, Moby-Dick or, the Whale

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Situating the thesis

This thesis is concerned with the philosophy of sports science and rarely engages with the wider philosophy of sport, unless mentioning it as an aside. This thesis sits closely to, and is informed by, much work in medical epistemology. Particularly, it draws on work by the Evidence-Based Medicine Plus (EBM+) group, and their 2018 publication *Evaluating evidence of mechanisms in medicine: principles and procedures.* Accordingly, the thesis also engages with work of both Evidence-Based Medicine (EBM) proponents, and commentators (e.g. Cartwright, 2007; Gillies, 2017b; Guyatt et al., 1992; Howick, 2011b; Sackett et al., 1996). However, despite drawing largely on, and providing arguments relevant to, current work in medical epistemology, its main focus is sports science, and looking at the application of what we know about medical epistemology to the sports sciences. Work of this ilk is hard to come by. In fact, at the time of writing, only Jukola, 2019 engages directly with similar questions to those addressed in this thesis. Some sports scientists are still concerned with the importance of philosophy to sports science. This thesis often draws on works in the edited collection *Philosophy and the* *sciences of exercise, health and sport* (McNamee, 2004) for philosophical discourse on scientific matters in sport. Outside this, others (e.g. McFee, 2009) have written philosophically about research in sport, but are concerned with different questions to the ones I contend with. Particularly, they are concerned with the methods of research investigating sport itself, rather than the methods of research in sports science.

1.2 Thesis outline

This thesis has one main, overarching, aim: to argue that evidence from mechanistic studies should be taken seriously in the sports sciences. I argue that this should be the case both when establishing and explaining causal claims, and when informing practice. This thesis has three complementary parts. In Part I, I argue that neither RCTs, nor RCT-style N of 1 trials in sports science can be assumed to provide high-quality evidence. In Part II, I argue that when establishing causal claims in the sports sciences, it is necessary to establish the existence of both a correlation, and a mechanism. Given the difficulty of establishing the existence of mechanisms from non-mechanistic studies, this motivates the importance of assessing evidence from mechanistic studies. In Part III of this thesis, I argue that going beyond just helping to establish the *existence* of a mechanism, providing details of mechanisms has other highly beneficial uses in sports science. As I will explain, this is counter to trends in evidence-based fields. In these fields, there is often a trend towards conducting research with a view to finding effective interventions without worrying about being able to explain why those interventions work. I give two reasons for the importance of being able to explain, at least in part, how mechanisms work by providing mechanism details. Firstly, by improving our theories and understanding of causal relationships, we are better able to explain and interpret the results of investigations, for example by aiding in questions of extrapolation. Secondly, providing details of relevant mechanisms can be used to improve and streamline how we develop and test interventions for practice by potentially making it quicker and cheaper. Just like Part II of the thesis, the arguments from Part III of the thesis also motivate the importance of assessing evidence from mechanistic studies as a source of evidence of details of mechanisms.

Parts I and II of this thesis are motivated by the EBM+ research project, particularly the collaborative publication *Evaluating Evidence of Mechanisms in Medicine* (2018). This project argued in favour of, among other things, the importance of assessing evidence of mechanism and evidence from mechanistic studies in medicine. This was intended to challenge and improve the current EBM practice of privileging Randomized Controlled Trials (RCTs) and dismissing the importance of evidence from mechanistic studies in establishing causal claims. The move to EBM in medicine has prompted the call to move to an evidence-based framework in sports science, often called Evidence-Based Practice (EBP). This call, and what EBP entails, is discussed in section 1.4. Just as EBM is concerned with integrating the best available evidence, such as evidence that establishes the safety and efficacy of treatments into practice and care (Straus et al., 2019, Guyatt et al., 1992), the primary goal of EBP is to rely on the best possible evidence when informing practice (Amonette et al., 2016). This idea is uncontroversial and is a key tenet of EBM and EBP. What is controversial is what counts as the *best possible evidence*. EBP practitioners would claim that the best possible evidence, as highlighted by the evidence hierarchies discussed in subsection 1.4.1, comes from RCTs and systematic reviews and meta-analyses of these. In Parts I and II of this thesis, I argue against this claim. I argue that if we assess evidence from mechanistic studies in addition to evidence from RCTs when assessing causal claims in sports science, this better constitutes relying on the best possible evidence. I will call the claim that this better fulfils the EBP aim of relying on the best possible evidence the **Better Evidence Thesis**.

A large part of current EBP methodology involves relying on and 'privileging' RCTs as a high-quality method of evidence gathering (Ivarsson and Andersen, 2016, 11), and the idea that using evidence sourced from RCTs is a good method of fulfilling the goal of relying on the best possible evidence. Part I of this thesis assesses the quality of evidence that may be obtained with the use of RCTs and RCT-style N of 1 trials in sports science. In chapters 2 and 3 of this thesis, I argue that RCTs and single subject RCT-style trials (N of 1 trials) often provide low-quality evidence in the sports sciences. This is because, alone, they often fail to exclude possible alternative explanations for observed outcomes, thus failing to rule in the tested intervention as an explanation of observed outcomes. This claim helps to make up the **Excluded Explanations Argument**, which is discussed in depth in chapter 2. I, then, argue that this is troubling for EBP as it looks like, by relying on RCTs as a primary evidence gathering source, they are attempting to rely on the best possible evidence by relying on evidence that is not of a high quality, or at least not *the best*.

Given the problems associated with relying primarily on evidence from RCTs, Part II of this thesis looks at what we can do better when assessing causal claims. Chapters 4 and 5 of the thesis are concerned with suggesting how EBP may progress how it assesses evidence for causal claims in order to advance the **Better Evidence Thesis**. I argue for the **Better Evidence Thesis** on the grounds that if evidence from mechanistic studies is assessed alongside evidence from association studies, this helps those assessing evidence to avoid key pitfalls outlined in the **Excluded Explanations Argument**. In aid of this, in chapter 4, I introduce the Russo-Williamson thesis from the philosophy of medicine. This is, broadly, the claim that establishing causal claims in medicine generally requires establishing both a suitable correlation, and the existence of a mechanism responsible for that correlation (Russo and Williamson, 2007). In chapter 4, I argue that the RWT is true in sports science, and I defend it from prominent critics in medicine. Following this, in chapter

5, I argue that in order to establish causal claims in medicine, given the RWT, we will generally need to assess evidence from mechanistic studies, as well as association studies, in order to establish both correlation and mechanism existence. The outcome of this is that, by assessing evidence from both mechanistic and association studies, we may avoid our evidence falling foul of the **Excluded Explanations Argument**. This means that EBP practitioners may inform practice and make decisions based on better evidence than if they relied on current EBP methodology. As such, EBP should take seriously evidence from mechanistic studies. Importantly, I call it the **Better Evidence Thesis** because I do not want to preclude the idea that other research in the future may indicate further ways in which we can modify EBP methodology beyond this. The argument of this thesis is intended to build on and improve current EBP methodology. I do not claim that the arguments of this thesis will perfect it.

In Part III of this thesis, I argue that the importance of providing details of mechanisms in sports science extends beyond how this may help to establish causal claims. This is in contrast to Part II of the thesis, where I only argue in favour of using evidence of mechanism to establish causal relationships. I do this by utilising two in-depth case studies and arguing that we cannot provide good explanations of observed results without invoking details of mechanisms to do so. This goes against a trend observed in evidence-based fields, and one which is adopted into EBP. The trend is that understanding the mechanisms that underlie causal relationships is unimportant, partly as it is not seen as an important part of being able to determine efficacy and prescribe interventions. This trend often leads to publications that prescribe specific interventions without being able to say what it is about the intervention that makes it effective. In chapter 6, I provide a case study of the FIFA 11+ injury prevention programme, which happens to be one of these cases. The 11+ is interesting because being prescribed the FIFA 11+ significantly reduces injury rates in some populations, and not in others. In chapter 7, I introduce the case of exercise interventions for obesity, and the fact that unsupervised interventions are rarely as effective as supervised interventions. The FIFA 11+ case focuses on physiological mechanisms, but also touches on some social/psychological mechanisms. The obesity intervention case focuses on social and psychological mechanisms. It is important to consider cases that touch on the social/psychological and physical/biological. This is because sport has both social and physical factors. In both instances, I argue that if we do not understand the mechanism by providing details of that mechanism, we cannot provide good explanations for the observed difference in outcomes in each case. This point is relatively uncontroversial. The controversy arises, however, when I argue that EBP, which is often unconcerned with providing explanations, should instead be concerned with being able to provide explanations. In aid of this, in both cases, I argue that providing mechanism details can improve and streamline intervention development and our ability to interpret the results of trials. Both of these outcomes, I argue, are important to the goals of EBP. Finally, in chapter 9, I argue that we can extrapolate the lessons from both of these case studies to the general case and that providing details of mechanisms in EBP is more important than trends in evidence-based fields would suggest. I argue that the arguments relating to the FIFA 11+ and exercise interventions for obesity extend naturally to general cases in sports science. The work of chapter 8 is informed by the works of Donald Gillies, and Rani Lill Anjum and Stephen Mumford.

In the remainder of this introduction, I introduce why there is a move towards EBP, and focus on the doctrine of using the best possible evidence. This will set up the next chapter, and the rest of the thesis. It explains why it is important that I argue that evidence from RCTs alone may not be the best possible evidence, especially not in sports science.

1.3 What is sports science?

Generally, sports science is taken, often tacitly, with exercise science to be the science of sports and exercise. In this thesis, I construe it in this broad sense too, but use the simple term ‘sports science’ for brevity. According to The British Association of Sport and Exercise Sciences (BASES), it can be broadly thought of as having three disciplines which can be considered to come under the remit of sports and exercise science when the research relates to sport and exercise (BASES, n.d.). These are (BASES, n.d.):

- biomechanics, which is concerned with how and why the body moves as it does, using a mechanical lens,
- physiology, which examines the effects of exercise and training on the body, and which is often linked with nutrition, and
- psychology, which looks at human behaviour as it relates to sport and exercise.

Much research in sports science is also interdisciplinary, involving some or all of these disciplines.

BASES explains sports science in the following way:

> Sport (sic) science tends to refer to the application of sport and exercise science principles within high-performance sport, where the application of science is concerned with maximising the performance of an athlete or team (BASES, n.d.).

BASES defines the exercise sciences in the following way:

> Exercise science refers to the application of sport and exercise science principles within health and fitness, where the application of science is primarily concerned with the improvement of physical and mental health through exercise. This covers both the role that exercise can play in preventing poor health and chronic diseases, such as coronary heart disease and diabetes, and the role of exercise in treating a variety of physiological and psychological disorders (BASES, n.d.)

In some more niche areas of the literature, a further distinction exists. This distinction is discussed by McFee in his book *Ethics, knowledge and truth in sports research: An epistemology of sport* (2009). He raises a distinction between 'sports science' research that is conducted in a 'naturalistic' sense, and sports research that is sociological, historical, and philosophical in its research (McFee, 2009, 4). This first kind of research in sport includes areas that follow the methods of the natural sciences: biomechanics, sports psychology, and physiology. This is the type of research that this thesis is concerned with. This second type of research includes topics such as investigating the religious attitudes of long-distance runners, or explaining the culture of football fans. Whilst McFee does occasionally discuss what he calls naturalistic sports science in this text, it is sports research of the second type that McFee attempts to provide an epistemology of in his book. In light of this, despite appearing on the surface to discuss similar topics, this thesis is not in opposition to, or even in conversation with, McFee's work.

Research in the sports sciences can be basic, applied, or some mixture of the two (Cooper and Nevill, 2005 117). According to Nevill, basic research is concerned with investigating theories that underpin phenomena, whereas applied research is concerned with questions about whether some intervention has a worthwhile effect on some outcome in real-world settings (2005, 117). In addition to these points made by Nevill, basic research is also concerned with providing explanations and models, and applied research also involves prediction and diagnosis.

Many parts of sports science have similarities with medical sciences, and often, such as in the case of sports medicine, there are overlaps. There are clear differences, however. Primarily, and most obviously, sports science is concerned with research relating to sport and exercise, and whilst these outcomes may be health-related outcomes, they will often not be. One key characteristic of sports science that differentiates it from medicine is that research is often conducted to see how performance can be improved, normally restricted by the criteria of a specific sport, and often beyond normal human functioning. Although sports science does aim to increase health on a national level, such as aiming to reduce the (as of 2012) £8.7 billion cost to Britain of inactivity (Select Committee on Science and Technology, 2012a, 8, paragraph 3), it also works on more niche problems that affect very few individuals, aiming to find improvements in sports-specific performance as small as a fraction of a percentage which can make all the difference when it comes to podium placement in the Olympics (Atkinson and Nevill, 2001).

In fact, in some cases, such as in the case of elite athletes, sports science interventions or decisions may be utilised that improve performance at the *expense* of physical and mental health. For instance, despite it being against the rules in most sporting federations, the use of illegal performance enhancing drugs, such as anabolic androgenic steroids (AAS), occurs in sport.[1] What also occurs is the use of non-illegal performance enhancing drugs through *therapeutic use exemptions*. This is where it is determined that an athlete can use a medication if it is *'proved'* that they need it for therapeutic reasons. As an example, it was revealed that the number of Olympic swimmers with a therapeutic exemption for asthma medication far outweighs population levels of asthma, this allows the use

[1] Some sports, such as bodybuilding and powerlifting, do have *untested* federations where the use of AAS is not *technically* permitted for sponsorship and legal reasons, but where there is an understanding that they will be used by all athletes because no drug testing is done to determine if an athlete is 'clean'.

of potentially performance enhancing asthma medications for all of those athletes (Herzog, 2017, 47).[2] Drugs used with the intention to improve performance, particularly AAS, often have negative mental and physical health outcomes (Pope et al., 2014, 341).

Other legal interventions used in sports to improve performance can also do this at the expense of health. In many sports, such as endurance sports like long-distance running, performance can be improved by lowering body-weight to improve mechanical efficiency. This practice often leads to negative changes in physiological function including 'metabolic rate, menstrual function, bone health, immunity, protein synthesis, cardiovascular and psychological health' (Mountjoy et al., 2014, 491). This is, of course, not to say that sports science is never concerned with long-term health. These are simply useful examples to illustrate key differences between medical and sports sciences.

1.4 Why EBP in the sports sciences?

In this section, I will introduce the rationale behind adopting an EBM-like methodology in the sports sciences. I will, then, explain what EBP in the sports sciences involves, and what cues it takes from EBM. I will also introduce a pre-EBP example that illuminates the dangers of utilising low-quality evidence to inform practice. This serves as motivation to adopt EBP principles in evidence assessment: the explicit evaluation of the best possible evidence in sports science.

EBM was introduced into medical practice with the aim of ensuring the: 'conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients' (Sackett et al., 1996, 71). This can be seen as a stepping stone in the history of medicine to-

[2] For an interesting philosophical overview of therapeutic use exemptions in sport, see Pike 2018.

wards practising based on higher quality evidence. Medicine has moved from being informed by expert opinion and authority, to observation and experiment, to EBM, which provides a structured framework for evidence and its assessment. Just as EBM is an improvement on previous methodologies, EBM+ seeks to extend this by providing a better way in which evidence can and should be assessed in the health sciences. EBM+ looks at the epistemological and practical foundations of EBM and argues that evidence from mechanistic studies should be taken seriously by medicine in the assessment of causal claims, in addition to the other sources of evidence it relies heavily on. Discussions of this are given in depth in chapters 4 and 5. Classic EBM either places little emphasis on evidence derived from mechanistic studies and non-experimental sources, or dismisses it entirely, as they claim that: ‘these routinely lead to false positive conclusions about efficacy’ (Sackett et al., 1996, 72). At the same time, RCTs, and reviews of them, are seen to be ‘gold standard’ methods of gathering evidence, given that they are seen as ‘so much more likely to inform us and so much less likely to mislead us’ (Sackett et al., 1996, 72). In chapter 2, I explain why RCTs are perceived as being a gold standard method of gathering evidence.

As will be discussed in subsection 1.4.1, classic EBM utilises hierarchies that can be used to rank the quality of evidence for a claim by method of evidence gathering. These can then supposedly be adjusted based on the quality of individual instantiations of that method of evidence gathering (see for example: Ebell et al., 2004; OCEBM Levels of Evidence Working Group, 2011; Schünemann et al., 2013). For instance, a poorly conducted RCT may be ranked down, and a well-conducted observational study ranked up. The perceived success of EBM has encouraged many other fields to adopt evidence-based approaches. One simply needs to perform a Google Scholar search for one of many practical fields to find papers outlining how to engage in that field in an evidence-based way. For instance, policymaking, policing, and teaching all return numerous search results. In the sports sciences, a call has been made, that has been gaining traction ever since, to adopt an evidence-based approach (see for example: Knudson et al., 2014; MacAuley, 2000; Steves and Hootman, 2004). A collective aim espoused by those wanting to adopt an evidence-based methodology in the sports sciences is the desire to promote practices in sport that have their foundations in science. The practices in sport should be justified by proven efficacy, determined according to an EBM-like methodology of evidence gathering and evaluation (Knudson et al., 2014, 196).

The call for an EBP paradigm in the Sport Sciences has left the realm of theory and entered practice (Knudson et al., 2014, 195). There are myriad examples of this. Many sports organisations seek to promote, teach, and publish evidence-based guidelines, practices, and research methodologies. For instance, BASES in the UK, and the National Strength and Conditioning Association (NSCA) and the American College of Sports Medicine (ACSM) in the USA. Further to this, many private organisations related to the sports sciences have sprung up with the aim of distilling or interpreting evidence for their clients so that they may inform their practice with high-quality evidence, without necessarily needing to be able to interpret that evidence themselves. Well-known companies touting evidence-based methodology include: Precision Nutrition, Renaissance Periodization, and Stronger by Science.

As has already been stated, the adoption of an EBP framework in sport and the sports sciences marks a move towards adopting evidence gathering and evaluation principles from EBM. A foundational idea in both EBM and EBP is that practitioners should use the best possible evidence, so that they can be sure their practice is likely to cause the desired outcome, and to know to what extent the practice will be productive of that outcome. This is, in part, a response to how practice in sport has been informed historically. As it was put by MacAuley, in the case of sports science, in a paper calling for evidence-based practice: ‘Until we can provide science to underpin clinical practice, we are open to the accusation that we are simply making it up as we go along' (2000, 258). Importantly, the call for evidence-based practice in sport is not a dismissal of all practices informed by what EBP considers to be low-quality evidence. Instead, the claim made by EBP proponents is that we should *now* seek to inform and recommend practice using the best possible evidence.

In the sports sciences, as well as placing an emphasis on understanding and explaining phenomena, much research is intended to be put to practical ends. EBP is almost entirely concerned with practical ends, as its key goal is informing practice in sport. Sport is a results driven field. The practices engaged in are responsible for at least some of those results. So, it is clearly important that practitioners employ practices that we can justifiably claim can produce the best results. The kinds of practices that those making the call for EBP see as benefiting greatly from being informed by the best possible evidence include:

- Promoting safe exercise (Hootman, 2007, 13)
- Providing the best exercise programmes that can be tailored to individuals goals (Amonette et al., 2010, 454)
- Allowing clinicians to improve the level of care and quality of interventions they can provide patients to return athletes to the field sooner (Steves and Hootman, 2004, 84, Prentice, 2014, 20-21, McKeon et al., 2006)
- Developing better diagnostics for sports therapy and training use (Bleakley and MacAuley, 2002, 124, MacAuley, 2000)
- The improvement of physical and mental health through exercise (BASES, n.d.)

There are also reasons outside pure sports applications for employing EBP in the sports sciences: if sports practitioners and researchers can engage in EBP, and provide research that contributes to this, there is a hope that this will elevate the reputation of the sports sciences (Steves and Hootman, 2004, 84, MacAuley, 2000, 258, Bleakley and MacAuley, 2002, 125, Amonette et al., 2010, 454). It has even been argued that, as sports science takes on evidence-based principles, and it becomes more possible to engage in practice based on higher-level evidence, it can be seen as an equal to other fields where EBM principles are employed, and even medicine (MacAuley and Best, 2007, XIV).

As I have previously stated, EBP can be seen to be a reaction to the way that practice was informed in sport in the past. When EBM was emerging, people rightly asked: if medicine is based on evidence now, what was it based on before (Howick, 2011b, 3)? Given the call for EBP in sports science, this same question can be asked of sport, as it calls into question the evidence used to inform historical practice. This is a large part of what provides strong motivation for the adoption of EBM-like principles, such as the idea that we should rely on the best possible evidence. The evidence-based framework would consider much of the evidence that informed practices prior to the widespread adoption of EBP to be low in quality. What is meant by the 'quality of evidence' is discussed in detail in section 1.5. Broadly, however, low in quality meaning that, as evidence, it does not provide strong justification that a causal relationship exists between practice and putative effect. The implication being that, in employing the practice, one can be less sure it will be productive of the desired outcome than if its effectiveness has been justified using high-quality evidence. For instance, much practice has been informed by unstructured observations, as opposed to structured research (MacAuley, 2000, 255 Bleakley and MacAuley, 2002, 124). Anecdotal evidence, publications without peer reviews, and personal experience were also common reasons given for engaging in certain practices (Amonette et al., 2010, 450). Practice without an informed evidence base is particularly troubling outside higher-level sports. As put by Amonette et al.:

> Exercise science is susceptible to misinformation and bogus claims – perhaps more than any other field. This is evident from a cursory knowledge of the personal training industry. (2010, 450)

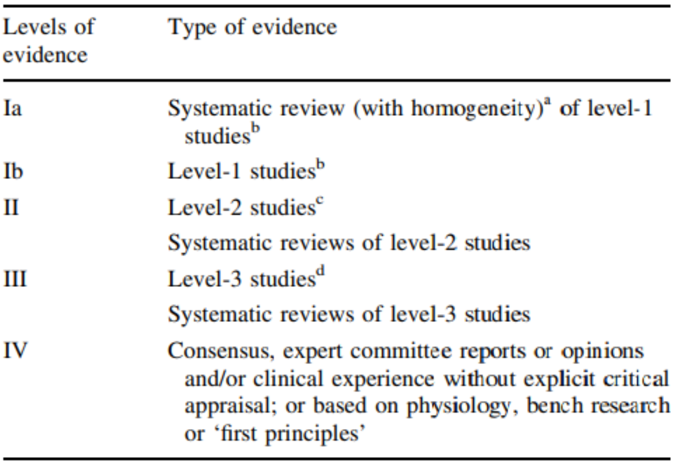
Also, as sports science is a relatively new field and is evolving rapidly, it is often the case that practice is employed that is based on research that was relevant during a practitioner's education, but which has since been outmoded (Amonette et al., 2010, 451). This leads to the adoption or retention of inferior or even harmful practices, not supported by high-quality evidence. For instance, many of us will remember having the importance of stretching before exercise being inculcated into us during childhood sports lessons in school. Now, however, most high-quality evidence suggests that this is either pointless, or can harm certain sporting outcomes (Barbosa et al., 2020). Pre-EBP (and even now, although it is less common), some professional journals for sports practitioners did not even include citations to support the practices they promoted, or made recommendations that failed to account for, or even went against, the larger body of evidence (Knudson, 2005, 215).

So far, we have seen the types of practice adopting an EBP framework can be useful for, and why it was necessary for a call to be made for practice in sport to be based on the best possible evidence. What remains to be seen is, what engaging in EBP looks like.

1.4.1 Evidence hierarchies

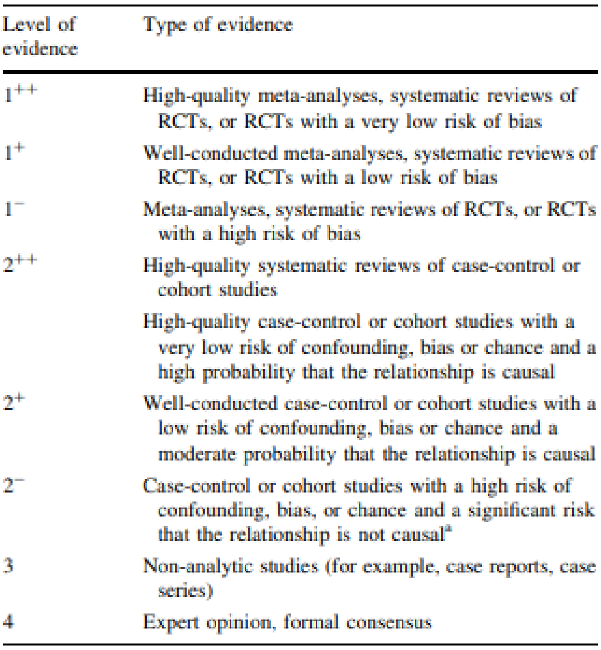
Just like EBM, engaging in EBP involves both understanding how strongly different types of evidence provide support for causal claims, and how to interpret evidence to inform practice. Practitioners need to be able to understand study results and how to interpret them, with particular importance being paid to understanding how to evaluate the quality of evidence supporting a recommendation for a particular practice (McKeon et al., 2006, 42). This involves understanding study type and design, and from that, being able to infer the meaningfulness of its results (Medina-McKeon and McKeon, 2009, 4). The claim is made in both EBM and EBP that not all types of study can be taken as providing evidence of equal quality. Different types of study are, purportedly, more or less susceptible to different types of weakness and bias that can influence their results (Knudson et al., 2014, 197). It is also claimed that the strength of evidence provided by different types of study can be ranked according to how susceptible they are to different types of bias. In order to help practitioners in medicine and sport interpret the quality of evidence research provides for a causal claim, a number of *evidence hierarchies* have been produced that aim to rank the assumed strength of evidence these different research methods provide for those causal claims (Knudson, 2005, 214, Amonette et al., 2010, 452, Prentice, 2014, 22, Medina et al., 2006, 38). One key reason why different methods of gathering evidence are ordered into hierarchies is so that it can help people determine if the research they are reading indicates that a proposed causal relationship is genuine. As such, practitioners who wish to employ a practice should, according to the hierarchical way of seeing, engage in practice supported by evidence types at the top, rather than the bottom of the hierarchy. The ordering, then, is intended to make this easier to accomplish.

Many evidence hierarchies exist: Figure 1.1, Figure 1.2 and Figure 1.3 are typical of the type of hierarchy that one may see in medicine. Popular hierarchies in medicine include the Levels of Evidence from the Oxford Centre for Evidence-Based Medicine (see: Howick et al., 2011), and the Strength of Recommendation Taxonomy (see: Ebell et al., 2004). Similar to evidence hierarchies, what also exist are evidence rating tools such as the Grading of Recommendations Assessment Development and Evaluation (see: Guyatt et al., 2011, Prentice, 2014, 22). These tools aim to provide a method by which one can determine the quality of evidence a particular study or set of studies provides in favour of a claim by allowing it to be rated up or down based on deficiencies and strengths in those studies. Others have also sought to develop their own, more specific to sports issues, such as the hierarchy produced by Knudson et al. (2014), as seen in Figure 1.4.



Levels of evidence	Type of evidence
Ia	Systematic review (with homogeneity)[a] of level-1 studies[b]
Ib	Level-1 studies[b]
II	Level-2 studies[c]
	Systematic reviews of level-2 studies
III	Level-3 studies[d]
	Systematic reviews of level-3 studies
IV	Consensus, expert committee reports or opinions and/or clinical experience without explicit critical appraisal; or based on physiology, bench research or 'first principles'

Figure 1.1: The evidence hierarchy of diagnostic studies outlined by NICE (NICE, 2006, 48).



Level of evidence	Type of evidence
1^{++}	High-quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1^{+}	Well-conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias
1^{-}	Meta-analyses, systematic reviews of RCTs, or RCTs with a high risk of bias
2^{++}	High-quality systematic reviews of case-control or cohort studies
	High-quality case-control or cohort studies with a very low risk of confounding, bias or chance and a high probability that the relationship is causal
2^{+}	Well-conducted case-control or cohort studies with a low risk of confounding, bias or chance and a moderate probability that the relationship is causal
2^{-}	Case-control or cohort studies with a high risk of confounding, bias, or chance and a significant risk that the relationship is not causal[a]
3	Non-analytic studies (for example, case reports, case series)
4	Expert opinion, formal consensus

Figure 1.2: The evidence hierarchy of intervention studies outlined by NICE (NICE, 2006, 47).

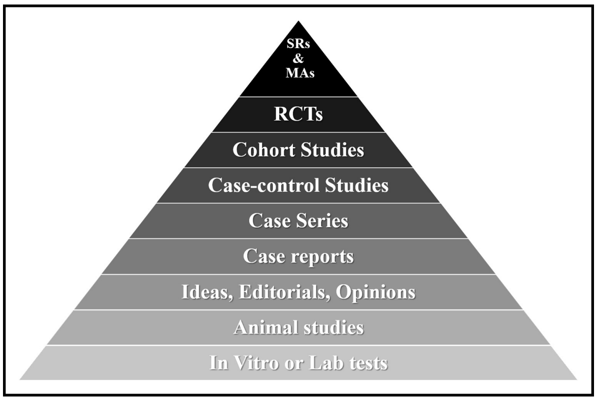
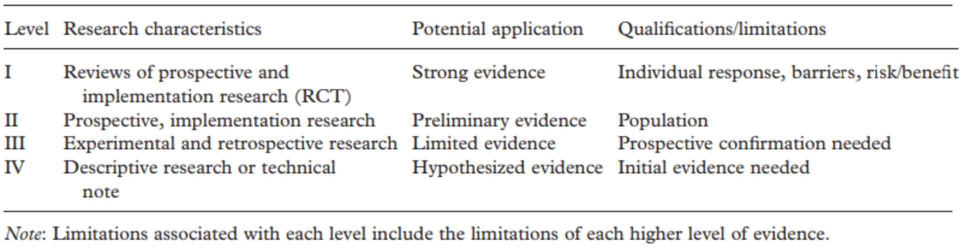


Figure 1.3: An evidence pyramid typical of those seen in medicine, (Mulimani, 2017, 2).



Level	Research characteristics	Potential application	Qualifications/limitations
I	Reviews of prospective and implementation research (RCT)	Strong evidence	Individual response, barriers, risk/benefit
II	Prospective, implementation research	Preliminary evidence	Population
III	Experimental and retrospective research	Limited evidence	Prospective confirmation needed
IV	Descriptive research or technical note	Hypothesized evidence	Initial evidence needed

Note: Limitations associated with each level include the limitations of each higher level of evidence.

Figure 1.4: An evidence hierarchy specific to sports science, reproduced from Knudson et al., 2014

Hierarchies generally rank the same types of study similarly. It is almost always taken, within literature in the sports sciences discussing evidence hierarchies, that RCTs, and systematic reviews and meta analyses of RCTs, provide the highest quality of evidence (see for example: Bleakley and MacAuley, 2002; Knudson et al., 2014; MacAuley and Best, 2007; McKeon et al., 2006; Medina-McKeon and McKeon, 2009; Prentice, 2014). This means that EBP takes RCTs as one of the best primary sources of evidence.[3] This is because, in theory, RCTs should be a good tool for ruling in interventions being tested as the cause of observed outcomes. This is largely due to the perceived ability for RCTs to be able to rule out explanations for an observed correlation between an intervention or exposure, aside from that the outcome was caused by the interven-

[3] RCTs are a primary source of evidence in that they do not rely on obtaining evidence from other studies in the way that systematic reviews and meta-analyses do.

tion or exposure being tested. In cases where no correlation is observed, RCTs are also considered to be good at ruling that, in these instances, the tested intervention or exposure has no effect. I argue, in chapter 2 and 3, that in sports science, it is often unlikely that RCTs and N of 1 trials can do this well. Expert opinion, and mechanistic studies such as bench research, are often rated as types of study that provide the lowest quality evidence of effectiveness, and are thus often inadequate to inform practice (Amonette et al., 2010, 451). Knudson et al. claim that evidence from mechanistic studies, or descriptions of how mechanisms lead to an outcome, provides no 'actual evidence of potential outcome' (2014, 200).

When we consider that before EBP much practice in sports was based on anecdote, expert opinion, and unstructured observations, and that these sources of evidence are ranked lowly or not at all in the hierarchies, from an EBP point of view, this speaks poorly for much past practice, and helps to explain why sports science is concerned with basing practice on evidence of a higher standard. As such, as EBP sees it, relying on evidence from sources at the top of evidence hierarchies is a major part of fulfilling the aim of relying on the best possible evidence. This is contrary to what I claim with the **Better Evidence Thesis**.

1.4.2 Hydration case study: the dangers of informing practice with low-quality evidence

One of the key motivators for adopting EBP, just like EBM, is how badly we can get things wrong when low-quality evidence is used. As an example of recommendations based on low-quality evidence, I will introduce a case now that I will use again in subsubsection 2.3.2.1 and again in section 8.5. In subsubsection 2.3.2.1 I discuss the problem of active controlling trials with harmful or useless treatments. In section 8.5 I discuss how providing details of relevant mechanisms is useful as it helps us to interpret results of trials. Here, however, I explain how poor evidence has been used, prior to a more widespread adoption of EBP principles, to inform poor recommendations, the type of thing EBP aims to avoid. The following case study on hydration recommendations is informed by the discussion of hydration research in Tim Noakes' work: *Can we trust rehydration research?* (2004). The example of rehydration research is useful here because, whilst probably not as dangerous as some past medical practices, such as using bloodletting for almost all maladies, it is an example of the potential dangers of informing practice with low-quality evidence.

From 1975 until the late 90s, the ACSM, the US Army, and the National Association of Athletic Trainers began to offer recommendations about how athletes should hydrate during exercise. This recommendation was to drink as much water as was tolerable whilst exercising. This was intended to help athletes replace all fluids lost through sweating, and to reduce risk of heat illnesses. Some guidelines recommended almost as much as 2 litres of fluid an hour during exercise. Not only was this meant to reduce heat illness risk, but it was also believed that it would help maximise potential performance. The research that prompted these recommendations proceeded to set up a 'foundational myth', and, thus, it was a widely accepted 'scientific belief' that maximal tolerable fluid ingestion was necessary to maximise performance and reduce heat illness risk (Noakes, 2004, 138).

It has since come to light that these recommendations were not only useless, in some cases they were harmful. High and sustained rates of fluid ingestion can lead to brain swelling and dysfunction (hyponatraemic encephalopathy), and, in some recorded cases, death. According to Noakes (2004), it has also since been shown that fluid intake during exercise is not necessary to reduce heat illness in sport. In fact, dehydration is now known to not reduce sweat rate, and thus, does not impact on the rate of cooling and heat illnesses in athletes. This is demonstrated by the fact that athletes regularly experience moderate levels of dehydration during sport without negative health consequences, and that even those who experience more severe dehydration, such as those walking 8-hours in desert conditions, tolerate dehydration well.

The advice is also not what would now be referred to as ‘evidence-based’ (Noakes, 2004, 137). This is because the evidence that supported it was not high-quality: it had not been peer reviewed, and was not obtained from ‘properly conducted’ studies which should ‘exclude all other possible interpretations’ (Noakes, 2004, 137). The evidence that supported these guidelines arose from a study where a correlation was observed between weight loss over the course of a running race, and rectal temperatures at the end of the race. The assumption was then made by researchers that inadequate fluid intake caused the high rectal temperatures, and that those high rectal temperatures indicated an increased risk of heat illness. Noakes claims that the researchers did not have sufficient high-quality evidence to assume that the relationship between fluid intake and rectal temperatures was directly causal, and that if the evidence had been evaluated using evidence-based guidelines, this relationship would not have been proposed as causal (Noakes, 2004, 141). In fact, high-quality evidence now suggests that higher internal temperatures and sweat rates (and thus acute weight loss) have a shared causal factor: metabolic rates (Noakes, 2004, 143). Those who ran faster both had increased sweat rates, and higher internal temperatures, caused by the metabolic stress from running faster, and not caused by increased dehydration. Thus, the recommendation to increase fluid intake would not combat internal temperatures or sweat rate, as dehydration at these levels has no direct effect on either of those outcomes. It is instances like this, that by adopting EBP, sports science aims to avoid. An EBP proponent would claim that if proper studies were conducted and the quality of evidence were interpreted correctly, a potentially dangerous practice would not have become so widespread.

Of course, one cannot interpret this case study as standing for all historical sports practices. Not all non-evidence-based practices have been useless or dangerous. Much of past practice has been based on what EBP sees as low-quality evidence, but this does not mean that the quality of practice did not improve before EBP. If we measure performance in sport by world or Olympic records, or the feats that athletes are able to complete, we can see that sports performance has largely been moving forwards. We need only look at an athletics record table or compare videos of the types of performance gymnastic athletes are capable of now in comparison to the 1900s to see progress. These improved performance outcomes are likely, in part, because coaches and athletes were able to determine which training programmes are more beneficial to performance than others, what performance enhancing drugs and protocols to use, and what technologies better facilitate performance. In order to determine this, practitioners must have used *some* types of evidence to inform their practice, no matter how low quality, such as unstructured observation and expert opinion. Using lower quality types of evidence does not necessarily mean that effective practices will not be developed. It just means that, among other things, when we employ them we cannot be as sure that they will be useful.

Coaches, for example, who developed training programmes for their athletes would not always have had access to rigorously controlled RCT data or reviews and analyses of relevant literature to inform their practice (Gilbert and Trudel, 2004, 388). However, coaching practices have changed (Gilbert and Trudel, 2004, 396, Carpenter, 2012), and athlete and team performances have improved. The evidence upon which many training programmes have been based is often comprised of observational data from coaches and their athletes, mechanistic reasoning, and intuitions (Day, 2011, 179, Carpenter, 2012, 172). Whilst it is reasonable to dispute the quality of evidence produced by expert judgement, if a coach performs that role for their entire adult life as a full-time commitment, they will likely have many individuals on which to base their observations, access to the same individuals for long periods of time to make measurements and see changes, and other field experts to share and discuss methodologies with. In fact, in a recent pithy Tweet by leading sports science researcher Brad Schoenfeld, when discussing how we inform practice for sports where strength and muscular development are important, he claimed that sports science will always be catching up with what happens in practice:

> The best applied research comes from the field; thus, research will always be "catching up" with what bodybuilders do in practice. The goal of science is to systematically and objectively test the validity of these practices to draw evidence-based opinions vs relying on anecdote (2020).

The sentiment of this Tweet can be applied across sports science. As has been stated, in the eyes of EBP, these methods of observation, case reports, mechanistic reasoning, and expert opinion are seen to be poor methods of gathering evidence by evidence hierarchies. However, they may still produce some reasonably effective practices. As performance has improved, what is viewed as poor evidence must still have informed some good practice. Regardless, the move towards adopting a methodology where we rely on the best possible evidence is still important. It is having a strong evidential basis for the claim that an intervention will actually have the desired effect that is important. We may end up employing the same intervention, but when employing the best possible evidence, we will be justified in our use of it, instead of risking using something suboptimal.

To give an example of 'poor evidence' informing good practice, we may look at the use of interval training in sports. What are now typically seen as low-quality methods of gathering evidence informed the use of interval training in sports, particularly running, through the 20th century (Noakes, 1991, chapter 8). Interval training is the practice of exercising for shorter amounts of time, with a rest between each exercise *interval*, rather than one extended amount of time. It now a highly researched and used training modality, with this research indicating its established effectiveness (Laursen et al., 2002, 1801, Daussin et al., 2008), but this was not always the case. Mechanistic reasoning in particular was used to inform the inclusion of interval training in running training (Noakes, 2004, 276). The reasoning behind the inclusion of interval training in sport being that if one ran 20 lots of 100 meters with a rest, rather than one straight 2000 meters, your running speed for the total distance would be greater, which should then translate by improving your straight 2000-metre time (Noakes, 1991, chapter 8). Mechanistic reasoning, such as this, is placed at the bottom of the hierarchies of evidence, but, as in the case of interval training, has informed practice still widely and effectively used. However, this is not to say that the increased effectiveness of interval training as opposed to, or in addition to, continuous training was established based on the reasoning that was originally used to inform its use. An EBP proponent, may, for example, argue that in these cases, good practice is arrived at by chance and that whilst the practice is good, it was not right to adopt it as the evidence in its favour was insufficient to justify it. Alternatively, it may be the case that, in this instance, the mechanistic reasoning was supported by an un-masked mechanism, and was sufficient to justify the use of interval training.

1.5 Clarifications

Before continuing with the thesis, I must briefly make some clarifications.

1.5.1 Metaphysics and epistemology

In terms of philosophical methodology, this thesis takes what might be termed an *epistemology first* view of questions relating to evidence, causation, and practice. This means that this thesis focuses on epistemological questions relating to evidence, causation, and practice. This thesis shows how far we can get with this type of methodology without introducing or doing any metaphysics. This is, of course, not to say that metaphysical questions, such as what a mechanism is or what causation is, are not important. What the methodology of this thesis shows is just how far we can get without answering metaphysical questions, and how productive this methodology is.

As this thesis is concerned only with the epistemology surrounding questions of causation, evidence, and practice, I will constrain my discussion of medicine and sport to the epistemological. One may argue that it is impossible to ask and answer these types of questions without having some metaphysical implications. For instance, Anjum and Mumford (2018, 246) claim that the use of any method to research causal claims makes metaphysical claims about causality, even if it is implicit. They claim that, if one conducts research to determine if the presence of A makes a difference to B, this says something about what it is for something to cause another thing. As such, in making claims about needing evidence of mechanism and correlation to establish causation, my work *may* have metaphysical implications. I will raise some metaphysical questions that arise in this thesis in the conclusion, but will leave addressing them for future work.

1.5.2 Mechanistic studies

Mechanistic studies are studies which provide evidence that a cause gives rise to an effect by giving evidence for details of the mechanism by which the cause gives rise to the effect (Parkkinen et al., 2018, 14). As will be seen throughout this thesis, and particularly in part III, it is this ability to provide evidence for details of mechanisms between cause and effect that is what is so important about mechanistic studies. Conducting mechanistic studies often means conducting bench research in a lab, but can include things like: ‘in vitro experiments, biomedical imaging, autopsy, established theory, animal experiments and simulations’ (Parkkinen et al., 2018, 14). An important note, and one which will be discussed again in subsection 4.5.2 is that association studies can, by providing evidence of features of a mechanism between cause and effect, also be mechanistic studies (Parkkinen et al., 2018, 14). For instance, if an association study provides evidence that a proposed cause gives rise to some variable that mediates the path between the proposed cause and effect, this provides evidence of mechanism between that proposed cause and effect.

1.5.3 Efficacy, effectiveness, and validity

Efficacy, effectiveness, and validity will be touched on throughout this thesis. In the most simple terms, efficacy is whether an intervention has its putative effect in a study population, where effectiveness is whether an intervention will have its effect in a ‘target population’ (Parkkinen et al., 2018, 5). We may conduct so-called efficacy and effectiveness trials. Efficacy trials are intended to measure the outcome of an intervention in laboratory settings, and effectiveness trials are intended to measure the outcome of an intervention in ‘real life’ settings (Ernst and Pittler, 2006). Efficacy and effectiveness trials are sometimes referred to as explanatory and pragmatic trials respectively (Wasan, 2014). This distinction can be contrasted with the external and internal validity distinction. Internal validity is whether the results of a trial indicate that an intervention had its effect in a trial (Cartwright, 2007). External validity is whether the results of a trial will apply outside of a trial; if a trial provides evidence that an intervention is effective in one situation, does that provide evidence it is effective in others (Cartwright and Munro, 2010)? So, internal validity asks, does a trial provide evidence that an intervention had its effects in that trial, and external validity asks if those effects would be seen in a group outside the trial.

Effectiveness can be thought of as the coincidence of efficacy and external validity, as Parkkinen et al. put it: ‘Typically, one establishes that a causal claim holds in a target population by establishing the claim in a study population and then extrapolating that claim to the target population’ (2018, 5). Because an efficacy trial is very strict, it may often have high internal validity, but be a poor match for real life settings, so have low external validity (Cartwright, 2007). Conversely, an effectiveness trial may have low internal validity, as it fails to control for many important, trial relevant factors, but have greater external validity than an efficacy trial because it more closely mirrors ‘real life’. Parts I and II of this thesis are primarily concerned with internal validity and efficacy; in practice, do RCTs in sports science provide strong evidence that observed outcomes can be attributed to interventions tested in those trials? I argue that very often, the answer is no. This contributes to the fact that relying on evidence from RCTs alone as a primary evidence gathering source does not provide the best possible evidence. Part III of this thesis is more concerned with using details of mechanisms to aid in exploring external validity questions than it is with questions of internal validity.

These distinctions can be a little fuzzy. For instance, simply knowing that one is in a trial can influence observed outcomes. Knowing that one is in a trial can help to ensure that one complies with a prescribed intervention more than would be the case outside a trial. So, even though effectiveness trials aim to measure intervention effects in ‘real life’ settings, one cannot discount the influence of trial effects on observed outcomes. One subtlety that arises here, and that is particularly important in Part III of this thesis, is adherence: whether people will comply with a prescribed intervention. After all, in a trial we can observe outcomes for an intervention when: A) it is complied with, or B), when it is prescribed. An intervention may have significant observed outcomes when compliance is assured, but seem to be ineffective when it is simply prescribed, for instance if it is difficult to comply with in virtue of it being time-consuming or overly physically demanding. A point I will make now, that I reiterate in Part III, is that even if an intervention would have its effect if it were complied with, if it is too difficult or convoluted to be adhered to, it is not a good intervention. This is one reason why measuring prescription without ensured adherence is useful, and is similar in some ways to *intention to treat analysis* (Soares and Carneiro, 2002) used in medicine, which includes those who drop out of the study in RCT results.

1.5.4 The quality of evidence

Regularly, throughout this thesis, I refer to evidence by its quality. I adopt the definition given by Parkkinen et al. (2018, 26) for what the quality of evidence is. I do this because, as I defend their views in chapter 5 relating to the quality of evidence, I should also use terms in the same way as they do.

The evidence that supports claims can be ranked according to its quality, that is, how good it is (Parkkinen et al., 2018, 25). This is not to be confused with a system that ranks evidence gathering methods. The quality of evidence, as put by Parkkinen et al. (2018, 26), is ascertained by determining how likely *in principle* it is that future evidence will impact our confidence in a causal claim, and by how much. This, then, ranks evidence quality by *stability*. Parkkinen et al. explain this using a scale that ranks evidence by quality level, which is informed by the GRADE Working Group (2004). The table they produce is reproduced in Table 1.1. If evidence quality is very low, it means that, in principle, future evidence could significantly change the confidence we have in a claim. High-quality evidence is the opposite. If evidence is high-quality, it means that future research is very unlikely to impact significantly the confidence we have in a claim.

To illustrate, let us imagine a case where all of our evidence, once evaluated, indicates that A is a cause of B. However, imagine that, in this

Quality level	Interpretation
High	Further research is highly unlikely to have a significant impact on our confidence in the claim.
Moderate	Further research is moderately unlikely to have a significant impact on our confidence in the claim.
Low	Further research is moderately likely to have a significant impact on our confidence in the claim.
Very Low	Further research is highly likely to have a significant impact on our confidence in the claim.

Table 1.1: Levels of evidence table, reproduced from Parkkinen et al., 2018 (26).

case, we know that all the evidence is flawed in some way. As this is the case, this evidence is insufficient to establish that A is the cause of B. We know that, in principle, it would be possible to have evidence in the future that does not have these flaws. This un-flawed evidence would, in principle, change how confident we are in the claim that A causes B, either positively or negatively. The fact that, in cases such as this, we *could* have evidence that is likely to change our confidence in a claim, means that the current evidence we have is either of low, or very low quality, according to this scale. This does not mean that in cases where we cannot get better evidence in practice, the evidence we currently have is of high quality. This scale works on the idea that evidence could, in principle, change our confidence in a claim. Parkkinen et al. provide a further, numerical, example that can help to explain this:

> Suppose current evidence warrants 75% confidence in a causal

> claim. One then learns that there is further evidence which warrants a 25% change in confidence, but one does not know the direction of this change. i.e., one does not know whether this new evidence warrants 50% confidence or 100% confidence. The 75% confidence is not sufficiently stable for the claim to be considered established or even provisionally established. This is because future evidence may be likely to decide between the 50 and 100% confidence, leading to a large change in confidence either way (2018, 26).

The implications of this definition for the **Excluded Explanations Argument** are discussed in section 2.2.

1.6 Conclusion to the introduction

In the introduction of this thesis, I outlined how this thesis will proceed. I also introduced and motivated the importance of key arguments of this thesis: that evidence from mechanistic studies in the sports sciences should be taken seriously, leading to the **Better Evidence Thesis**, that evidence primarily derived from RCTs does not provide the best possible evidence, and that providing details of mechanisms is useful in the sports sciences. I also introduced, discussed, and motivated, the move towards EBP in sports science. In the next two chapters, I will argue that evidence from RCTs and RCT-style N of 1 trials in sports science often do not provide high-quality evidence, motivating the idea that evidence from those sources alone does not provide the best possible evidence.

Part I

The problems

Introduction to Part I

As was seen in chapter 1, evidence from RCTs is often privileged in EBP. This is often at the expense of other types of study, such as mechanistic studies. Relying on evidence from RCTs is seen as a good way of fulfilling the EBP goal of relying on the best possible evidence. This can be seen, for instance, in the use of evidence hierarchies. Part I of this thesis argues that RCTs, and RCT-style N of 1 trials which are sometimes offered as a solution to problems facing groups RCTs in sports science, will often provide low-quality evidence. As will be argued in chapters 2 and 3, this is because of difficulties associated with sample sizes, placebo controlling, and blinding that, in many instances, cannot be avoided. This challenges the idea that evidence from these types of trials should be privileged in EBP, particularly at the expense of evidence from other types of trial. It also challenges the idea that relying on them counts as relying on the best possible evidence.

Chapter 2

Problems with sports science RCTs: why the evidence might not be that good

2.1 Introduction

In this chapter, I introduce the idea of the *ideal* RCT. As will be seen, the qualities an RCT has when it is close to ideal are what supports the idea that RCTs provide strong evidence. These qualities are intended to provide strong evidence for a causal claim by ruling out explanations for observed outcomes, other than the intervention or exposure being tested, ruling in the intervention or exposure being tested. In this chapter, I argue that, due to the nature of the sports sciences, it is often difficult, and sometimes impossible, for RCTs to adequately fulfil the conditions an RCT must fulfil in order to provide strong evidence. These conditions are: having an adequately large sample size, effective placebo controlling, and adequate blinding. I argue that, as a result of this, RCTs in the sports sciences may often not provide evidence that is as strong as the hierarchies suggest. I argue, even, that RCTs in sports science often produce low-quality evidence as a result of failing to rule out alternative explanations for observed outcomes. I argue that in many instances this is unavoidable due to the nature of the science. The argument I will give in favour of this claim, I will call the **Excluded Explanations Argument**. This brings into question the privileging of RCTs in EBP. This chapter sets up the main argument of Part II of the thesis, where I argue that evaluating evidence from mechanistic studies, in addition to evidence from RCTs, will often provide stronger evidence than evaluating evidence from RCTs alone.

It is important to note that in this chapter I am not making the argument that all RCTs provide less-than-strong evidence. In fact, in chapter 4, I argue that it *is* possible for evidence from association studies such as RCTs to establish causal claims, which of course requires strong evidence. This chapter, instead, sets up how difficult it is for RCTs in sports science to provide strong evidence in practice and indicates some areas where this will be the case particularly often. This means that the need to reform evidence evaluation processes in sports science may be greater than in medicine, where the EBM+ group have already championed the cause. Finally, I finish the chapter by considering the suggestion that we could address this problem by utilising evidence rating schemes, like GRADE, from medicine for the assessment of evidence in sports science. I argue that it may not be appropriate to rate the quality of evidence provided by sports science RCTs using a system like GRADE as are they are too conservative in their down-rating of RCTs.

2.2 Ideal RCTs

In this section, I discuss what an *ideal* RCT is. I also discuss the quality of evidence RCTs may produce. I do this with reference to the evidence quality system given in section 1.5. I discuss what features an RCT requires in order to be considered to be ideal. Also, I explain why the quality of evidence a non-ideal RCT provides us with is lower the further from ideal an RCT is, and higher the closer it is to ideal. The quality of evidence produced is based on how well RCTs can eliminate non-C explanations for trial outcomes. Seeing why the further an RCT is from ideal, the lower the quality of evidence that may be gleaned from it is important as it helps to explain why evidence produced by RCTs in the sports sciences may often unavoidably be less strong than is suggested by evidence hierarchies. This is a key premise in the **Excluded Explanations Argument**. This section is not intended to be a critique of RCT methodology, or whether close-to-ideal RCTs deserve their place at the top of evidence hierarchies. It is instead intended to set up a discussion about the quality of evidence RCTs provide when they are far from ideal.

RCTs are used to provide evidence for the claim that a proposed cause, C, has some putative effect, E, in the population being studied, P (Cartwright, 2007, 15). Whilst they reach different conclusions about the ability for RCTs to provide strong evidence for causal claims, Cartwright (2007, 2010), Stegenga (2014), and Worrall (2002, 2007) all explain that the quality of evidence an RCT provides for a causal claim hinges on, in part, its ability to rule out non-C explanations for observed outcomes. Certain features in RCTs are meant to help them in ruling out non-C explanations for observed outcomes, and ruling in C as a genuine explanation. These are: large sample sizes, adequate blinding, and adequate placebo controlling. It is the assumed occurrence of these features in practice, and their ability to eliminate non-C explanations for observed outcomes, that is meant to justify the claim that RCTs provide strong evidence for causal claims. It is this that is taken to justify the position of RCTs at the top of evidence hierarchies (Stegenga, 2014, 318).

A typical, but simplified, RCT would compare measured outcomes between two groups. In one of these groups, 1, an intervention C that is proposed to cause effect E would be present. This is sometimes called a trial or test group. In the other group, 2, C would not be present to the same extent as in group 1. In some situations it may be possible to completely eliminate the presence of C from 2, but not all. For instance, in an exercise trial it is unlikely that it would be possible to stop participants exercising at all, or in a vitamin trial to stop them having any of that vitamin in their diet. Instead of receiving the proposed cause C, in group 2, a placebo or active control is utilised. Placebos are discussed below, and in subsection 2.3.2. This group, 2, is often called the control group. Participants in the trial, and those conducting the trial, should ideally not know who has been assigned to trial or control groups. This is called *blinding*. Ideally, the only thing that may give rise to a difference in measured outcome in E between groups is the proposed cause under investigation, C. If there is a sufficiently different measured outcome in the trial than in the control group, we can be said to *observe* a correlation. Where this can be attributed to the presence of C, it provides evidence that C is a genuine cause of E. The quality of this evidence is dependent, in part, on the quality of the RCT.

Relevant, non-C differences between groups in trials that may give rise to E, or that could be used to explain differences in measured outcomes, are called confounders (Howick, 2011b, 34). If it is possible that confounders exist, the possibility of their existence can be invoked to explain at least part of the difference in measured outcomes between groups. Consider a trial investigating the effects of a particular training programme on 100 m running times. If all those in the test wing of the trial are untrained, and all those in the control wing are highly trained, we may reasonably expect that those in the test wing of the trial will see far greater improvements in 100 m time than the highly trained athletes. This is because it is far easier for an untrained athlete to improve their 100 m time by a large margin than a highly trained one. As, in this imaginary trial, how well-trained an athlete is originally can provide an alternative explanation for differences in observed outcomes, it is a confounder.

An ideal RCT is one where, by its design and application, it is able to rule out all non-C explanations for observed outcomes in a trial. In other words, in an ideal RCT, the only net difference between groups that should have some effect on outcome E in the trial should be C.

Unfortunately, we cannot conduct a perfect, ideal RCT. For instance, no RCT can ever have an infinitely large sample size, which would be necessary to totally rule out chance as an explanation of observed outcomes. This can be explained with a coin. A fair coin that is equally likely to land on heads as it is on tails may be flipped 100 times and appear to favour heads by chance. It is not until the number of flips nears infinity that we can be sure that the measured outcomes will stabilise accurately. Just as this is the case, until we have an RCT with an infinitely large sample size, we cannot be sure that the outcome measure of the trial is in no way affected by the chance distribution of confounders. However, just because no RCT can be truly ideal and rule out all confounders does not mean that they cannot provide strong evidence. When the criteria mentioned above are adequately met, even non-ideal RCTs are often seen as good tools to help rule out alternative explanations for differences in measured outcomes between groups. As such, they *can* provide strong evidence in theory. But, as it is their ability to help rule out non-C explanations, and rule in C as a genuine explanation for observed outcomes, the strength of evidence an RCT can provide is conditional on how well the features of an RCT allow this to happen.

A number of theories have been put forward to explain how RCTs can provide evidence for causation. For instance, on Cartwright's view, evidence from RCTs can be used to establish the existence of a correlation between cause and effect. Cartwright claims that, for an RCT:

> If the probability of an 'outcome' O is greater with a putative cause T than without T once all 'confounders' are controlled for in some particular way, that is sufficient for the claim 'T causes O' in that particular setting of confounding factors.

> (2007, 15)

On her view, RCTs can establish causation for the trial population by ruling out confounders and showing that an outcome is more likely in the presence of the proposed cause. For Cartwright, establishing the existence of a special type of correlation, one that is conditional on all confounders, is sufficient to establish the existence of a causal relationship. In chapter 4, I will argue that RCTs, in addition to being able to establish the existence of a correlation, can sometimes suffice to establish the existence of a mechanism linking cause and effect by ruling in the existence of a mechanism.

A key point to note about RCTs is that, regardless of what theory of causality one holds, the less strong the evidence an RCT produces is, the less well it contributes towards establishing causality. For instance, Cartwright explains RCTs in terms of the probabilistic theory of causality, in which establishing causality requires establishing the existence of a correlation. It is easy to see how an RCT may fail to establish causation in this instance by failing to rule out alternative explanations for difference in effect sizes. Whilst being far from the only person to advance a probabilistic theory of causality, (see for instance: Reichenbach (1978), Good (1959), and Suppes (1973)), I offer this explanation in Cartwright's terms as she offers one of the clearest and most rigorous philosophical, probabilistic, overviews of RCTs. One may alternatively subscribe to a mechanistic theory of causality, which requires the existence of a mechanism to be established in order to establish causality. Whilst it is not as clear here why RCTs would be useful to proponents of a mechanistic view of causality, I argue in chapter 4, that evidence from RCTs may, in some cases, be sufficient to infer the existence of a mechanism by ruling out any explanations for observed outcomes other than that a mechanism must exist. Proponents of a mechanistic theory of causality include: Salmon (1998), Dowe (2000), and Machamer, Darden, and Craver (2000). Alternatively, one may subscribe to the epistemic theory of causality, such as that advocated by Russo and Williamson (2007), which requires that both the existence of a suitable correlation and mechanism be established in order to establish causality. In all of these instances, an RCT provides evidence for *something*. Broadly, regardless of which theory one takes to underpin establishing causation, RCTs can provide some level of evidence for this by ruling out confounders, and having a sufficiently different value in outcome measure in control and trial groups to rule in a specific cause. The key point here is that, whatever one's theory of causality requires in order for causality to be established, the less strong the evidence from an RCT is, the less well it provides evidence for that thing. For this reason, I stay relatively agnostic towards all theories of causality in this chapter, so that one need not buy a particular theory of causality in order to buy the argument of this chapter. In chapters 4 and 5, I will argue in favour of the application of the Russo-Williamson thesis for sports science, which contends that in order to establish causation we need to establish the existence of both a mechanism and a correlation. This has no bearing, however, on the argument of this chapter.

So, whilst eliminating possible non-C explanations for differences in measured outcome between groups ensures that an RCT can provide good evidence, not doing so sufficiently will limit the quality of evidence an RCT can provide (Worrall, 2002, 318, Wootton, 2007, 485). Why is this? The less well confounders are controlled for, the greater the probability that C is not the cause of observed outcomes, and the lower our confidence that C caused E should be. As Cartwright puts it, in the case of RCTs: 'the probability of the conclusion can be no higher than that of the weakest premise' (2007, 14). An observed correlation may have a few explanations other than that C gives rise to a causal relationship which explains the correlation (Williamson, 2021, section 3). For instance, there is, according to Clarke et al. (2014, 343), a 'realistic chance of stumbling across coincidental correlations in RCTs'. Association studies are also subject to other problems, such as where what is assumed to be causal and what is assumed to be caused, may actually have common causes (Clarke et al., 2013, 745). Williamson (2019, 38), identifies two problems that occur when assessing data, such as that data which may arise from an RCT. Problem one is falsely inferring the existence of a genuine correlation, and problem two is falsely inferring that C is a genuine cause of E from an observed correlation. The failure to eliminate alternative explanations, other than that C causes E, which can explain observed correlations in RCT data, illustrates why evidence from RCTs alone may be insufficient to establish that a correlation is causal.

I will illustrate this with a hypothetical example. We may observe a correlation in an RCT, for instance, between an intervention C and its putative effect E, where we cannot exclude a non-C possible explanation of the observed correlation. This means that we will not be able to determine if the observed correlation is a genuine result of C causing E. It may be hard to exclude, particularly in small trials, the explanation that the pre-trial treatment randomisation accidentally sorted the trial groups such that those predisposed to E were all in the treatment wing of a trial. This may happen where we do not know all indicators that one is predisposed to E, and are unable to exclude those people from registering for the trial. RCTs like this, whilst they may reduce the probability that alternate explanations may be given for observed outcomes, are unlikely to be able to rule them out sufficiently well to establish causal claims, unless they meet the criteria proposed previously: adequate placebo controlling, adequate blinding, and an adequate sample size. This illustrates how evidence from RCTs, and other types of association study, alone, may fail to rule in C as the *only* possible cause of an observed correlation with E, providing low-quality evidence that C is a cause of E.

Thus, if the effects of confounders on E cannot be ruled out with a high degree of certainty, what is being measured in an RCT may not only be whether, or by how much, C makes a difference to E. This is a key point in the **Excluded Explanations Argument**. An important note here is that very high quality RCTs that have good levels of blinding, good placebos, and large sample sizes, may still provide strong evidence for causal claims. This is because, in virtue of this, they may rule out confounders to a high degree.

2.2.1 Features that help an RCT rule out confounding, bias, and chance

As can be seen from the name, randomization is a key part of RCTs. By randomizing participant assignment to trial groups, it is assumed that with sufficiently large sample sizes, the influence of prognostic factors on measured outcomes will be balanced between groups. Prognostic factors, a type of confounder, are factors that influence the course of a disease. These can include things such as age, sex, and number of comorbidities. In sports science, unless it focuses on disease, randomization is intended to balance relevant factors across the trial arms that could influence measured outcomes in a trial, these can include things like age, sex, and injuries or illnesses. Randomization can also be *restricted*. Restricted randomization ensures that some factors are evenly distributed between groups before randomization occurs (Howick, 2011b, 184-185). Randomization is also intended to give both trial and control patients the same expectation of recovery (ignoring intervention effects), which is important given that the expectation of recovery can aid recovery (Urbach, 1993, 1422).

Some, e.g. Papineau (1994), hold that randomization is important in investigating causal claims. Papineau claims that the randomization of treatment allocation helps us draw causal conclusions from trials (1994, 440). Papineau claims that once we have randomized, balancing confounding factors, the inference 'from probabilities to causes, is quite infallible, in virtue of the randomization of the treatment in the experiment at hand' (1994, 447). He calls randomization a 'sure-fire guide to causal conclusions' (1994, 447).

Opposed to this, commentators such as Worrall hold that randomization holds no special power in detecting causes (2002, 2007). Worrall holds that the biases which we would use randomization to eliminate are not eliminated by some special power held by randomization (2007, 466). Worrall goes on to say that the supposed importance of randomization as a method of eliminating bias shows that eliminating bias is important in finding causes, not that randomization itself is what is important. In fact, or so says Worrall, we may eliminate bias in a number of ways and, as such, randomization should not be given the epistemic privilege afforded it in many instances.

I will not provide an argument in favour of, or against, randomization. It will be seen in this chapter that whether randomization plays an important role in discovering causes is unimportant to the conclusions I draw. Further, because of the deficiencies faced by RCTs in the sports sciences, particularly those as a result of having small sample sizes, randomization will often be unlikely to be able to balance confounders equally across groups. This means that even if randomization is an important and useful part of finding causes, it may not have its desired effect in many sports science trials.

As well as by using randomization, the three features mentioned above help to rule out confounders:

- Blinding - A trial is blinded if participants do not know whether they are assigned to the test or control group. A trial is double blinded if neither participants, nor those administering the experiment, know who is assigned to which group. Blinding 'helps ensure that no differences slip in between treatment and control wings due to differences in attitudes, expectations or hopes of anyone involved in the process.' (Cartwright, 2010, 63)

- Placebo controls - These are given to those in the trial groups so that they are less likely to realise they are not receiving the intervention under investigation. They are ideally indistinguishable from the intervention under investigation. They are 'supposed to ensure that any 'psychological' effects produced by the recognition that a subject is receiving the treatment will be the same in both wings' (Cartwright, 2010, 64). They help blind the trial, but can also be used to isolate the proposed *characteristic* feature of the intervention under investigation.

- Large sample size - This is how many participants, or things from which data is sampled, are in a trial. This is to rule out chance as an explanation of differences in measured outcome (Williamson, 2019, 37-38), but as will be seen in the next section, is also necessary to distinguish smaller effect sizes.

All of these factors, and how not implementing them sufficiently in an RCT can impact the quality of evidence produced, will be discussed in detail in section 2.3.

Now we have some key elements of the **Excluded Explanations Argument**.

- An RCT is meant to provide high-quality evidence on the grounds that it rules out alternate explanations for trial outcomes, such as bias, confounding, and chance.

- If things other than the intervention or exposure being tested can explain the outcome of a trial, we cannot rule in the intervention or exposure being tested as the only explanation for observed outcomes.

- The less well RCTs rule out alternate explanations, such as bias, confounding, and chance, for differences in observed outcomes, the

> lower the quality of evidence in favour of the causal claim those RCTs produce. This is, in part, because we have less confidence in our ability to rule *in* the intervention or exposure being tested as an explanation of those differences.

These premises, along with the arguments made in the bulk of this chapter, that RCTs in the sports sciences often do not meet the requirements needed to rule out bias, confounding, and chance, make up the **Excluded Explanations Argument**, the conclusion of which being that RCTs in the sports sciences often do not produce high-quality evidence, often unavoidably. This can be seen in light of how we may define the quality of evidence given in section 1.5. It was raised that low-quality evidence for a claim is evidence where there is low confidence in a claim, and, in principle, future evidence is likely to change our confidence in that claim. Where an RCT produces evidence that does not rule out confounders, we will have low confidence in a causal claim based on that evidence, and it is likely that future evidence could, in principle, change our confidence in the claim supported on the basis of these far-from-ideal RCTs. So, where far-from-ideal RCTs may give us low confidence in a causal claim, the possibility that, in principle, high-quality evidence *could* be gathered that rules out confounders and rules in an intervention as causal to a high degree tells us that where RCTs cannot rule out confounders, the evidence they produce is of low quality.

One may wonder, given that, as I shall argue, sports science RCTs are likely to be far from ideal, how we can tell how strong the evidence an RCT, or set of RCTs, can provide for a causal claim is. What if high quality RCTs disagree? I do not want to provide a definite answer here. However, the obvious thing to bring up at this point is that evidence rating schemes which account for the varying quality of evidence provided by RCTs, and allow us to rank them according to how well they fulfil certain criteria, already exist. The GRADE hierarchy is one such example. By allowing experts to rate the quality of evidence produced by individual RCTs based on its susceptibility to bias and confounding, this puts the onus of determining if causality is established on experts. Further to this, many methods of conducting systematic reviews and meta-analyses of the evidence for a claim exist. These can be used by experts to determine how much evidential justification causal claims have. I argue later (in section 2.4), however, that these hierarchies, as they are designed for medicine where RCTs are likely of higher quality than sports science, may still be insufficient to rank the quality of evidence in sports science. This is left until later in the chapter as it is necessary to first highlight the severity of the potential limitations to evidence in some sports science RCTs. Only then will it be clear how GRADE, as it is designed with medicine in mind, may be insufficient as it currently stands, to evaluate evidence for causal claims in sports science. However, something like the GRADE system of evidence evaluation may be useful in answering this question.

2.3 Limitations in sports science RCTs

In this section, I will explain why not having sufficient blinding, placebo controlling, and sample size, can introduce bias, confounding, and chance, as potential explanations for trial outcomes. I will also argue that these inadequacies are present relatively often in RCTs in the sports sciences, and that they may often be unavoidable, due to the nature of the research. The aim of this is to help provide an argument that often, due to the nature of the sports sciences, evidence provided by RCTs alone is less-than-strong, and is relatively often insufficient to establish the existence of a causal relationship. This leads to the claim that, by relying on evidence from RCTs, EBP is encouraging reliance on evidence that, in the real world, is not good evidence at all. Later, this will be taken with claims from chapter 4 and chapter 5 to argue in favour of the **Better Evidence Thesis**.

2.3.1 Sample sizes

Within the sports sciences, it is difficult to recruit and maintain study participant numbers for trials. As a result, sample sizes tend to be small. This is regularly commented on in sports science literature, and reviews of sports science as a field; see for example: Select Committee on Science and Technology (2012b, 11-12), Nevill et al. (2008, 422), and Schweizer and Furley (2016, 114). In this section, I explain how small sample sizes reduce the quality of evidence an RCT can produce. Also, I argue that this is the case for sufficiently large quantities of sports science research that it calls into question the assumption made by evidence hierarchies that sports science RCTs provide strong evidence for causal claims. The prevalence of, and problems caused by, small sample sizes in the sports sciences can be highlighted with a few cases:

- In a systematic review of intervention studies in the sports sciences that considered all 'articles providing information on the recruitment of adults into interventions involving sport and reporting physical activity or participation outcomes', Cooke and Jones found that, of those studies examined, only half reached their recruitment goals (2017, 1)

- Many studies, even those presented at international conferences for instance, have small sample sizes (Pyne et al., 2010, 4). These are often too small to reasonably infer effects from – some include sample sizes of less than 10 (Pyne et al., 2010, 4).

- An analysis of four leading sports science journals also found that, of all studies published between 2009 and 2013, in a large proportion of these experimental studies, sample sizes were too small to detect 'small-to-medium effects' (Schweizer and Furley, 2016, 114).

- In a review of sports medicine studies, all RCTs of orthopaedic and arthroscopic surgery meeting simple inclusion criteria published between January 2005 and October 2015 were reviewed. In most RCTs reviewed, it was possible to reverse statistical significance by changing the outcome of only a few patients (Khan et al., 2017, 2164), meaning that the result of the trial may be explained by chance. How representative this is of RCTs in sports science in general could be questioned given that these studies are in sports medicine rather than sports science but, if anything, this highlights how difficult it is to achieve large sample sizes in much of sports related research.

- Whilst it does not concern internal validity and the quality of evidence produced by RCTs, so long as participants are evenly balanced between arms of a trial, it is worth mentioning here that there is often an uneven distribution of ages, genders, socio-economic status, disability, and education, represented in trials (Cooke and Jones, 2017, 1). In other words, many sports science trials are engaged in by only specific types of person. A systematic review found that sports science trial participation was biased towards affluent, middle-aged, white women (Cooke and Jones, 2017). This means that, as well as trials having few participants, the results of these trials may lack good external validity. This means that where good quality trials are conducted, there may not be many people in the sporting world for whom the results are relevant.

I contend that there are three types of sample size limitations in the sports sciences. First, there are trials where the sample size is unavoidably small and where this can not be changed. For instance, trials investigating elite athletes will necessarily have small sample sizes due to the tiny number of these athletes that exist, and that are willing to be in trials (Select Committee on Science and Technology, 2012b, 11-12, Buchheit et al., 2019, 1327). This will also include trials involving those with uncommon injuries or rehabilitation needs, and also trials on niche sports. These cases, particularly that of elite athletes, demonstrate that large portions of sports science research must necessarily be conducted with small sample sizes. The second type of limitation to trial size is where the limitation is a methodological limitation that is difficult to avoid, but which could be overcome in some instances. This includes limitations such as those due to the ability to widely advertise a trial. It also includes being able to find participants who are willing to adhere to trial interventions and placebo instructions. These are methodological in nature because they could *technically* be overcome, unlike limitations due to the small number of elite athletes. The final category is a cross-section between the two previous ones. These are limitations that are methodological, but which are almost impossible to avoid in the current sports science climate. These may sometimes intersect with those in the previous category, depending on the type of research being conducted. For instance, for some sporting populations it may currently be almost impossible to convince large samples of people to adhere to an intervention, where, for others, it is simply difficult. For instance, it may be easy to convince large numbers of youth athletes to engage in an intervention as one can encourage coaches to make it a mandatory part of training. Conversely, it may be hard to encourage adults to perform the same intervention as they may disagree with their coach on the importance of performing an intervention where a child would likely not. This final category also includes limitations such as funding. In a different sports science climate this limitation may be overcome, but it currently often cannot be because research budgets in sports science are generally small. They are particularly small when compared to those in medicine, where pharmaceutical companies stand to make lots of money with new interventions.

These instances mentioned above highlight how common it is for trials in the sports sciences to have small sample sizes. These cases also highlight how small sample sizes can affect trial outcomes. First, small sample sizes increase the likelihood that chance is a potential explanation for observed results. This is particularly noticeable where, by changing the results of a few patients, we can reverse the statistical significance of trials. Importantly, with small sample sizes it is very difficult to balance confounders across trial groups, meaning that one group may be more disposed to some outcome than the other. Second, small sample sizes mean that small to medium effect sizes may not be observed. This can, unfortunately, lead to studies being published that purport to show small to medium effect sizes through an abuse or misunderstanding of statistics, but which actually do not reach statistical significance. In their systematic review previously mentioned, Schweizer and Furley (2016) show that the reporting of false positives in Sports Psychology may exceed 50%, just as Simmons et al., (2011) argues is the case in Psychology in general, part of the argument that lead to the replication crisis (for more on the replication crisis see: Shrout and Rodgers, 2018). Knudson (2009, 97) also argues that Sports Biomechanics suffers from an 'inappropriate use of statistical analyses' and reporting of results. This means studies which report that they provide evidence against the null hypothesis, that an intervention has no effect, may often be too small to have actually done that. As an example of the abuse of statistical methods, it is not uncommon for researchers to report that a trial 'approached statistical significance' as if this provides evidence against the null hypothesis, and for a causal relationship. In one specific instance, an investigation into exercise addiction in sports science students (Szabo and Griffiths, 2007, 25), reported that as their study has a p-value of 0.09, it 'approached statistical significance' and that the 'findings raise the possibility that sports science students may be more susceptible to some components of exercise addiction than exercisers more generally'. This is, of course, a misrepresentation of findings. Something does not approach significance, it is simply not significant and is therefore insufficient to provide evidence against the null hypothesis. Thus, when a trial has a small sample size, the possible explanations for observed effects and reported results can include chance, and improper use of statistics, weakening any causal claim the trial purports to provide evidence for.

An aside on P-values

To properly understand why this is an abuse of statistics, I will need to briefly explain how p-values work. A p-value statistically measures the chance that the measured outcome would have happened if an intervention is ineffective. This is often set at a p-value of 0.05 (Stovitz et al., 2017). The results of a trial are said to be statistically significant where a p-value indicates that it is highly unlikely that the measured outcome would have occurred if the treatment was ineffective. This means that a low p-value indicates that if the null hypothesis is true, the results observed are highly unlikely. In an intervention study, the null hypothesis is that an intervention has no effect. Thus, where a p-value is very low, the observed outcome would have been highly unlikely if the intervention was ineffective, which can be taken as some indication that the results infer that the intervention had some effect. It is also important to note that p-values do not represent the magnitude of an effect or outcome, so a very low p-value does not mean that an intervention is particularly effective. So, the study into exercise addiction previously mentioned (Szabo and Griffiths, 2007) did not produce results that raise the possibility that sports science students suffer from exercise addiction more than other students because the p-value indicates that there is a likelihood that the results observed could have been observed if sports science students did not experience exercise addiction more regularly. The problem here is misrepresenting statistics in such a way that they appear to evidence an outcome which they do not. In addition to this one example given, the problem of so called p-hacking in sports science is considered to be so large that The Society for Transparency, Openness, and Replication in Kinesiology (STORK) has been founded to combat it (Society for Transparency, Openness, and Replication in Kinesiology, 2020). P-hacking is the manipulation of the way results are analysed until they appear to be statistically significant. For instance one can choose to report only outcome measures that show significant results. This is, in part, so that publication can be easier and research appears to be more important in its findings. STORK has researchers commit to a certain research methodology and plan of analysis before conducting their research, so they cannot change how they report and analyse their results after the fact to force statistical significance.

Limitations to the quality of evidence provided by RCTs due to low sample sizes can also be seen in some areas of medicine. We can draw lessons for sports science from this. In cases where experiments are used to determine the effectiveness of some treatment on rare diseases, the fact that those diseases are rare means that sample sizes are necessarily small. In medicine, it is accepted that not having adequate sample sizes to conduct high quality RCTs can limit evidence quality in the case of rare diseases (Behera et al., 2007, Lilford et al., 1995). In cases where sample sizes are small, p-values will not be significant unless the treatment effect is large. The difficulty of obtaining adequate sample sizes in these instances is exemplified in a paper by Gallin et al., who took ten years to reach a sample size of 39 to test the use of itraconazole for severe fungal infection in patients with chronic granulomatous disease (2003). In instances where samples sizes will not reach sufficient numbers, it is taken that the evidence they can provide is poor and, as such, they will often not be conducted (Behera et al., 2007). For rare diseases like this, treatments can be employed in practice without an RCT being needed to inform that decision (Behera et al., 2007, 163). This is because it is often better to do something than do nothing, but the efficacy of that treatment will be understood as unjustified. What this goes to show is that in medicine, where RCTs are typically of a higher quality than the sports sciences, if an RCT will unavoidably have a small sample size, it will either not be conducted, or the quality of evidence it provides will be considered to be poor. From this we can infer that the same problems with evidence from RCTs will be present in the sports sciences, and that small sample sizes should be taken seriously in sports science, just as they are in medicine.

It is interesting to note that a sample size as low as 39 to investigate an intervention for a disease is considered rare in medicine, where, as Pyne tells us (2010, 4), research on fewer than 10 participants is presented at international conferences in the sports sciences.

If mechanisms that lead to small sample sizes can be uncovered, it would help to explain if they are avoidable. However, it is difficult to fully explain the mechanisms that lead to poor study recruitment and to determine if they are unavoidable as the reporting of these mechanisms is very often poor, as is explained by Cooke and Jones, in an investigation of sports science recruiting and sample sizes (2017, 1). However, I will try to give the argument some weight by offering some reasonable intuitions one may begin an explanation with. All the following are reasons why a trial may only be able to recruit a few participants:

- The type of person who wants to be involved in a sports science trial, particularly given that they often do not provide financial remunerations, is likely an athlete or sportsperson. A trial on cycling will likely attract cyclists, for example. And then, not all cyclists, only those who are also interested in being part of a trial. This reduces potential sample sizes.
- Studies in particular populations, such as 'highly-trained' or 'well-trained' can only recruit from small pools of people (even if those pools are bigger than the number of elite athletes).
- In an RCT, there is a likelihood that any participant will be put in a control group with a placebo intervention or active control. Athletes who are concerned with performance will likely drop out of a trial, or engage in different interventions than those being tested in a trial if they think they are in the control wing of a trial (just as is the case in medicine).
- Athletes may not even sign up for a trial as there is a risk that their placement in any trial group may impact their performance.

For instance, if a control group is asked to be sedentary or perform minimal exercise, or if a trial intervention is potentially ineffective. If a sportsperson has spent years training to reach a certain level of performance, it is unlikely they will want to risk losing potential improvements in the time the trial takes, or de-training in the course of a trial.

- There is considerable money in medical trials, which can be used to help recruit participants and encourage trial adherence and participation. Even studies conducted at universities may have large corporate sponsors. This is not the case in the vast majority of sports science research. A difference in funding can help explain the difference in expected participant numbers in sports science and medicine trials, but the sheer lack of funding, particularly outside of elite sports science research, helps to explain the small sample sizes observed.[1]

- Trials can be very long and intensive, it is not uncommon for exercise trials to exceed 8 to 16 weeks, and that is a lot of commitment for participants who are often volunteers.

- To further the comparison to medicine, the percentage of a sample that effectively adheres to control or trial treatments may be lower in sports science than in medicine, as the barrier to effective completion of a prescribed intervention may be higher in sports science. For instance, adherence to a medical trial may be taking a pill a number of times per day, which is relatively easy when compared to a sports science trial which may involve exercising for multiple hours a week. Even those committed to exercise regimes before enrolling in a trial may struggle to adhere, as daily life can impact the ability to adhere to habitual exercise, let alone trial mandated exercise. For instance, in an examination of a football injury pre-

[1] I have been engaged in a number of sports science trials, and I have never once received any type of incentive or remuneration.

vention programme in adult men, it was found that very few trial participants actually adhered to the intervention, so it was not possible to determine if the intervention would have been effective had they adhered properly (Hammes et al., 2015).

What we can take from this is that, in general, a certain type of person will be interested in being involved in trials in sports science, which limits potential participants. Further, as trials are often conducted on certain populations, such as 'highly trained', as well as reducing the number of potential participants by selecting from a smaller portion of the population, some people from that population may also be unwilling to be involved if it could risk their performance. When explained in this way, it becomes apparent that RCTs in sports science may have small sample sizes unavoidably because either there are not many people to recruit from, or the types of people who could be recruited simply will not want to be part of a trial. This problem is unavoidable because, in many instances, sports science researchers do not have the resources to change the disposition of potential participants, for instance through financial remuneration, in the way that medicine does.

There are of course counter-examples where sports science trials have large sample sizes. For instance, in a later chapter, chapter 6, I discuss RCTs on a football injury prevention programme, the primary RCT of which included 1892 athletes (Soligard et al., 2008). However, the existence of these counter-examples does nothing to diminish the fact that some areas of sports science will necessarily conduct research on small sample sizes, be that forever, or in the current sports science climate. In fact, the football RCT I have just mentioned, which appears to have a large sample size, actually undershot the recruitment goal they suggested was necessary to see effect sizes by around 250 (Soligard et al., 2008).

In this section, I have provided a number of sources that demonstrate sample size problems in sports science. I have argued for three categories of sample size limit. I have also provided rationale that explains why sample sizes are often small, and have explained why this lowers evidence quality. What I intend to be taken from this section is that in the sports sciences, RCTs in some areas will unavoidably produce low-quality evidence because they cannot reach sufficiently large sample sizes to rule out chance and balance confounders. Further, they may also, unavoidably, be too small to see small to medium effect sizes. I have also demonstrated that, even though the problem of small sample sizes may be avoided in some instances, the prevalence of studies which unavoidably have small sample sizes raises questions about the ranking of RCTs in evidence hierarchies as providing strong evidence.

2.3.2 The difficulty of adequately placebo controlling

I will now argue that it is difficult and often impossible to adequately placebo control in the sports sciences, lending weight to the **Excluded Explanations Argument** by giving another reason why RCTs can fail to rule out alternate explanations. Placebo controls are meant to help to blind participants in trials and help reduce psychological confounding effects (Maddocks et al., 2016, 598). In addition to this, an adequate placebo control is essential to correctly estimating effect sizes, as will be explained. The intention of this is to help eliminate things such as guessing what trial group one is in as potential explanations for observed trial outcomes. In this section, I present an argument by Maddocks et al. that it can be difficult, and often impossible, to placebo control exercise trials. I, then, argue that this argument holds more widely for sports science. I do this in part by providing characteristic examples of sports science interventions and demonstrating how difficult they would be to placebo control. I also argue that potential solutions to the problem of placebo controlling trials in sports science: viewing treatments holistically, active controlling trials, and dose response trials, may also not be adequate solutions to the problem in many instances.

Many have attempted to define what it is to placebo control (see for example: Grünbaum, 1981; Grünbaum, 1986; Holman, 2015; Moerman, 1983; Shapiro and Morris, 1978 and Evans 2003), however, none of these attempts is universally accepted. Despite the absence of a universally accepted definition of placebo, the widely accepted adequacy criteria for a placebo, which Maddocks et al. tout as being philosophically, 'the most operationally useful conceptualization of placebo' are that it must (Grünbaum, 1981):

- Have no features of the treatment it is being compared to that may cause recovery, or must have none of the features proposed to cause recovery that are under investigation. These are called the characteristic features.
- Have every feature that is in the true treatment being tested, but that would not cause a recovery, or that is not under investigation. These are called the incidental features.
- Have no more features.

Clearly, in sports science, we are not only concerned with recovery as in from an illness or injury. As this is the case, in some instances we can interchange recovery for whatever outcome we are interested in. I will adopt this definition on the grounds that it is widely accepted, and captures the key elements of what a placebo is that are necessary for this thesis. Whether the definition of placebo given here is entirely correct is ultimately not what is at stake in this section. What is really important, as will become clear, is that in many instances in sports science it is difficult or impossible to provide something against which we can compare the outcomes of an intervention under investigation to, whilst maintaining the assumption that it has little to no unintended effect on outcome measures.

Jeremy Howick (2011b, 84), furnishes us with a clear medical example that illustrates the different types of features important in a placebo. In a medical trial to measure the effectiveness of fluoxetine hydrochloride on depression, those in the test wing of the trial will have not only a pill, but also consultations with doctors, and other factors that may influence the result of the trial as part of their treatment. In order to attempt to isolate the effects of fluoxetine hydrochloride alone, the placebo wing of the trial needs to be otherwise indistinguishable from the trial wing. Otherwise, differences between trial wings could be reasonably given as explanations for differences in observed outcomes. This involves including consultations with doctors, and the ingestion of a pill that is indistinguishable from the fluoxetine hydrochloride pill in every way, except the inclusion of that active ingredient in the placebo wing of the trial. Some trials even make sure that the placebo pill being taken has similar side effects to the real treatment. As can be seen, in this example all the incidental features not being tested should be present in the placebo wing of the trial, whilst none of the characteristic features are.

It is of particular concern to trials on things such as physical therapy or exercise, common in sports science, that adequate placebos are hard or impossible to produce (Maddocks et al., 2016). In order to demonstrate this, Maddocks et al. ask us to try and imagine how we could come up with a placebo for exercise (2016, 598). Jeremy Howick (2011b), gives an inexhaustive list of characteristic and non-characteristic features of an exercise programme that must be considered when a placebo is being developed. This list includes (Howick, 2011b, 88):

1. belief that one is being treated with exercise;
2. participant/investigator (fitness trainer or advisor) interaction;
3. other "psychological" benefits of exercise (distraction from daily routine and worry, the sense of achievement and social interactions);

4. increased metabolic rate;
5. increased body temperature;
6. increased heart rate for some prolonged period of time;
7. increased endorphin and epinephrine/adrenaline levels caused by exercise

An example will help make things clearer here. Let us try to imagine a trial to test the effects of daily bike riding on blood pressure. In order to adequately placebo control this trial, in the way given by the Grünbaum criteria, placebo participants would need to undergo all incidental features of daily bike riding, whilst not undergoing characteristic features. Incidental features include getting out of breath and sweaty, a feeling of tiredness after bike riding, increased appetite, and also the feeling of riding a bike. All of this must be achieved without participants riding a bike and improving their fitness in a way that could actually have an effect on their blood pressure, the characteristic feature. This example helps to illustrate that it may be difficult or impossible to adequately placebo control some types of trial in the sports sciences. Without being able to adequately placebo control a trial, it is difficult to isolate the potential cause under investigation as the only explanation for observed differences in measured outcomes between groups. For instance, what is used as the placebo in a trial may itself influence the result of the trial. Interestingly, in medicine, this was the case in a trial that examined the effects of an intervention on cholesterol. In one, now famous exemplar cholesterol trial, the placebo control was olive oil. Olive oil was later found to also actively reduce cholesterol, meaning that the trial possibly underestimated treatment effects (Howick, 2009, 36). So, in some cases, it is difficult to placebo control in a way that isolates the proposed cause under investigation as the reason for observed outcomes. Where RCTs do not adequately placebo control, they miss criteria that justifies them as being able to provide strong evidence.

Whilst exercise trials make up a large portion of sports science research, and placebo controlling those types of trials alone is highly difficult, not all sports science trials involve exercise. Nutrition, for instance, is a key part of sports science research. In some instances, this will also be difficult to placebo control. Small changes to the diet may be relatively easy to control, for instance the ingestion of sugar pills rather than vitamin pills, but entire diet overhauls may be difficult to placebo control. For instance, imagine a trial to investigate the rate of fat to muscle loss during weight loss on different diets. This type of investigation may be useful for many sports where categories are organised by weight, like boxing. It would likely be important, in this type of trial, to have a control group that eats the same type of diet as weight loss groups, but who do not lose weight to see if that diet, without weight loss, has an effect on fat to muscle ratios. This would be important to see if, outside of weight loss, the diet had any effects on body composition. There are, of course, certain things a participant may expect in a weight loss trial that cannot be replicated in a placebo group who eat the same diet but maintain the same weight to observe fat and muscle changes in those groups. For instance, one would expect, among other things, hunger, to physically eat less than one is accustomed to, and a lower amount of energy for exercise sessions.

It is important to see how inadequate placebo controlling can have an effect on the observed results of trials. Inadequate placebo controlling may influence trial results where there are characteristic features of the treatment in the placebo, and where there are too few incidental features of the treatment in the placebo. If the placebo used to control a trial has characteristic features of the trial treatment that is being measured, it may lead to underestimating the effects of the actual treatment (Maddocks et al., 2016, 599). Conversely to this, using an inadequate placebo with few or none of the incidental features of an intervention being trialled may lead to an overestimation of treatment effects (Maddocks et al., 2016, 599). Consider the bike riding example again. If the exercise component of bike riding is seen as the characteristic feature of bike riding, one may try to placebo control this type of trial by asking the trial participants to not perform any exercise that they would not otherwise perform. Those in the placebo group may even be asked to be sedentary. However, by placebo controlling in these ways, the placebo group lacks some important incidental features that the trial group is exposed to. For instance, bike riding is often done outdoors in nature, and if placebo participants are sedentary, they will likely be indoors more often. If time spent outdoors has an effect on blood pressure, for instance by reducing stress, this then introduces time spent outdoors as part of the possible explanation for observed outcomes. It also fails to isolate the exercise component of bike riding specifically as being the only explanation for observed trial outcomes. From this we can see how placebo controls with too few incidental features can lead to overestimating treatment effects. We can also easily extrapolate from the imaginary cycling trial to other, similar, types of trial to see how problematic placebos can be.

Many interventions in the sports sciences will be very complex and have multivarious factors that would need to be accounted for in placebo controls. Interventions may have social, biological, and physical elements that go beyond medical pill taking in complexity. As a result, I have argued, it can be difficult or impossible to create a placebo that adequately rules out confounders. A solution to this could be to view treatments holistically. This idea is similar to that seen in the practice of conducting *pragmatic trials*. These are trials that intend to investigate the effects of a treatment in a real-world setting rather than in rigidly controlled idealised settings (Patsopoulos, 2011). Whilst pragmatic trials are intended to increase the external validity of trial findings by conducting experiments in similar to real life situations, placebo controlling may also be easier. Placebo controlling would then be done based on a wide conception of what the intervention under investigation is. One way to do this can be to view some incidental features as characteristic. For instance, in the example of bike riding, being outdoors, and other features related to bike riding that are difficult to control may be considered a characteristic component of the treatment if viewed holistically. It is easier to ask someone to simply not bike ride than it is to develop a placebo that is indistinguishable from bike riding. This level of holisticity can be contrasted with the very non-holistic, fluoxetine hydorchloride trial example mentioned previously. In trials on that drug, the intention is to measure specifically the effect of fluoxetine hydrochloride on depression and, as such, a placebo must control for all other non-fluoxetine hydrochloride treatment effects. By broadening the conception of what counts as an intervention in sports science it may be possible to measure the effects of less specific interventions, but more adequately control them as less, theoretically, needs to be controlled for.

The best course of action may be to leave it up to experimenters to determine what counts as characteristic and incidental features of a treatment, and how holistically they view a treatment being investigated. There is, however, a major downside with viewing treatments holistically. It becomes difficult to isolate which part of the treatment *caused* the putative effect. This is troubling, both because it does not lead to a deep understanding of causal relationships that strict placebos can aid, but, also, because it makes giving recommendations for practice more difficult. As will be seen in Part III of this thesis, when we consider evidence from mechanistic studies alongside evidence from RCTs, this becomes less worrying. A holistically controlled study looking at the effects of bike riding on blood pressure may have to make a recommendation along the lines of 'given the results of the trial, we can recommend bike riding in such and such an area, at such and such a time of year, at such and such a time of day etc.' At this point it is not even clear that the bike riding was important as part of the treatment. If not viewed holistically, adequately controlling remains difficult or impossible but, where possible, we can achieve a deeper understanding of what parts of the causal relationship are important, and recommendations may be able to be made more easily. Recall, this is important as being able to make recommendations for practice is a key goal of EBP. Further to this, despite being intended to increase the external validity of trials, in a commentary on pragmatic trials Patsopoulos (2011, 220-221), points out that success of a treatment in one real-world setting does not mean that that treatment will be successful in another real-world setting, particularly when it is difficult to point to what elements in a trial have an effect on measured outcomes.

From the preceding points, we can see that there are cases in the sports sciences where placebo controls can be difficult or impossible in RCTs. This means that where placebo controls are employed, they may not be sufficient to rule out effects that can influence the trial's result that are not those being tested. As a result, trials may be published and results interpreted as providing evidence for a causal claim where the placebo controlling was actually inadequate for this claim to be made. The potential solution to this, of viewing treatments more holistically, may mean that it is difficult to determine what parts of treatments are important in order to make future recommendations. As a result, in these instances, the quality of the evidence provided by RCTs in practice is not as strong as that produced by ideal RCTs, and evidence from RCTs is not as strong as is suggested by evidence hierarchies. Next, I will argue that two potential solutions to placebo controlling trials also face issues in sports science that mean that the quality of evidence produced by trials that use them may be limited, just as placebo controlled trials are.

2.3.2.1 Solutions to the placebo problem: active controlled trials

Given what has been said in this section so far, we can see that it is the case that adequate placebos are difficult or impossible to produce in some instances of sports science research. Placebos are not, however, the only method of controlling RCTs. In medicine, rather than placebo controlling, we may conduct Active Controlled Trials (ACTs). In an

ACT, a treatment with a known effect is used as a control against which the effectiveness of a new treatment is measured (Howick, 2011b, 97). This can include comparisons between a trial treatment and current best treatments (Howick, 2011b, 97). Here, I argue that given the difficulty of determining the effectiveness of the active control being tested against in sports science without already having good placebo controls, an ACT may also lead to wrongly estimating treatment effects.

Within the philosophy of EBM there have been criticisms of ACTs, largely on the grounds that they will not have appropriate sensitivity to determine effect sizes. This is because the difference between effect sizes of the two treatments may be very small (Miller and Brody, 2002; Temple and Ellenberg, 2000a, 2000b). Where differences in effect sizes are very small, ACTs may lack sensitivity in their ability to provide evidence that one treatment is more effective than the other. Critics have also argued that the active treatment that a new treatment is being compared to may be no better than a placebo, or even harmful (Shapiro et al., 1999, 12, Wootton, 2007). In order for an ACT to possess adequate sensitivity to treatment effects, the assumption must be made that the active control treatment is effective. The active control must be assumed to be effective in order that any intervention compared to it, which has a similar effect size, can also be assumed to be effective. Even in cases where an intervention being tested is much more effective than an active control, the control must be assumed to be effective in order that we can properly estimate effect sizes. As such, any effect size shown in a trial where the active control is wrongly thought to be effective, or where it is actually harmful, will lead us to wrongly estimate an effect size and may simply be evidence that the treatment under investigation is less harmful than the active control (Howick, 2011b, 98). Jeremy Howick (2011b, chapter 8) tackles this problem in the instance of medicine. He argues that, as placebos are rarely legitimate, ACTs will not always lack sensitivity in comparison to placebo controlled trials. He claims that a good ACT can be just as sensitive, or more sensitive, than placebo controlled trials in many cases, and thus, ACTs are not always inferior to placebo controlled trials.

Let us apply this reasoning to sports science ACTs. Following Howick's argument, it seems to be the case that an ACT will be useful in sports science in instances where legitimate placebos are hard or impossible to find. He makes this claim on the assumption that the active control *is* effective. Further, his reasoning suggests that these ACTs will not lack sensitivity when compared to the equivalent placebo trials. This brings us back to a point made in the introductory chapter to this thesis. Much practice is, or was, informed by evidence now considered to be poor. This means that the evidence in favour of the effectiveness of an active control may be poor, and our knowledge of the effectiveness of it may be wanting. I contend that, in sport science, it may not always be possible to have a legitimate ACT. This is because the evidence that supports the active control as being effective may not be sufficient to justify its effectiveness as established. Unbeknownst to researchers, the active control may be useless, or even harmful. As a result, any ACT may lead to an overestimation of actual effect size if the active control has not had its effectiveness well established prior to the ACT, which, given the difficulty of placebo controlling, is also difficult. In instances where the active control is actually harmful, an ACT may only really show that the new treatment is simply less harmful. If we do not have adequate placebos with which to determine that the active control is effective in the first place, how can we be justified in using them as an active control. So, at least in the case of sports science, where the evidence in favour of an active control being active may be poor, Howick's point does not always stand. I will illustrate this with an example of a treatment that was employed dogmatically, which was eventually condemned as useless and harmful.

Consider, again, the case of recommendations made for hydration of athletes and those engaged in physical activity as outlined by Timothy Noakes (2004), and discussed in detail in the introductory chapter to this thesis. Recall that the advice given was to intake as much fluid as possible to improve performance and reduce risk of heat illness, and that some recommendations even suggested 2 litres an hour. Remember also that following this advice led to a number of deaths, and cases of brain swelling. Finally, recall that current evidence suggests that ingesting the maximum tolerable fluid does not have beneficial effects to health or performance. In fact, maximal fluid intake is harmful, both to measured outcomes like sports performance (due to discomfort and carrying extra water weight or performance loss due to brain swelling), and to health (potential for death).[2] Given that it was, for a time, considered that maximal fluid ingestion was best practice, it is not far-fetched to imagine that it, and similarly poorly evidenced interventions, could be employed as the active control in an ACT.

So, any intervention tested against maximal fluid ingestion that either does not harm performance, or that harms it similarly, but that has fewer instances of brain swelling associated with it, would appear to be a better intervention than maximal fluid ingestion. In reality, all this ACT would show is that another intervention was less harmful than maximal fluid ingestion. This would lead to an overestimation of effect sizes that would not occur if an adequate placebo existed. This is because the trial would have been conducted on the assumption that extreme rehydration was already more effective than a placebo. As is illustrated by this case, there is a worry that the active control in an ACT may be ineffective or harmful, and poorly evidenced, and is unsuitable to be used. If we want to know how effective an active control is, it needs to be compared to an adequate placebo first, meaning the problem of placebos still looms large, until we can be sure of the legitimacy of our active controls.

To further strengthen this point, and to counter potential objections, I

[2] To put it very plainly, in worst-case scenarios, it will be detrimental to performance if one dies of brain swelling during or after competing or training.

will now also consider a further example. It may be objected that, in this case, of course one would not use maximal hydration as an active control in a trial against heat illness *now*. We know the evidence that led to it being dogmatically taken as best practice is poor evidence. The objection would continue that we would only use something as an active control if we are sure the evidence justifying its use as an active control is sound. Problems with actually conducting high-quality trials to determine the effectiveness of active controls aside, there is a more damning rebuttal to this objection. It is possible that the evidence which supports an active control may still be considered to be high quality from an 'evidence-based' standpoint, but which actually falsely points to a large observed effect size. Thus, any treatment or intervention compared to it will have its effect size misestimated in relation. If we cannot be sure about the effects of even 'evidence-based' active controls, how can we be sure the effect sizes observed in trials are better than placebos? This is not simply a theoretical issue. As has been discussed in the previous section, papers do get published with misreported effect sizes, or misinterpreted statistics. I will give an example.

Through the 2010s, in strength sports, accommodating resistance training has become increasingly popular. This type of training can involve the use of, in addition to the weights themselves, elastic bands to add resistance to weighted movements as they increase tension as they are stretched, or long chains that unfurl from the ground increasing the resistance as they unfurl as less of their total weight is resting on the ground. The intention behind this is to increase resistance to a movement as the movement progresses. In theory, as many weighted exercise movements are more difficult as they start, and easier at the end, by increasing the resistance as the movement progresses with the use of elastic bands or chains, the movement remains difficult throughout. This is because, during the beginning part of a movement, 'neural intermuscular and intramuscular coordination' is the least efficient, 'resulting in a reduction in the force sustained' (Soria-Gila et al., 2015, 3260). This can be seen in the changing velocity of the barbell being lifted during these areas of lower force output. For instance, the most difficult part of a deadlift, a resistance exercise where a barbel is lifted from the floor to waist height by standing up with the bar in hand (see: Figure 2.1), is in the few inches as the bar leaves the floor. As a result, when the barbell is above this position, the lift becomes far easier. However, elastic bands increase in tension as they are stretched, thus, increasing the resistance on the barbell as the movement progresses. Accordingly, the velocity of the lift should remain constant throughout the lift as an increased ability to apply force to the bar is met with an increased need to apply force to move the bar. The intention of employing this type of training is that by ensuring the entirety of a lift is taxing, not just during areas of mechanical disadvantage, athletes can better provoke adaptions that lead to a greater expression of maximal strength, measured by the maximum weight lifted for one repetition. For more on this, see: Soria-Gila et al., 2015.

A 2015 meta-analysis appeared to find the technique to provide a significantly better maximal strength gain in untrained and strength trained athletes than did traditional, non-banded, lifting, as measured by a maximum one repetition lift (Soria-Gila et al., 2015). For instance, comparing increases in non-banded back squats for the two groups found greater increases in one repetition maximal lifts for those who incorporated banded exercises in their programmes. Soria-Gila et al. (2015, 3260) gave the recommendation, given their meta-analysis, that using ‘elastic bands attached to the barbell emerged as an effective evidence-based method of improving maximal strength both in athletes with different sports backgrounds and untrained subjects.’ Meta-analyses are seen to provide strong evidence to inform practice, as demonstrated by the Knudson hierarchy (Figure 1.4, 2014). Given that the effect size was shown in a meta-analysis, one could also call this active control ‘evidence-based’, where one may not say the same for the hydration example. So, as the evidence justifying the use of elastic bands in strength training seems

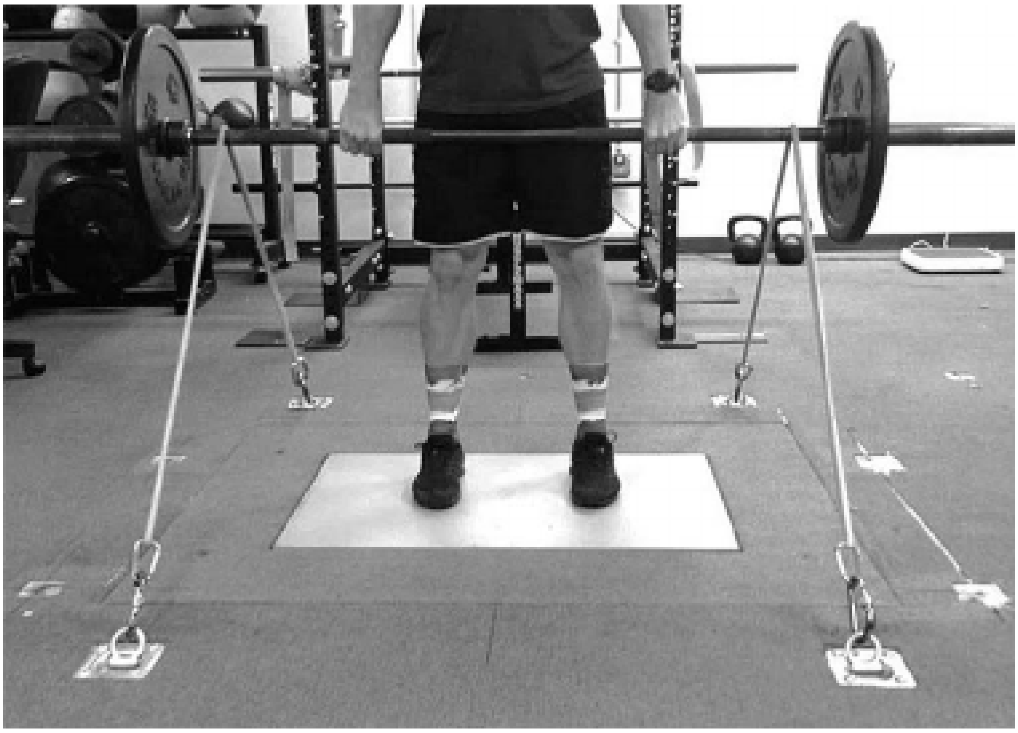


Figure 2.1: An athlete standing in the finish position of a deadlift with elastic bands. Reproduced from Galpin et al., 2015.

strong, it would also then seem like a potential treatment that could be used in ACTs as a current best practice against which to compare new interventions that aim to improve maximal strength. However, further research indicates that the effect size found by Soria-Gila et al. (2015) was due to an incorrect transcription of results between spreadsheets (dos Santos et al., 2018, e54). When, in 2018 this error was found and corrected in the calculations that were used to find an effect size, it was then determined that the results used actually show no effect size (dos Santos et al., 2018). As such, for three years, a meta-analysis seemed to provide strong evidence in favour of a practice as having a significant effect size, when in reality, this is not what the results of the studies considered actually showed. Thus, for a time this was considered to be an effective, evidence-based intervention. Any trial that utilised banded resistance training as an active control that found an effect size of a new intervention would overestimate the effect size of that new intervention. So, still, being unable to find adequate placebo controls against which to test the effectiveness of treatments is a problem for RCTs as one cannot be sure the active control is effective in the first place. This means that Howick's arguments in their favour are not effective.

2.3.2.2 Solutions to the placebo problem: dose response trials

Dose response trials can be used to determine if a measured treatment related response changes as dose increases. Hill, who set out an early list of criteria one may look for when determining if causal relationships exist in medicine, suggested that evidence of causation can be considered to be more compelling if it shows a dose response relationship or biological gradient than if it does not (Hill, 2015). A 'dose-response relationship' can refer to how an outcome or risk changes with increasing exposure (Emilien et al., 2000, 33). Dose response relationships are useful when investigating complex causal relationships, such as that between smoking and cancer. Where there is a relationship between cause and effect, and where measured outcomes are larger with a greater dose of the cause, this can put causal relationships in a ‘clearer light’ (Hill, 2015). A good dose response trial still needs to be randomized, blinded, have an adequate sample size. Rather than comparing to an active control, as has just been discussed, it may be possible to compare different doses of the treatment under investigation in order to determine if increasing dose increases effect size. For instance, comparing differing levels of vigorousness or time with a massage, or perhaps comparing bike riding for exercise where one bike has a secret motor that assists the rider, and where one is powered only by the rider may be ways of implementing different doses in sports science trials. For a more detailed overview of dose response trials, see Emilien et al., 2000. It must be noted that dose response trials also require the use of placebos or active controls. A significant trend without the use of a placebo or active control in a dose response trial is not seen to be strong evidence for drug effects (Ruberg, 1995, 2). This is because, without a placebo we may think there is a trend between dose and response, but a placebo is needed to indicate if this is the case or not. This is illustrated by Figure 2.2. In this graph, we can see that without the inclusion of a placebo dose, we may be likely to interpret the results as indicating a trend (Ruberg, 1995, 3). I have already argued that this is problematic in the sports science. But, even disregarding this point, there are further problems with the effective use of dose response trials in the sports science.

If we take the previous example, hydration in sport, a dose response trial may be useful in this instance. If, for nothing else, showing the uselessness of the treatment, that heat illness does not decrease with increased fluid intake in the way suggested by original recommendations. However, a dose response trial requires adequate sample sizes in order to show effects (Emilien et al., 2000, 33), and, as has been shown in the previous subsection, this is often problematic in the case of sports science trials. Thus, whilst in some instances, a dose response trial may theoretically be an effective solution to the problem of placebo controlling

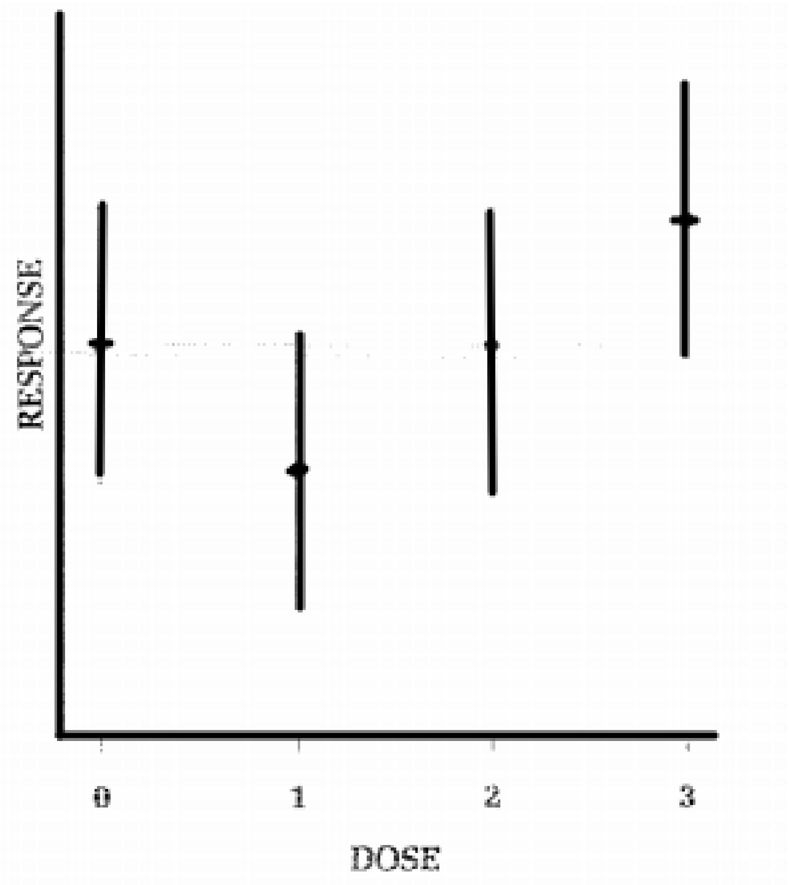


Figure 2.2: An illustrative dose response graph reproduced from Ruberg, 1995, 3

in sports science, it becomes difficult in practice.

In some types of trials, such as exercise trials, dose response trials may be hard to conduct for further reasons. It is easy to imagine that if an exercise trial sought to measure the effects of increasing the amount of exercise completed, there may be difficulties in ensuring adherence of athletes in both low-volume and high-volume groups. If we consider the type of person who is likely to engage in sports science trials: those who are athletes themselves, those in low exercise volume groups may be unwilling to limit the amount of exercise they complete, at risk of under-training or de-training influencing their performance outside of the trial. Those in high-volume groups may be unable, or unwilling, to complete large amounts of exercise, especially at the required intensity. The lack of data needed in some instances for higher training volumes in exercise studies is illustrated in a meta-analysis of the dose response relationship between weekly resistance training volumes and muscle hypertrophy. It was found that there were insufficiently many studies performed using higher training volumes to draw conclusions about upper bounds of the relationship once weekly resistance training dose reached a certain point (more than 12 sets per week) (Schoenfeld et al., 2017, 1080).

Finally, as I have already argued, there is a problem with fielding sufficiently large sample sizes in sports science to see small effect sizes, and to rule out chance as a confounder. Different types of dose response trial need large sample sizes to be conducted properly. Parallel assessments, where different groups take different doses of an intervention at the same time, need the largest sample sizes. Parallel type trials are the most common type of trial. Crossover trials, where the same participants receive different doses of the same intervention over time, need the smallest sample sizes. In a meta analysis of 2103 dose response trials in medicine, the average sample size for crossover trials was 37, with the lowest sample size of 31 and the highest of 91 (Huang et al., 2015). In the same meta analyses, the average sample size in parallel trials was found to be 151.

Given the difficulty of fielding sufficiently large sample sizes to see effect sizes in sports science trials generally, and that dose response trials need sufficiently large sample sizes to accurately measure effects at different doses, including the placebo arm of a trial, dose response trials in sports science may also suffer from the same sample size problems as typical trials. Thus, they may not be a good solution to the placebo problem in many instances.

2.3.3 Blinding in the sports sciences

As has been explained, most would consider that an ideal RCT should be *double blinded* to help rule out potential psychological confounders. EBM textbooks claim that it is 'necessary' (Straus et al., 2019, 131). A trial is double blinded when neither those conducting an experiment, administering an intervention, nor the trial participants, know if they are in the test group or the control group (Howick, 2011b, 68). The official textbook of EBM claims that blinding patients will help to stop them having 'hunches' about which group they are in and reporting symptoms differently (Straus et al., 2019, 131). The blinding of people administering or assessing treatments also prevents treatments being administered differently or the misinterpreting of symptoms (Straus et al., 2019, 131). The intention of double blinding is to remove the effects that knowing which treatment group one is in may have on the results of the trial.

The usefulness of double blinding is not uncontested. Ney et al. (1986, (119)) for instance, argue that in situations where double blinding is used, Philip's paradox will apply. They state this paradox as follows: 'the more potent a therapeutic variable the less likely its efficacy can be 'proven' in a double-blind study' (1986, 119). The idea here being that the more effective a treatment is, the less suitable it is for testing in a blinded trial because it is harder to effectively blind. Howick counters that, regardless of this, blinding is 'an instrumental good: it is valuable insofar as it rules out potential confounders arising from participant and caregiver knowledge of who receives the experimental intervention' (2011b, 69). As such, blinding should still be attempted. The outcome of this, however, is that to be useful the masking must succeed. In this section, I will argue that in many types of sports science trial, it does not.

It is particularly difficult, in many types of sports science research, to double-blind a trial. Even if a patient does not know what is being studied, they may have hunches about what group they are in based on what they are asked to do. Consider a study on massage, there are many incidental features of massage that a participant will expect, manipulation of their muscles, reduced soreness, increased mobility, perhaps pain or relaxation. In a study where massage is being applied, it would likely be easy to work out for a participant that they were not really receiving a massage. If someone who enrols in an exercise study is told not to exercise, they may be able to infer that they are a control. A participant in a nutritional investigation may have a good hunch about what side of the trial they are on based on how they are told to eat. The participant having an active role in the intervention - having to act in a certain way - will likely be a part of how they will be able to have a hunch about what group they are in. The difficulty of blinding these types of trials can be compared unfavourably to medical trials where the active involvement of the participant in the trial may be as simple as taking a pill, and, thus, the blinding of receiving the treatment is far less complicated. It must be noted that participants are not always told what intervention is being studied, so these issues may not always arise.

Perhaps a stronger case for the problem of double blinding in sports science is that many trials in the sports sciences will actively involve clinicians or professionals as part of the intervention. If someone administering an intervention is an active part of that intervention, it will be almost impossible for them to not know what group a patient is in. For instance, a masseuse in a trial on massage will likely know if they are giving a sham massage, and a professional coach will be aware they are asking a participant to perform a control exercise regime. It may be, again, in part, this active involvement as part of the intervention, which can sometimes be seen in many sports science interventions and trials, that makes double blinding difficult or impossible.

The problem of blinding is closely related to that of placebo controlling. A good placebo is essential to ensure trials are well blinded. Blinding is likely easier than adequately placebo controlling, however. Even a moderately successful placebo can be used to blind a trial, particularly if participants do not know what is being measured. The difficulty of creating a placebo in many sports science investigations that is able to blind *and* fit the Grunbaüm criteria is a much greater ask.

2.4 Rating RCTs in sports science

Now, all the key elements of the **Excluded Explanations Argument** have been given. It can be summarised as follows:

From section 2.2 we know that.

- An RCT is meant to provide high-quality evidence on the grounds that it rules out alternate explanations for trial outcomes.
- If things other than the intervention or exposure being tested can explain the outcome of a trial, we cannot rule in the intervention or exposure being tested as the only explanation for observed outcomes.
- The less well RCTs rule out alternate explanations for differences in observed outcomes, the lower the quality of evidence those RCTs produce is. This is, in part, because we have less confidence in our ability to rule *in* the intervention or exposure being tested as a

cause of those differences.

None of these points, so far, are controversial.

From the arguments made in section 2.3, we know that due to the nature of the types of research conducted in the sports sciences, and the types of interventions and causal relationships under investigation, it can be difficult or impossible to conduct high quality RCTs. This is because the conditions required to rule out bias, confounding, and chance, as explanations for observed correlations in trials can be difficult or impossible to fulfil. The conditions are that a trial must have:

- adequate placebo controlling,
- a sufficiently large sample sizes, and
- effective blinding.

Where this is the case, the RCT will be far from ideal and produce less-than-strong evidence. The controversial point made in this chapter is not that low quality RCTs produce low-quality evidence. This is readily accepted. Given what is indicated by evidence hierarchies, particularly the sports science focused evidence hierarchy produced by Knudson et al. (2014), the controversial point that I have demonstrated in this chapter is how limited the evidence provided by RCTs in sports science is in practice. For instance, consider Knudson's claim that 'In general, meta-analyses, systematic reviews, and randomized controlled trials (RCT) provide the most trustworthy evidence' (2012, 131). Also consider that RCTs in the sports sciences are often 'privileged' above other types of evidence (Ivarsson and Andersen, 2016, 12 and 19), and are seen as a 'gold standard' tool (Ivarsson and Andersen, 2016, 12). The claim that in general RCTs in the sports sciences provide trustworthy evidence seems troubling, when I have argued that the evidence they produce will often not be of a high quality. Many key areas of sports science research face at least one limitation to the quality of evidence they can produce. Studies on elite athletes, and exercise and nutrition studies for instance, make up large swathes of the research conducted. Based on this, it is not appropriate to label RCTs as providing strong evidence in general, when they often do not. It is not justified to claim that relying primarily on evidence from RCTs will regularly provide sufficient justification to establish causation or inform practice in reality. It is also not appropriate to claim that relying primarily on evidence they produce constitutes relying on the best possible evidence, as will be seen in chapter 5.

Given the discussion so far in this chapter, we can see why the EBP practice of relying on evidence from RCTs as a high-quality evidence gathering technique may have some problems and may not count as relying on the best possible evidence. RCTs can not always be assumed to provide strong evidence. Thus, relying on them dogmatically, particularly at the expense of other types of evidence, may mean practice is *not* informed with the best possible evidence. Given this, some advice can be offered. The evidence individual RCTs provide should be evaluated. The dogma that RCT evidence may be taken as strong grounds for practice, or can always justify practice, on the grounds that in general they *may* provide strong evidence, should be avoided. Instead, due to the risk of poor evidence, the quality of RCTs should be assessed for quality by experts on an individual RCT basis, rather than blindly following hierarchy suggestions which rank them on a general basis. In chapter 5, I will argue that this evaluation, to really rely on the best possible evidence, should include the evaluation of evidence from both association studies and mechanistic studies. This point about mechanisms aside, there are a range of tools to assess the quality of evidence to help make these types of recommendations in medicine; one has already been briefly mentioned: the GRADE hierarchy (Guyatt et al., 2011). Unfortunately, as the quality of evidence produced by RCTs in sports science is so often limited, these medical rating systems may not be sufficient to rate sports science

RCTs.

GRADE ranks the quality of evidence to support a claim from high to very low, with adjustments being made based on the evidence gathering method, and quality of individual instances of evidence gathering methods under consideration (Schünemann et al., 2013, chapter 5). For instance, an RCT is, at first, assumed to provide high-quality evidence. Once limitations are assessed, they can be used to downgrade the perceived quality of evidence that may be derived from individual RCTs (Schünemann et al., 2013, chapter 5). This seems like a useful tool to evaluate the evidence in sports science. However, it must be noted that GRADE and other rating systems are developed with medical research in mind. This may mean that they are not entirely suitable as tools to rate sports science RCTs. For instance, where GRADE suggests ranking evidence down one or two categories for medical trials (Schünemann et al., 2013, chapter 5.2), this may be insufficient to account for the potential for, and magnitude of, limitations to evidence quality in some sports science trials. It is recommended in the GRADE handbook, for instance, that '[o]ne should be conservative in the judgement of rating down.' (Schünemann et al., 2013, chapter 5.2). However, as a sample size of 10 is not uncommon in sports science (as is suggested in Pyne et al., 2010), this recommendation may lead a reviewer to not rank down the quality of evidence from that study sufficiently as that sample size does not seem abnormal. This can be compared to medicine, where expected trial participant numbers are much higher than this, and would then be marked down. For example, a US regulation body, the FDA, recommends a sample size of 300 to 3,000 in trials to determine the effectiveness and side effects of drugs (US Food and Drug Administration, 2016). This number may be higher than is necessary for sports science, given the potential risks of medical drugs and need to find side effects, but note must be made of the disparity between sample sizes and how recommendations are made. So, whilst RCTs should be rated individually in sports science to assess the quality of evidence they provide, rather than blindly assuming they provide strong evidence, current rating systems developed for medicine may be almost as misguided as the evidence hierarchies already borrowed from medicine.

In chapters 4 and 5, I will go into greater detail about what can be done to establish causality in the sports sciences, particularly in areas where RCTs can not provide evidence of a high quality. But what I will discuss there must be alluded to here to make an important point. Given the limitations to evidence produced by RCTs in the sports sciences, it may be impossible for evidence from RCTs alone to ever provide sufficient evidence to establish causal claims in some areas of research. This is because there may be instances where evidence from RCTs will never be sufficient to establish the existence of a mechanism, which I argue is essential in chapter 4. This means that even where the quality of evidence RCTs can provide can be rated, if these rating systems do not take seriously other sources of evidence, like mechanistic studies, they will not be able to account for the establishing of causal claims.

2.5 Conclusion

In this chapter, I introduced the idea of an ideal RCT. I introduced, also, the idea that the further an RCT is from meeting the ideal criteria, the lower the quality of evidence it produces. I made this claim using the **Excluded Explanations Argument**, which argued that because trials in the sports sciences often cannot have adequate samples sizes, placebo controlling, and blinding, they are often far from ideal. The outcome of this being that they fail to rule out confounders sufficiently to rule in interventions or exposures being tested as being responsible for any results observed. Thus, outcomes observed from trials cannot be said to be the result of an intervention or exposure under examination with a high level of confidence.

In the next chapter, I introduce single subject RCT style trials, N of 1 trials, which are sometimes suggested as ways to overcome the limitations to evidence posed in sports science by RCTs. I argue that although they can overcome some limitations to evidence quality, they also suffer similar problems. This means that the **Excluded Explanations Argument** can still apply to them. This, then, leads me to Part II of the thesis, where I argue for a solution to these problems: in order to establish causal relationships, and inform practice based on the best possible evidence, we will often need to assess evidence from both association and mechanistic studies.

Chapter 3

N of 1 trials: not a solution to the problems of group RCTs

3.1 Introduction

The key reason for introducing a discussion of N of 1 trials into this thesis is that, in a paper criticising the quality of evidence used to inform elite athlete nutrition, Jukola (2019) suggests that N of 1 trials may provide a solution to this problem. Jukola's paper advances similar criticisms to elite athlete nutrition specifically to those I made for sports science in general in the previous chapter. Jukola suggests N of 1 trials as a solution to the problem very briefly, and only for elite athlete nutrition. In this chapter, I engage in a much more in depth look at N of 1 trials, and my discussion of them is not limited to elite athlete nutrition.

In this chapter, I argue that, in some instances, N of 1 trials may provide evidence of a higher quality than groups RCTs provide. Accordingly, N of 1 trials may be useful tools in some areas. I argue this on the grounds that, unlike group RCTs: N of 1 trials do not lose information about individual treatment effects that may be relevant to intervention outcomes, they can be used to personalise interventions, and they do not suffer with problems stemming from small sample sizes or lack of representativeness. Ultimately, however, I do argue that N of 1 trials will suffer from the same problems with adequately placebo controlling that group RCTs in the sports sciences do. This means that the **Excluded Explanations Argument** does still apply. I also concede that whilst N of 1 trials conducted on the target individual will provide evidence for effectiveness in that target individual, that evidence may have limited generalizability outside that individual. These points have two implications. Firstly, whilst N of 1 trials may overcome some problems that arise in group sports science RCTs, they still provide evidence that is less strong than that produced by ideal RCTs, and that is less strong than evidence hierarchies suggest RCTs should produce. Secondly, the scope and practical use of N of 1 trials may be limited by difficulties faced when trying to draw general conclusions.

3.2 What are N of 1 trials?

In sports science, N of 1 trials have been employed in many areas, for instance: in behavioural sports psychology (Martin et al., 2004), in the development of weight management and physical activity strategies (Kwasnicka et al., 2017; McDonald, Vieira, et al., 2017; Ordovas et al., 2018), in sports nutrition (Jukola, 2019), and in sports performance (Guyatt et al., 2000). Single subject studies can be conducted in a number of ways. For instance, they can be conducted as single subject case studies, observational studies, and as RCT style studies. In this chapter, I am concerned with RCT style N of 1 trials. There are a few important factors that set the N of 1 style RCT apart from other single subject research studies: the use of a control, randomisation of control and intervention delivery phase sequences, attempted blinding of participant and administrator, and defined outcome measures (Tate et al., 2013, 621).

In many ways, an N of 1 style RCT is similar to a group RCT. Further, there is the assumption that if the trial is well controlled, a difference in outcome measures can be attributed to the effects of the intervention. A key difference between group and N of 1 style RCTs is that in N of 1 trials, as there cannot be separate control and trial groups, the individuals on whom a trial is conducted must act as their own control. This means that participants must receive both the control and the intervention at different times so that outcome measures can be compared (Tate et al., 2013, 620-621, Gabler et al., 2011, 761). In order that this can happen, participants receive the control in the *baseline* phase and the intervention in the *intervention* or *trial* phase. Throughout both phases, participants are repeatedly measured, which allows intervention effects to be inferred (McDonald, Quinn, et al., 2017). Group RCTs compare average outcome measures between control and intervention groups in an attempt to find a generalizable effect size. In contrast, N of 1 trials, by comparing outcome measures between phases in an individual, attempt to determine the effect size of an intervention for an individual. This is called the Individual Treatment Effect, or ITE (Gabler et al., 2011, 762).

As has been mentioned, a key part of all of these types of trial is establishing a control in a baseline phase against which outcome measures in intervention phases can be compared. For a baseline phase to be considered a good control, the outcome measure must show low variability, and no visible change in trend (Sands et al., 2019, 6). There may be a trend, such as a steadily improving sports performance, because it may not be feasible to totally eliminate every factor that can influence an outcome measure. What is important is that this trend must not change. For example, if in a baseline phase a 100 m athlete routinely decreases their 100 m time by a 10th of a second every month, this is an acceptable baseline phase. If their change in performance varies wildly, this would not be considered to be a good baseline phase against which to make reliable effect size estimates.

3.2.1 Trial design

The design of RCT-style N of 1 trials may differ (Kinugasa et al., 2004; Sands et al., 2019). In order to explain this, I will give hypothetical and real life examples of N of 1 trials in sports science. The most simple trial design is the AB trial. In the AB design trial, a baseline outcome measure is taken across the A phase, and in the B phase, the intervention is introduced. Following the trial, outcome measures can be compared between phases with the aim of estimating an effect size of the intervention being tested (Kinugasa et al., 2004, 1037). This can include the comparison of outcome measure trends through phases, and outcome measures at the end of phases. As this type of trial takes baseline measures before intervention measures, the order of phases cannot be randomised. Instead, what can be randomised is the length of phases. If this is done well, participants should not know when they are in baseline and intervention phases of a trial. AB trials can also be extended into ABA and ABAB trials, and trials with growing numbers of A and B phases. In these types of trial, interventions are introduced and withdrawn. This allows for outcome measures to be taken for multiple baseline and intervention phases. This also allows for changes in outcome measures as interventions are introduced and withdrawn to be observed (Kinugasa et al., 2004, 1037), which may provide important benefits in some instances. This may be the case, for instance, with physiotherapy interventions. It may be important to observe if pain returns to athletes after an intervention is withdrawn and the athlete stops engaging in it. Perhaps an intervention is only effective when regularly utilised, for instance. Withdrawal and reintroduction trials may also be useful in the case of altitude and heat alleviation trials. This is because once heat and low oxygen adaptations have been made and then lost by athletes, they are retrained more easily than they were trained initially (Gibson et al., 2020, 19)[1]. In other types of N of 1 trial, additional interventions may be

[1] I would like to thank Neil Maxwell of the University of Brighton for furnishing me with this example in private correspondence.

tested in further phases (Kinugasa et al., 2004, 1037). For instance, an ABC trial would involve taking baseline measurements, measurements when intervention B is employed, and measurements when intervention C is employed. Intervention C could be a totally different intervention to B. It could also be a variation on B, which may help to investigate individualised interventions, a topic I will come to in subsection 3.5.1.

The following is an example of a hypothetical trial which illustrates why an intervention withdrawal trial, and ABA trial, may be conducted. There may, for instance, be evidence from group trials that a certain style of training may have the effect of increasing the number of hours slept by athletes when averaged over a group, which could contribute to an improved recovery from exercise and readiness to exercise again. A team sports scientist may, then, want to determine if an athlete under their charge will sleep more whilst undergoing that type of training. Under researcher instruction, the athlete begins to report nightly sleep duration. After a baseline has been established, phase A, the coach is then instructed to employ the potentially-sleep-improving training modality, phase B, without telling the participant training has changed, or what it may change. The training modality can, then, be removed and introduced at randomized intervals, with sleep data being collected for those intervals, data being collected for multiple A and B phases. If nightly reported sleep time is greater during periods where the new training modality is used, and then decreases upon withdrawal, this may be taken as evidence that the new training modality did cause improved length of sleep duration in that individual. This is, of course, a simplistic example. Modifying training may have effects on important performance factors other than improving sleep, and improving sleep may not be directly linked to improved performance via recovery. The new training modality may be less effective at improving sport specific skills, for instance. A more complex type of N of 1 trial is necessary to measure this, the multiple baseline trial, which is described in the following paragraph.

The most common type of trial employed in single subject research is not the AB trial, it is the multiple baseline trial (Kinugasa et al., 2004, 1038). These trials take baseline measurements for multiple different outcomes A, A', A" etc. and measure the outcome for those different measures in intervention phases B, B', B", etc. (Kinugasa et al., 2004, 1038-1039). This can be done in parallel, or sequentially, which I will explain shortly. Multiple baseline trials are useful when researchers want to examine the effects of one intervention across different measures, multiple interventions over different measures, or when multiple interventions need to be assessed without the withdrawal of others (Kinugasa et al., 2004, 1039). As data is sampled for all outcome measures before the introduction of interventions, it is supposed to be possible to see if specific interventions are associated with changes, or changes in rate of change, in non-target outcome measures (McDonald, Quinn, et al., 2017, 313). For instance, if we test two interventions and their effects on different outcomes, if one intervention affects both outcomes, this can theoretically be observed. This does raise concerns about the ability to obtain clear evidence for effects of interventions independent of one another in multiple baseline trials. However, as these trials are often conducted when seemingly effective interventions cannot be withdrawn, such as for ethical reasons, this may not always be a concern.

When done in parallel, measures are taken of outcome variables A, A', A" etc. in the baseline phase. Each of these measurements is considered a different baseline. Following this, an intervention is introduced, and the outcome measures are taken again for each corresponding outcome B, B', B". etc. This allows researchers to observe changes in multiple outcome measures when a single intervention is introduced. To re-use the sleep intervention example, a parallel multiple baseline trial could be used to establish baselines for different outcome measures, as well as sleep, in order to provide evidence for the effects of an intervention on other important factors such as sport specific skills. In this instance, outcome measures could be A, sleep, A', time to run a certain distance, and A”, maximum jump height. If, once the intervention was introduced, A improved but A’, and A” stagnated or got worse, researchers could use that information to try and determine if the new sleep improving training modality, whilst having a beneficial effect on sleep, overall had negative performance outcomes and was not beneficial to the athlete.

To illustrate sequential multiple baseline studies more clearly, an example of measurements taken during a hypothetical multiple baseline study of an intervention on football skills given by Kinugasa et al. (2004), is replicated in Figure 3.1. In this hypothetical scenario, the interventions being tested are the adoption of new specific training programmes for three different football skills. Here, baseline outcome measures are taken for each skill, and specific coaching interventions for those skills are added sequentially. By ensuring that only the outcome measure that is intended to be affected by each intervention varies with the introduction of that intervention, there can be an amount of assumed independence of intervention effects. This independence can be tested statistically. For instance, as we can see from Figure 3.1, when the dribble skill is introduced and A moves from baseline to intervention phase B, the value of the outcome measure varies. However, the values of the outcome measures in A’ and A” remain relatively invariant until the relevant intervention for each is introduced, and they enter their respective intervention phases. This allows baselines to be established for each skill, and evidence for the effect of each intervention on its relevant skill to be collected and the assumption that each intervention only effects its relevant skill to be maintained. Other types of N of 1 trial are also utilised in sports science, such as those used to measure the effects of multiple interventions on one outcome, but these are not well represented in the literature (Kinugasa et al., 2004, 1039), and for this reason will not be explained.

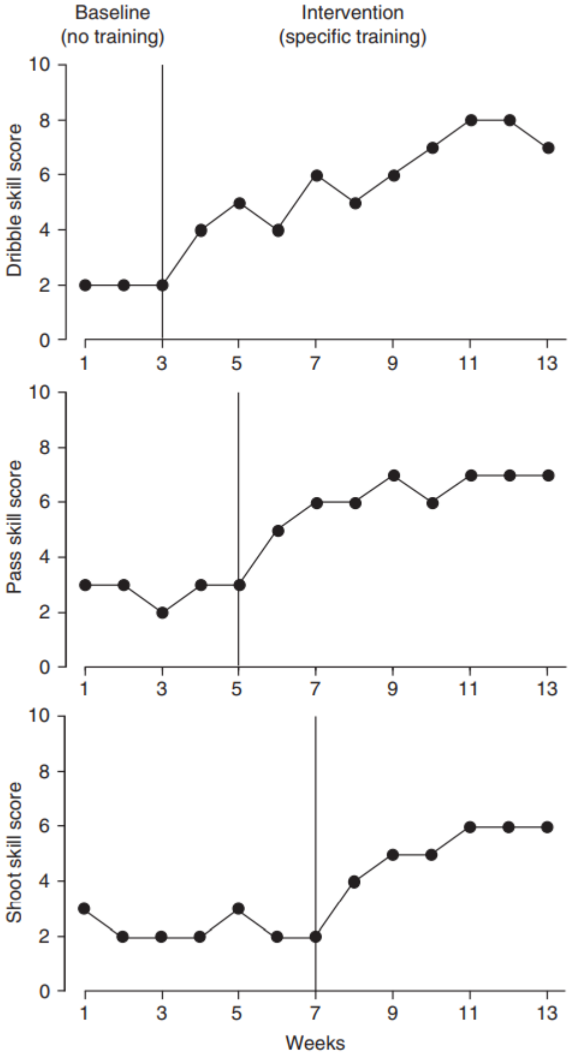


Figure 3.1: A hypothetical multiple baseline trial measuring the effects of 3 different football skill specific interventions on those 3 different football skills.

3.2.2 N of 1 trial result assessment

The results of N of 1 trials in sports science need to be assessed once measurements are taken (Sands et al., 2019, 6). According to Kinugasa et al. '[t]his complex evaluation ... is currently based on the coach's intuitive judgement or subjective visual analysis', although statistical analysis is also used (Kinugasa et al., 2004, 1041). Beyond visual and statistical analysis of intervention effects, it is sometimes practice to subjectively infer from a benefit, as perceived by an athlete, that an intervention is effective (Kinugasa et al., 2004, 1047). This is problematic in light of EBP methodology, that wishes to borrow from EBM and move away from expert and subjective judgements. Relying on opinion and subjective judgement is worrying as it can fall foul of many biases, such as those explicated by Tversky (1974), Kahneman (2000), and Detmer (1978). Expert opinion is seen as so damaging in medicine that Sackett, who advocated against its use, retired as soon as he believed he was an expert (Howick, 2011b, 148). This, however, is not the death knell for N of 1 trials. New and rigorous methods of statistical analysis are emerging, particularly with the availability of statistical computing software like SPSS. Further, the importance of giving a strong statistical education is growing in sports science departments. Until recently, the teaching of statistics in sports science departments was either minimal, or not done at all at undergrad levels.

One important benefit of utilising statistical analysis in N of 1 trials is that we can employ powerful statistical techniques, such as statistical process control (SPC). Using SPC, it can be determined whether an outcome measure under investigation is within normal limits *throughout* the course of the intervention phase, or if it is varying to an extent that indicates the intervention had an effect. SPC allows this to be done with more precision than intuitive or subjective judgement. In SPC, an outcome measure is considered to be 'in control' if it only varies within normal, expected limits. SPC is used to identify patterns in data by continually sampling outcome measures in an individual (Sands et al., 2019, 11). There will be an expected, daily, monthly, and yearly variability in outcome measures (dependent on what the measure is) which do not require explanation, and may be as a result of factors such as chance or expected biological fluctuations (Sands et al., 2019, 11-12). Determining this involves sampling outcome data regularly throughout the trial to observe patterns and variation that occur in outcome measures. If, once an intervention is introduced, outcome measures sufficiently change beyond normal expected variability, determined by SPC, it may be able to be inferred that this is a result of the intervention and that the intervention is having an effect on that outcome (Sands et al., 2019, 11). Used in this way, SPC has similar utility to the use of p-values in group RCTs discussed in subsection 2.3.1. Just as p-values indicate how likely it is that an outcome measure could be achieved if the intervention was not effective, SPC can be used to determine the likelihood outcome measures would change as they do if an intervention was ineffective.

This statistical technique is useful as it allows both changes in measured outcomes throughout the trial to be observed, but also because it allows for the analysis of outcome measures to be made whilst also explicitly taking into account normal variability in outcome measures. For instance, in a trial on 100 m running times, the fastest time to run 100 m in an athlete's training each day may vary from the mean by 0.1 seconds in the baseline phase. Then, when an intervention is introduced, the daily fastest time is regularly 0.2 seconds faster than the mean established in the baseline phase, statistical process control can be used to determine if this variability is expected under normal conditions, or if this variability in outcome measures is sufficiently large that it is unlikely to have occurred without the introduction of the intervention.

As has been mentioned previously, in some instances, data in N of 1 trials may be evaluated subjectively, and values of outcomes measures may be obtained subjectively. For instance, in the football skill example (Figure 3.1), the ability to perform football skills was assessed on a subjective, points based system. McDonald et al. (2017, 315) claim that subjective interpretation and measuring is 'prone to a number of errors and biases', but that this concern about the quality of evidence is diminished as technology advances because it becomes easier to measure more outcomes objectively and confidently. They concede, however, that some outcomes will always be subject to subjectivity and self report, and therefore errors and bias. I note that if an investigation must be carried out with subjective data collection in an N of 1 trial, it is highly likely that other types of trial using the same outcome measure will also use self reported or subjective data. Therefore, where this may cause a problem for N of 1 trials, it will also cause a problem for other types of trial. For example, if an N of 1 trial on an intervention must use subjective measures for outcomes, a group trial would be subject to the same limitations to evidence quality for this reason. Thus, the problem is not specific to N of 1 trials and, instead, is a problem for any research in the field trying to use that measure.

3.3 Example N of 1 trials

Before I argue that there are instances where N of 1 trials are better than group RCTs, it will be helpful to give some examples of real N of 1 trials from the sports science literature. Below, I include two: a physiotherapy example, and a sports performance example. It must be noted that despite having multiple individuals in the trials, the following trials are still conducted as N of 1 trials to obtain evidence for intervention effects in individuals rather than groups. Perhaps it is better to think of them as multiple N of 1 trials happening concurrently, investigating the same interventions.

A 2011 study by Bernhardsson et al. on the effects of a home-based, and physiotherapy-supported, intervention for subacromial impingement

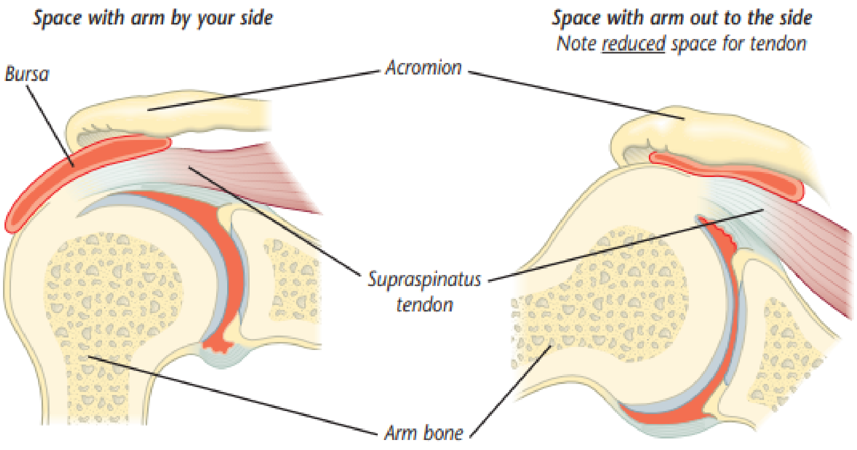

Space with arm by your side
Space with arm out to the side
Note reduced space for tendon
Bursa
Acromion
Supraspinatus tendon
Arm bone

Figure 3.2: This image demonstrates reduced spaced in the shoulder joint for the supraspinatus tendon to move (reproduced from Oxford Shoulder and Elbow Clinic, 2004).

syndrome can be used to exemplify how N of 1 trials may be used in sports science. Subacromial impingement syndrome accounts for around half of all shoulder pain complaints at physician visits (Umer et al., 2012, 79). It is the impingement of tendons in the rotator cuff against parts of the acromion, coracoacromial ligament, and the acromioclavicular joint (Bernhardsson et al., 2011, 70). The rotator cuff consists of the muscles responsible for shoulder stability, the supraspinatus, the infraspinatus, teres minor, and the subscapularis (Frederic, 2017, 328), essentially, the muscles on the back of the shoulder, that go around the shoulder blade. The general anatomy of the rotator cuff is shown in Figure 3.3. As a simple explanation of the impingement, it occurs when the space in the shoulder joint through which tendons move is narrowed enough to cause swelling of, and pressure on, the tendons. This is illustrated by Figure 3.2.

The trial was conducted as follows. First, over three weeks baseline outcome measures were taken for assessed shoulder function (such as mobility), and for subjective pain levels. Following this, a 12-week eccentric shoulder strengthening routine was prescribed and carried out. It

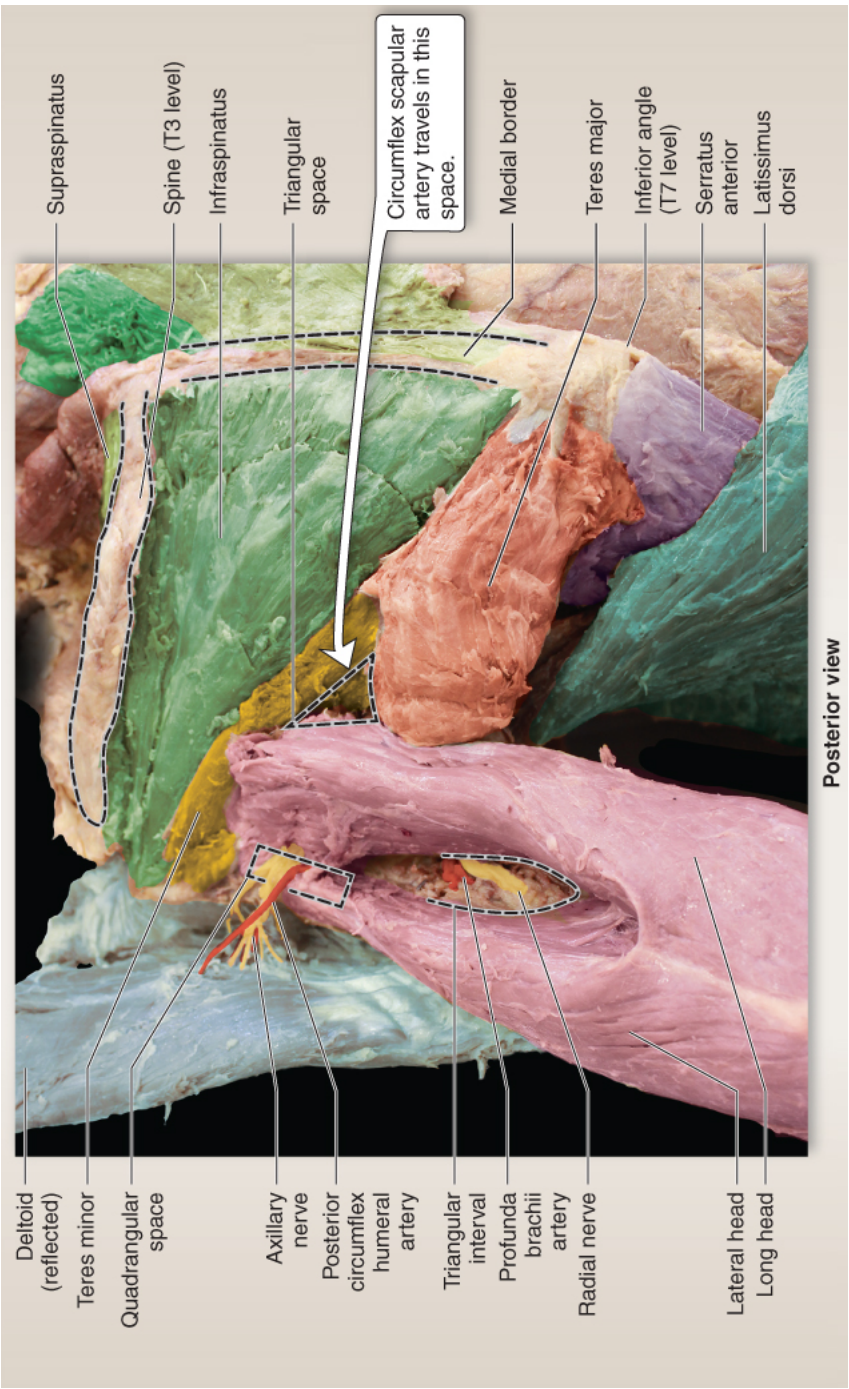


Figure 3.3: This image shows the muscles of the shoulder and rotator cuff (reproduced from Harrell, 2019).

involved daily strengthening exercises for the muscles in the rotator cuff. Outcome measures were assessed throughout the course of the 12-week intervention phase. Unlike in a group RCT, where one group is a control group against which outcome measure changes are compared, subjects were their own control, with no between participant comparisons being made. Instead, participants had their own outcome measures compared between A and B phases in order to estimate effect sizes of the intervention in those individuals. Bernhardsson et al. (2011, 78) claim that by utilising the single subject design they were able to evaluate the process of treatment in individuals rather than simply end results, and that this can be useful to help identify new treatment methods for individuals. This allowed, for instance, those who had increases in pain whilst in the intervention phase to drop out of the study.

In another, perhaps more obviously sport related example, Scott et al. (1999) observed differences in outcome measures when applying different mental strategies to indoor rowing performance. The trial compared the utilisation of awareness and distraction strategies to 40 minute maximum distance performances on an indoor rowing machine. Awareness strategies employed an increased and concentrated focus on rowing, movement, and pace. In opposition to this, distraction strategies allowed athletes to focus on things other than rowing, such as music and videos. Results were compared for individuals between baseline performances, performances using awareness strategies, and performances using distraction strategies, measuring distance rowed as the outcome. All 9 study participants showed improved performances when utilising awareness and distraction techniques, in comparison to baseline performances, but to varying degrees. Some athletes performed better with distraction techniques, and some athletes performed better with awareness techniques. This exemplifies the problem of generalisability by highlighting how whilst the results of an N of 1 trial may be relevant to one person, this does not mean they will be relevant to another. One person may benefit more from a distraction intervention, but that does not mean a further person will not benefit most from an awareness intervention. I will discuss this problem in detail in section 3.6. Scott et al. (1999, 67) claimed that the single subject nature of the research was useful as it allowed for greater analysis of intervention effects in individuals, and comparison between different performance strategies in individuals. This type of research may be useful for coaches when multiple conflicting interventions may be useful in order to determine which different athletes under their charge would be best suited to.

3.4 N of 1 trials in sports science versus medicine

A prevailing theme through this thesis is that sports science should not be considered to be so similar in research type, goal, or method, to medical research, that conclusions drawn about the quality of evidence in medicine may always, without argument, be said to apply in the case of sports science. This has been the case, for instance, with evidence hierarchies in general, and the quality of evidence provided by group RCTs in particular. It is the case here, also. Sports science research is sufficiently dissimilar from medical research that conclusions drawn about N of 1 trials in either discipline are not automatically relevant in the other. There are a number of novel aspects and considerations in sports science that mean that N of 1 trials have special uses as a research tool unlikely to occur in medicine.

As I construe sports science broadly, this means that I include research undertaken by team scientists for elite and professional teams, by sports industry scientists, scientists involved with broader sports organisations (such as FIFA), and by university research teams, as parts of sports science. The research goals of these areas will often be different. Consider, first, team scientists. There is a good deal of literature that discusses some potential benefits of utilising N of 1 trials for elite athletes (see for example: Jukola, 2019, Sands et al., 2019, Kinugasa et al., 2004, Kinugasa, 2013). There are a number of reasons why N of 1 trials hold a particularly special place here when compared to medicine, and especially when compared to group RCTs. As has been mentioned, elite athletes are both rare, and are likely to respond to interventions differently to non-elite athletes. This means that group RCTs conducted on other populations may not apply to them. Further to this, it may not be possible to have elite athletes in sufficient number to be able to conduct high-powered group RCTs on them. In addition to this, the interventions that may be under investigation by team scientists on elite athletes may be both novel, and secret, in order to attempt to gain a competitive edge. In this area, researchers can neither conduct large scale, high-quality group RCTs in order to gain good evidence, nor do they necessarily care about the effects of an intervention outside the athletes under their charge. As such, N of 1 trials have a pragmatic virtue over group RCTs: they are a type of trial it is both possible to conduct, and possible to conduct privately. Conclusions drawn from these trials will likely lack generalisability (see: section 3.6), but this will likely not be of concern to team scientists. In medicine however, even in the case of rare diseases, where group RCTs can be difficult to conduct, and N of 1 trials may be employed in their stead, there can be an assumed shared goal amongst researchers to provide evidence for a treatment or intervention's generalizable effects, even outside those currently in need of treatment. There is an assumed duty of care for medical researchers to publish evidence relevant to intervention effects that may help others. This means that the difficulty of generalising conclusions in medicine is problematic. There are also good reasons not to conduct research in secret in medicine: funding, fame, publishing, improving population health, etc.

Other than establishing efficacy, one key reason why we want to be able to draw general conclusions about interventions is to establish the harms associated with that intervention. In the following paragraphs, I argue that the potential harms associated with interventions in sports science are likely less critical than those associated with interventions in medicine. I call this the **Harm Profile Thesis** and refer to it again in later chapters. The relative lack of harm associated with interventions in sports science, compared to medicine, means that we need not be as concerned with finding out about, and generalising, the harm profiles associated with sports science interventions. This means that where in medicine we may need to conduct large scale group-RCTs to uncover harm profiles, this is not as pressing in sports science. What this shows is that, whilst being unable to generalise conclusions is a limitation of N of 1 methodology, it is not a limitation that is as important in sports science as it is in medicine. Accordingly, whilst evidence from N of 1 trials will not provide strong evidence that an intervention will be effective outside a trial participant, if we did happen to adopt an intervention on the basis of evidence from N of 1 trials, we need not be as worried about potential harms in sport as we would be in medicine. Further, we may conduct N of 1 trials in sport without being as concerned as we may be in medicine that preliminary trials intended to uncover harm profiles have not been conducted.

Allow me to explain. Health related harm profiles associated with sports interventions will generally be lower than those associated with medicine. The harm profile associated with an exercise intervention is likely less dramatic than that associated with taking a medication, and many interventions, such as injury prevention programmes, actually have a reduced harm profile when they are used compared to when they are not. For instance in football, some injury prevention programmes are very effective in some populations for reducing hamstring injuries (Soligard et al., 2008), and in rugby, the use of injury prevention programmes can have a dramatic effect on the incidence of spinal injuries (Quarrie et al., 2007). To put it broadly, it's unlikely that many sports interventions, such as exercise interventions, will have critical harm profiles. Conversely, as Stegenga (2018, 144, and throughout) argues, many medical interventions have particularly troubling harm profiles, and these are often underestimated. Given the potential risk of employing medical interventions, it is important to have potentially generalizable evidence about the risks of an intervention. Having group RCTs or collecting group data for medical interventions can provide some sort of evidence to this effect that may not be as necessary in the case of sports science interventions given their less extreme harm profiles. This type of evidence is difficult to obtain from N of 1 trials.

I will illustrate this with an example of preventative interventions in both sports science and medicine. For sports science, I will consider a hamstring injury prevention programme for football, the FIFA 11+, that I will dedicate the entirety of chapter 6 to later in the thesis (Soligard et al., 2008). For medicine, I will borrow some examples of drugs with serious harm profiles from Jacob Stegenga's book Medical Nihilism (2018). In this book, Stegenga actually argues that we regularly underestimate the harm profiles associated with drugs. This is in part, because investigations into the harms caused by drugs often fish for results that indicate them not being harmful, in order to be able to get through regulation procedures and sell more. Consider rosiglitazone, marketed as Avandia, a best-selling medication used in the treatment of diabetes. The harm profile associated with Avandia includes an increased risk of cardiovascular disease and death (Stegenga, 2018, 137). In a large-scale trial, performed by the manufacturer of Avandia, the outcome measures, all hospitalisations and deaths from any cardiovascular causes, were chosen to be very broad. This was so that both control and Avandia groups would have large numbers of this outcome, watering down the chance of the trial showing any statistical significance that could indicate that Avandia was harmful (Stegenga, 2018, 137).

We can also consider statins, a drug used in the treatment of cardiovascular disease. Whilst a large percentage (around 90%), of the reported instances of harms from statins are now suggested to be as a result of the nocebo effect (Pedro-Botet et al., 2019), this does still leave a number that are not. The harm profile associated with statins includes: muscle symptoms such as myalgia and cramps, diabetes mellitus, memory loss, tendon rupture, decreased renal function, interstitial lung disease, lowered testosterone, and depression (Thompson et al., 2016). Rare or not, these are serious harms. For instance, muscle symptoms occur, observationally, at a rate of 7 to 29% (Laufs et al., 2017) and 1 in 50 people develop diabetes as a result of statins (the NNT, 2013). Because statins are given to patients as a preventative measure, one cannot know if an individual patient has had a heart incident prevented by the statins. Their effect can only be observed over a population. Interestingly, in the case of statins, in those with known heart disease, 1 in 83 people have their life saved as a result of the drugs and 1 in 39 have non-fatal heart attacks prevented (the NNT, 2013). This puts the odds of getting diabetes from statins close to the odds of it having a beneficial effect. As parts of this harm profile are rare, it is important to obtain large-scale data to inform risk.

In contrast to this, and as an illustration of the lack of serious harms from many sporting interventions, a systematic review and meta-analysis of the FIFA 11+ including over 7,500 participants found only one report of adverse effects, a hamstring strain (Thorborg et al., 2017, 569). A relatively non-serious harm profile, in comparison. So, with the caveat that exceptions may exist in extreme cases like massively abnormal dietary interventions, as adverse outcomes of interventions in sport are unlikely to have such serious harm profiles associated with them, the importance of conducting group trials to find these harm profiles is lower. This means that N of 1 trials may be conducted more feasibly in sport.

As well as medicine facing the problem of relatively harsh harm profiles, there may also be areas of sports science where it is more feasible to conduct research with a view to generating individual treatment effects (ITEs) than it is in medicine. We can, once again, make the comparison between the feasibility and utility of conducting N of 1 trials in medicine and high level sports science. Whilst general research funding in the sports sciences remains low, high-level professional sports practitioners likely have the time and money to conduct N of 1 research with a view to optimising an athlete's training. In contrast, most medical professionals, such as General Practitioners, will likely not have time to conduct N of 1 style research in a systematic and controlled manner for every patient, particularly not to the thorough degree that high-level coaches in sport may. A high-level coach needs an understanding of the effects of an intervention in individual members of their squad, not average effects in a different population. This is unlike some examples that may be seen in medicine where treatments are prescribed widely, given their effectiveness in a group, even if they may not have a positive ITE for every member of that group.

Statins are prescribed in order to help prevent cardiovascular disease (CVD), and are prescribed for patients who are over 75, or are under 75 but have a 10-year risk of CVD of 10% (Ricciardi et al., 2019). Given that statins can be prescribed when CVD risk is only 10%, and that, according to Ricciardi et al., practitioners often prescribe statins when there is no CVD risk, a practitioner cannot always know if it will ever have a positive treatment effect in any individual patient, never knowing if a patient would necessarily have contracted CVD if they didn't receive statins. In fact, according to thennt.com, a group that collects information about drugs to find out how many people have to have that drug for one person to have its putative effect, it requires a group of 104 people with no known heart disease to take statins before one of those people is prevented from having a heart attack (the NNT, 2013). In addition to this, it would take an infinite number of people with no known heart disease to take statins before one person has their life saved from a heart attack by the statins (the NNT, 2013). Further to this, in people *with* known heart disease, to prevent one death from heart attacks, 83 people must take statins (the NNT, 2013). As has been stated, statins are prescribed because they have a population-level benefit. It is not feasible to determine if they have an individual benefit in every case. The benefit of an N of 1 trial, in sports instances then, is that relevant outcome measures of an intervention can be measured in individuals to evidence where there is a positive benefit. Whilst the incorporation of a non-beneficial intervention in sport may be less harmful than say, risking diabetes with statins, if an athlete has limited time and energy in the day to train, it is likely that a coach would not want them to engage in an intervention that is effective on a group level if no benefit is observed in that individual.

I would like to make two small points relating to what I have said in this section. First, it must be noted that unsystematic single subject trials are often conducted in medicine, between practitioners and patients. This can involve, for instance, trying various medications for chronic illnesses and assessment by practitioners and patients to determine which appears to be the most effective. This, however, is done after group trials have been used to indicate that an intervention is relatively safe (although Stegenga would dispute whether they actually do manage this (Stegenga, 2018)). Given the lack of extreme harm profiles in sport, the necessity of waiting for group trials before conducting an N of 1 trial is not so important. The point here is that, because of the danger of utilising medical treatments, this type of trial is very rarely suitable as primary research, where in sport it may be.

Second, it may be slightly misleading to say there is no harm profile associated with sports science interventions. Depending on an athlete's proximity to eliteness, they may view using a useless, or even less than optimum intervention as harmful. For instance, if one intervention an Olympic athlete could perform would take 0.2 seconds off their 100 m time in a year, and others would only take off 0.1 seconds over the same period of time, they may view the second intervention as harmful. If the difference between an Olympic medal and fourth place in an Olympic final is 0.1 seconds for that athlete, they may view the less optimal intervention as particularly harmful. However, the number of athletes or people engaging in sport that will view harms to performance like this in the same way they may view harms to health from medical interventions is very small. Thus, in general, the point about sports interventions having small harm profiles in comparison to medicine still stands.

3.5 Where can N of 1 trials provide better evidence than group RCTs in sports science?

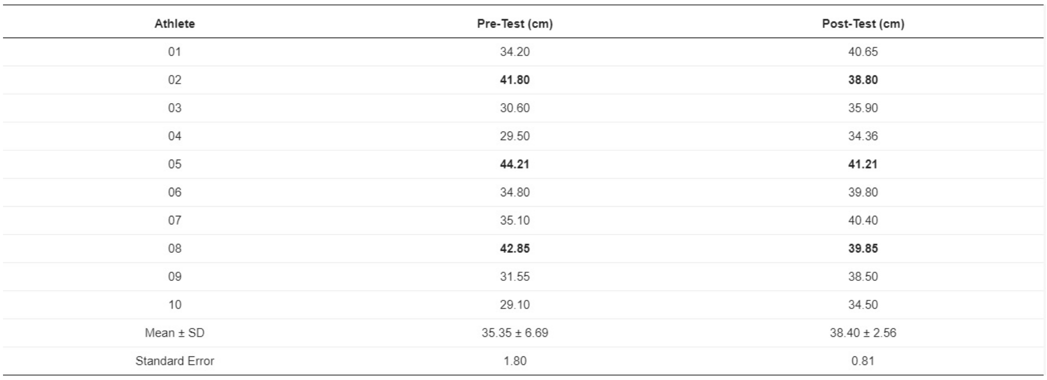
In this section, I will argue that N of 1 trials may produce higher quality evidence than group RCTs for individuals in some instances. This is because, in some instances, they can: provide evidence for claims about ITEs rather than average effects in groups; overcome limitations to evidence due to small sample sizes where group RCTs cannot; have a more representative sample where the target is the trial participant; and because they can be used to personalise interventions.

3.5.1 Individual treatment effects

As has previously been explained, where a group RCT can provide evidence for average intervention effects in a group, an N of 1 trial provides evidence for ITEs. In the case of individual sports performance, and in elite athletes in particular, this is very useful as practitioners are often interested in improving the results of the few under their charge. This is unlike parts of sports science concerned with making general conclusions about intervention effects, such as areas that inform public policy.

If sports science group-trials only report average effects, even if conducted in the same cohort in which an effective intervention would then be applied, they risk losing important ITE information (Kwasnicka and Naughton, 2020, 2). For instance, a sports performance intervention may show an average positive effect that reaches statistical significance in a group trial, but an average positive effect does not rule out the possibility of negative effects in individuals, obscured by the average. Sands et al. (2019), highlights this with a hypothetical trial example in which a coach employs a strength training intervention to improve jump height for a cohort of athletes. The table giving the data about jump heights pre- and post-intervention is replicated from Sands et al., 2019 in Figure 3.4. Across the cohort, there is an 8% increase in average jump height after the intervention, which reaches statistical significance (Sands et al., 2019, 4-5). However, the highest performing members of the cohort before the intervention show a decrease in individual jump heights post-intervention, which cannot be seen if only group averages are reported. If participants are treated individually within the trial, a negative performance effect, or no effect over the course of that intervention, may be seen. This illustrates that a practitioner trying to inform their practice based on group trial evidence may see that an intervention is highly effective on average, but not note that in trials there were some for whom the outcome measures were worse after adopting the intervention, or they may not know if it will be harmful to the individuals they wish to apply it to. A coach would not want to adopt an intervention for every member of their cohort that was on average beneficial, if it would harm the performance of, say, the only medal hopeful in a cohort (Sands et al., 2019, 4-5). If, however, a trial of the same intervention was conducted on that same cohort, and ITEs were established, a coach may then prescribe that intervention in individuals where positive effects were observed, and apply different or altered interventions for those where no benefit or negative performance effects are seen. As N of 1 trials can give rise to evidence about ITEs where group trials cannot, this supports a benefit of N of 1 trials over group RCTs.

It could, of course, justifiably be raised that conducting a group RCT,



Athlete	Pre-Test (cm)	Post-Test (cm)
01	34.20	40.65
02	**41.80**	**38.80**
03	30.60	35.90
04	29.50	34.36
05	**44.21**	**41.21**
06	34.80	39.80
07	35.10	40.40
08	**42.85**	**39.85**
09	31.55	38.50
10	29.10	34.50
Mean ± SD	35.35 ± 6.69	38.40 ± 2.56
Standard Error	1.80	0.81

Figure 3.4: Theoretical jump heights for athletes pre- and post-intervention. An average jump height improvement of 8% is seen in group data, but in bold are athletes who showed a decrease in jump height post intervention (Sands et al., 2019, 4-5).

and then performing subgroup analysis, could also give rise to information about negative effects of interventions. In medicine, for instance, subgroup analysis has been employed in the evaluation of the use of antihypertensives in order to establish which ethic subgroups should be prescribed different treatments for high blood pressure (Clarke et al., 2014, 347). In Figure 3.4, there is a subgroup of three athletes who responded negatively to the intervention. Subgroup analysis may also highlight subgroups that show varying degrees of positive response that may be identified. The relevance of this information may be questionable, however. Given the deficiencies of group RCT data already discussed, it is not certain that subgroup analysis can be extrapolated outside the cohort on which it was conducted to other cohorts. Particularly given the small size of many sports science RCTs, effectively identifying criteria that mark out a subgroup may be difficult, particularly when compared to large scale medical trials. If, however, subgroup analysis is conducted in the target cohort, the information is, of course, relevant. If research is being conducted on the target cohort, however, and there are enough data and resources to conduct subgroup analysis, it may be more beneficial to treat trial data individually, as if from concurrent N of 1 trials, and obtain ITEs rather than averages across subgroups. This may help better generate, for instance, predictions for future performance in individuals.

Following this previous example, another benefit of N of 1 trials over group trials may be seen. As well as being used to establish ITEs of specific interventions, N of 1 trials can be used to provide evidence for the effects of personalised interventions, as is suggested by Guyatt et al. 2000, Kwasnicka et al., 2017, and Ordovas et al., 2018. This is unlike group RCTs where, in order to control the trial effectively, participants receive the same intervention (unless it is a rare case where the intervention being measured is 'the effects of personalised interventions for X'). Competitive sports-people want the competitive edge, especially at high levels, and the ability to change or adapt interventions as N of 1 trials progress, and measure the different rates of change to outcome measures can benefit this. For instance, the use of N of 1 trials can be used to help develop and measure the effect size differences for individualised nutritional plans that account for physiological and psychological differences between individuals, such as food preference, intolerances, and responses to certain nutrients (Ordovas et al., 2018, 2).

3.5.2 Sample size and relevance

In the previous chapter, I argued that a problem seen in group RCTs in sport is that small sample sizes are a common limitation to evidence quality in sports science, sometimes unavoidably. Although not exclusive to this area, elite athletes, who are rare by definition will, unavoidably, present small sample sizes. Reaching sample sizes where small to moderate effect sizes in elite athletes may be seen, and chance can be ruled out as a cause of effect size in trials can be impossible. However, in cases where sample sizes are small, N of 1 trials with repeated assessments may be beneficial. Kwasnicka and Naughton (2020, 2) argue that an N of 1 trial with 50 to 300 assessments in an individual is comparable to a group trial with 50 to 300 participants. It is far more feasible to conduct an N of 1 trial with 50 assessments on an elite athlete than it is to find a cohort of 50 representative elite athletes within whom to research the same intervention.

A further problem with group sports science RCTs can be illustrated with the example of elite athletes. Some demographics are better represented than others in sports science research, and evidence for an intervention's effectiveness may not apply across groups. Just as Cartwright and Hardie (2012) argue in the case of RCTs for public policy, evidence from an RCT that something is effective in one instance does not tell you that results of that RCT are applicable elsewhere. This is the case for sports science. For instance, many argue that even a high-quality group RCT performed on non-elite athletes may not provide good evidence for effectiveness in elite athletes (Jukola, 2019, 5, Sands et al., 2019, 2, Burke and Hawley, 2018, 785). This is due, in part, to differences in psychological and physiological response, and differences in training (Jukola, 2019, 5). Particularly troubling, when trying to inform practice from group RCT evidence, is that there is seen in sports science, a problem where non-elite athletes are wrongly labelled as being elite in trials, further confusing efforts to inform practice (Sands et al., 2019, 2). However, this problem can be avoided by conducting N of 1 trials on individual elite athletes. By doing this, a practitioner no longer needs to worry about the applicability or extrapoloability of evidence from non-elite group RCTs to their athletes, as they will have evidence about the specific individuals they are concerned with and will not need to extrapolate. N of 1 trials allow researchers to obtain evidence about effectiveness in the individual in whom the intervention may be applied and see the ITEs of that intervention. This is particularly useful as, unlike group RCTs, one does not have to worry about whether the athlete in question is predisposed to have a greater or lesser outcome as a result of the intervention. One does not need to worry about balancing these factors across groups. Further to this, where practice evidenced by group RCTs may seem to be effective under controlled RCT settings, it may not be in real life settings. Unlike this type of RCT, an N of 1 trial, however, can show effectiveness in individuals in real-world settings (Kinugasa, 2013, 158). This does not, however, mean that evidence collected for one elite athlete in a given field will apply to any other elite athletes in this field, which brings us to the problem of generalising.

3.6 Evidence limitations: generalizing and making general conclusions

When a causal claim is generalizable, we mean that the same causal relationship operates in other, similar situations. There are clear scenarios where researchers can conduct N of 1 trials within the target cohort and not worry about whether general conclusions can be drawn, or whether research is relevant to other cohorts. These are the areas where I have, so far, intended to argue that N of 1 trials may provide more relevant evidence than group RCTs. Whilst philosophical debate exists questioning the practical limitations associated with generalising of some types of causal claims in sport, particularly in identifying what factors are relevant to generalisability (see for instance: McFee, 2009, Chapter 1 and throughout), general claims are regularly made. Generalisability, for instance, is important for drawing general conclusions about effects, generating theories that explain effects, and for furthering the knowledge set in sports science as a field, but is also important for research, funding, and publishing. University researchers, for instance, will likely care about drawing general conclusions about effects, unlike team scientists who may only care about effects in a few athletes. An exercise physiology textbook that only explains the biological workings of a single individual is not a terribly useful tool. Unfortunately, conclusions drawn for particular athletes in N of 1 trials should not always be considered to be relevant to other athletes (Kinugasa et al., 2004, 1047). For instance, in a review of the quality of evidence in sports nutrition literature for elite athletes, Jukola (2019) concludes that N of 1 trials may be a good tool to develop nutrition plans for specific elite athletes, particularly when compared to group trials, but does not conclude that evidence for effectiveness in a single athlete can be extrapolated to other groups.

Although single subject research has been conducted in a number of disciplines, there is a paucity of papers published using it, which may be due, in part, to difficulties in publishing and funding. For instance, a 2004 review of single subject trials found that in 30 years of behavioural sports psychology research, only 40 published papers met simple inclusion criteria including: using competitive athletes, measuring performance data, using reliable data gathering methods or objective statistics (Martin et al., 264-265). Further to this, in an overview of published sports performance research, Kinugasa (2013, 165) also comments that N of 1 trials are 'rare' in that field. This could be obscured by results being kept secret where a competitive edge is desired, however.

Kinugasa (2013), however, suggests a solution to the problem of generalisability. It is suggested that meta-analysis could be used in order to collect separate single subject data and, from it, produce *magnitude based inferences* of effects in order 'to establish generalizability of the findings' (164). Similar has also been argued by Sniehotta et al. (2012), and Araujo et al. (2016). There are problems faced by magnitude based inferences as an analysis tool. This thesis will not explain them in detail, but they include problems such as misinterpreting frequentist statistics, making overly optimistic inferences, making claims that can only really be made by Bayesian statistics as priors are bypassed, and others as explained by Welsh and Knight (2015) and Sainani et al. (2019). Even disregarding this, evidence from a small number of N of 1 trials may not be sufficiently strong to generalise conclusions from. Combined results from N of 1 trials on the same intervention will still likely suffer from the same small sample size problems that groups trials, and meta-analyses of these trials do. Thus, evidence from proposed meta-analyses may not be generalisable. This is because it will likely be just as difficult to find a large number of people of a suitably similar reference class to perform well controlled N of 1 trials on, as it is to find the same number of people to perform a group RCT on. They will likely not provide sufficiently strong evidence to overcome the fact that the mechanism by which an intervention has its putative effect in one group may be different, may interact with another mechanism changing its net effect, or may be over-determined such that the outcome would occur even without a new intervention in another population. I will come back to the use of mechanisms as an aid to extrapolating between groups in section 8.5.

As a final note to this, in areas where generalizing is important, I contend that N of 1 trials may still have a use, even if limited in scope. Just as in medicine, primary trials are conducted before moving to larger scale trials, conducting N of 1 trials as a preliminary to group trials may be beneficial in the sports sciences. Not only may this provide motivation and evidence that further research may be done in order to generalise conclusions, it may also be a way for researchers to test and improve experimental design and methodology.[2] This is particularly the case given the lack of danger participants of N of 1 trials face from harm profiles, as argued earlier.

3.7 Evidence limitations: inadequate placebos

So far, I have argued that N of 1 trials provide benefits over group RCTs in that the sample is representative, small sample size problems may be overcome, interventions may be personalised, and ITEs obtained from N of 1 trials may be better to inform practice for an individual than

[2] I can attest, personally, to the importance of this. I was a test subject in the early stages of a sports science trial that required immobilising the wrists in order to test grip strength. In early stages of the trial this was done using over-tightened, budget, dog collars that left welts on the wrists. My complaints helped the researchers improve their research design to medical wrist restraints that caused much less chafing.

measures of average outcomes ascertained from group trials. A problem seen in group RCTs still looms large, however. A key element of the **Excluded Explanations Argument** was the difficulty of placebo controlling in sports science. This can lead to misestimations of the effect sizes of new interventions when compared to placebos or active controls. This is, of course, a problem for N of 1 trials, also. If the same type of interventions are being researched, why would a placebo be more adequate in an N of 1 trial than a group trial? In this section, I will argue that as placebo controlling may be as difficult in N of 1 trials, particularly when used in order to aid blinding, and that using current practice as an active control does not solve this, evidence quality is still limited in N of 1 trials.

It is possible to conceive of N of 1 trials as eluding the placebo problem, unlike group RCTs, as follows:

> Group RCTs provide evidence for the average effects of an intervention, in comparison to the effects of a placebo or active control, across a group. However, regardless of whether the control is adequate, the comparison of effect size between a group trial control and intervention does not necessarily have bearing on the effect size of an intervention compared to an individual athlete's current prescribed or quotidian practices. After all, if an athlete wants to improve an outcome enough to seek better interventions, it is likely they are already engaging in an activity that they hope will improve it. By calculating baseline measures for a trial around current athlete practices, evidence for an effect size in comparison to current practice can be generated. So, whilst this may not provide widely generalizable evidence for effect size in comparison to a true neutral placebo, evidence can still be obtained in order to help determine if a new practice is more effective than an individual's current practice. From the potential for in-

> adequacy in comparison between intervention effects with a placebo, and the benefits of controlling using current practice in N of 1 trials, the lack of necessity of placebos in N of 1 trial can be seen. In an outcome driven field like sport, it may be more useful to have evidence that an intervention is associated with a positive difference in effect size than current practice in an individual, than that some intervention is more effective than a placebo in a group trial.

The reasoning behind this line of thought can be illustrated with the rotator cuff strengthening intervention described in section 3.3:

> Whilst not always the case, baseline or control measurements can be taken using current practice as an active control, allowing evidence to be gathered to produce estimates of difference in outcome measures between an individual's current and new practices. For example, in the Bernhardsson et al. case, where exercise was used to treat shoulder pain, an inclusion criterion for the shoulder pain was that it must have been chronic. This was to allow a baseline pain level to be established. Before the introduction of the intervention, participants were asked to act as normal for three weeks, whilst rating perceived pain, to establish this baseline (2011, 69). This means that any difference in outcome measure seen over the course of the B phase in individuals was attributed to, in part, the intervention. In this case, the N of 1 trial seems to be able to provide evidence, for individuals, that exercise regime improved outcome measures, when compared to pre-trial activity. Here, outcome measures were taken and compared to a baseline without the need for a placebo. Instead, evidence was gathered for the effects of an intervention in an individual, in a practical, real-world setting, in comparison to normal, pre-intervention practices as the active control.

However, this response to the problem of inadequate placebos in sports science that attempts to absolve N of 1 trials is flawed. One purpose of placebo controlling in a trial is to help rule out non-characteristic intervention effects on outcome measures, as was explained in chapter 2. There may be effects that influence outcome measures due to simply being in a trial, or using a novel intervention, for instance, which need to be controlled for. In an ideal situation, with a placebo that has none of the characteristic features of the intervention under investigation, in the placebo wing of a trial, all effects on outcome measures over the trial *should* be as a result of trial effects, and incidental features of the intervention. This means that, as both control and intervention phases are subject to the same trial effects, differences in outcome measures between control and intervention phases *should* provide evidence for difference in effect sizes regardless of trial effects where an ideal or adequate placebo is used. However, where the influence of these trial effects on outcome measures cannot be mitigated, where all differences in outcome measure between groups, or phases, can be explained by non-characteristic effects of the intervention, this limits the quality of evidence these trials can produce.

In addition to difficulties involved with placebo controlling N of 1 trials, blinding will also be difficult and can reduce the quality of evidence trials can produce. This issue, and its relevance to the **Excluded Explanations Argument**, are explored in detail in chapter 2. I suggested that blinding and placebo controlling in the sports sciences may be, in some instances, near impossible due to the active involvement of participants and practitioners in the trial. For a brief, simple example, an athlete likely knows when they are undergoing a novel training regime, such as may be the intervention in a B phase of an N of 1 trial. Rather than as a useful tool to excuse the need for placebo controlling in an N of 1 trial, active controlling may contribute to de-blinding trial participants. This means that using current practice may exacerbate the likelihood for an overestimation of intervention effects seen when comparing outcomes in A and B phases. Introducing a novel intervention in a B phase of a trial may, for instance, have psychological effects on that athlete that influences outcome measures. In an exercise intervention instance, for example, it would be difficult to rule out the effects intervention novelty may have on outcome measures in a B phase, if the active control used was current training practices. An athlete may approach a new training programme with more zeal than their current one, artificially inflating outcome measures. As novelty will wear off, and other trial effects will not be present when an intervention is used outside a trial, if a trial is insufficiently blinded to rule out these effects on outcome measures, this can lead to an overestimation of effect size. This, of course, refers to interventions that are employed over a long duration. Improving outcome measures acutely, such as on race day, with a placebo effect may be particularly useful in practice. However, trials to examine these effects may be difficult to carry out.

So, whilst the potential response to the problem of placebos in N of 1 trials claims that controlling with current practice means that a placebo may not always be necessary, this response does not work. Active controlling using current practice means that that differences in outcomes measures between baseline and intervention phases will not rule out trial effects as an explanation for that difference in outcome measures. The potential presence of trial effects outside of intervention effects raises a concern about the quality of evidence that an active control may produce when they are used without adequate blinding, as may be the case using current practice as an active control.

Therefore, where a trial is insufficiently blinded, the effects due to knowing one is in a trial, and engaging in a novel practice, may impact outcome measures in a B phase of a trial, but not in an A phase. This means that a comparison of outcome measures between control and intervention phases will not provide evidence for the effects of only the intervention in relation to the control if those effects are not also present in the A phase.

Also, where psychological effects due to novelty and knowledge that one is in a trial do not continue with the use of an intervention post trial, an N of 1 trial may provide misleading evidence that an intervention is, indeed, better than the active control. These preceding points lead us to the notion that the conclusion that may be drawn in an active controlled N of 1 trial, where the active control is normal practice, is simply that: when intervention B is utilised in trial conditions and where it is a novel intervention, it is associated with a change in outcome measures in comparison to baseline active intervention A which is not novel.

The shoulder pain example can be used to explain why controlling using current practice will not absolve an N of 1 trial from needing to placebo control. There may have been non-intervention-specific effects at play in the B phase of the trial in addition to intervention effects. Perhaps, for instance, knowing that one is undertaking shoulder exercise may mean that participants try to return to more authentic shoulder movement when they would not otherwise, which may have an effect on shoulder health in addition to the specific effect of the exercises. In addition to this, the belief that one is undergoing what should be an effective treatment for shoulder pain may have an effect on outcome measures. Therefore, differences in outcome between A and B phases may evidence a more tentative conclusion than *'these exercises are associated with a reduction in shoulder pain'*. Instead, they evidence that *'being prescribed this intervention under trial conditions is associated with an improvement in shoulder pain'*. This means that effects that are attributed to the intervention itself, and not trial conditions, may be overestimated. So, difficulties in blinding using current practice as an active control means that, in fact, evidence provided by this type of active control may be flawed.

This means that, whilst N of 1 trials have some benefits over group RCTs for individuals, they both still fall foul of problems related to controlling both with active controls and with placebos.

However, all hope is not lost for N of 1 trials. Where two or more novel interventions are used in a trial, such as in an ABC trial, all novel interventions will be subject to similar psychological and other trial condition effects, so, comparisons between them should go some way towards providing evidence for the effect of each in relation to the other, regardless of trial effects. The distraction and awareness techniques trial for indoor rowing performance provides an example of this. Both awareness and distraction techniques were employed under the same trial conditions, so the outcome measures obtained in B and C trial phases may be compared with reasonable justification. However, comparisons between B/C phases and the A phase are still problematic. Conclusions were drawn from comparisons between B and C trial phases for individuals, but, due to problems regarding intervention novelty, whilst being in a trial may encourage a slightly better rowing performance than normal in an A phase, comparisons between B/C phases and the A phase may overestimate treatment effects. Problems with active controls discussed in chapter 2 apply here, also. As is the case in group trials, N of 1 trials may not be sensitive enough to determine if a B or a C intervention is more effective, particularly if effect sizes are small or similar. Further, the evidence they provide may still not be sufficient to determine if either intervention is more effective than an adequate placebo if one cannot be used.

3.8 Conclusion

In this chapter, I introduced N of 1 trials, which have been suggested by some as a potential solution to evidence gathering issues faced by group RCT trials in the sports sciences. I argued that these types of trials do have benefits over group trials in some instances. For instance, they allow for the generation of evidence for ITEs, which is important in some areas of sports science, particularly for elite athletes. This can help to overcome issues with generalising results from group trials to individuals, where an N of 1 trial is employed on the same person who the intervention is intended to be used on. However, this came with some caveats. Firstly, the results of N of 1 trials lack generalizability, and even conducting meta-analyses of their results does not remedy this. Further, N of 1 trials suffer from the same issues with placebo controlling and blinding as group trials. This will mean that N of 1 trials are just as susceptible to elements of the **Excluded Explanations Argument** as group trials. The outcome, then, is that N of 1 trials both lack scope in applicability, and still face issues in the quality of evidence they can produce even for individuals. N of 1 trials will often, then, not provide strong enough evidence to fulfil the goal of relying on the best possible evidence.

In the following chapters, I bring in my argument in favour of the **Better Evidence Thesis** in sports science: that we rely on better evidence if we include the assessment of evidence from mechanistic studies as well as association studies, like RCTs, when assessing causal claims and informing practice. This involves, in the next chapter, introducing and defending the RWT, before arguing for its applicability in the sports sciences. Then, in chapter 5, I use the RWT, and examples from medicine and sports science to argue that the assessment of evidence from mechanistic studies in the sports sciences better helps us to establish causal claims.

Part II

Some solutions

Introduction to Part II

Part I of this thesis argued that group RCTs and RCT-style N of 1 trials in sports science often provide low-quality evidence. This challenges the idea that they should be privileged in sports science, particularly at the expense of other types of study, like mechanistic studies. Given the challenges to the notion that relying on evidence from RCTs well fulfils the EBP goal of relying on the best possible evidence, Part II of this thesis looks at how we might do things better. Chapter 4 of this thesis looks at what we actually need evidence of to establish causal claims, arguing in favour of the RWT: the idea that we need to establish a correlation and a suitable mechanism in order to establish a causal claim. Given this, chapter 5 looks at what we ought to do in order to best be able to provide evidence for both of these things. Chapter 5, motivated by EBM+ and the practice of IARC in medicine, argues that in sports science we would better fulfil the EBP goal of relying on the best possible evidence if we assessed evidence from mechanistic and association studies together when assessing causal claims than if we assess evidence from RCTs alone. This is the **Better evidence Thesis**.

Chapter 4

A defence of the RWT

4.1 Introduction

Establishing causal relationships is, of course, key to EBP. We generally want to know something works before we employ it in practice. However, given that I have argued for the general weak nature of evidence that may be derived from RCTs in the sports sciences, if causality is to be established, something other than a reliance on evidence from RCTs must be utilised if we want to rely on the best possible evidence. In the next chapter, I introduce the work of the Evidence-Based Medicine Plus (EBM+) group and argue for the applicability of this work to sports science. EBM+ argues that, when assessing causal claims, we should analyse both evidence from mechanistic and association studies. The work of the EBM+ group relies on the Russo-Williamson Thesis (RWT). This is the thesis that in order to establish a causal relationship in medicine, one needs to establish both the existence of a correlation, and the existence of a mechanism. This will be explained thoroughly in section 4.2. In order to argue for a new evidence assessment modality in sports science, based on EBM+, I must, therefore, provide a defence of the RWT. This chapter also uses case studies from sports science that help to exemplify how the

RWT applies to sports science. The example of creatine supplementation provides us with a historical case in sports science that conforms to the RWT. The example of caffeine supplementation provides us with an example that shows how association studies in sports science may also provide evidence of mechanism, just as I argue they may do in medicine.

In this chapter, I offer a brief overview of five key criticisms of the RWT that exemplify those in the literature. I will both provide original rebuttals to these criticisms, and draw on the work of others. The intention of this is to highlight that the RWT can handle key criticisms. Further, the intention is to show that in order to handle criticisms, it must be possible that the existence of a mechanism can be established based on evidence from association studies. This is because some criticisms of the RWT object that causality has been established in medicine without the need for mechanistic studies. Association studies include trials such as RCTs, and observational studies. Williamson (2019), provides an inference intended to show how these types of trial can provide sufficient evidence to establish the existence of a mechanism. Part of the work of this chapter will be to argue that Williamson's inference successfully lays out criteria which, when met by association studies, are sufficient to provide evidence that can be used to establish the existence of a mechanism.

4.2 The RWT

The RWT is an epistemological thesis concerned with causality. It is intended to be both descriptive and normative (Williamson, 2019, 34). The RWT is informed by historical cases, but also has strong epistemic rationale supporting it. The RWT was initially stated as:

> 'To establish causal claims, scientists need the mutual support of mechanisms and dependencies. The idea is that probabilistic evidence needs to be accounted for by an underlying

> mechanism before the causal claim can be established' (Russo and Williamson, 2007, 159)

The thesis is intended to be descriptive in that it aims to explain how causation is established in the health sciences. It is also intended to be normative in that it sets out epistemic criteria that *should* be met before something can be established. Central to the thesis is the idea that each of, both evidence of correlation, and evidence of mechanism, when assessed separately, are flawed as justifications for inferring causation (Williamson, 2019, section 1). These flaws mean that each type of evidence, alone, is insufficient to establish the existence of a causal relationship in the health sciences. I will explain this in more detail in the next chapter, where I introduce it as a key concept that is part of EBM+ methodology. Another key concept in EBM+ methodology, and one also propounded by Williamson, is that each kind of evidence, evidence of mechanism, and evidence of correlation, *cover* the flaws of the other type of evidence (2019, section 1). I will briefly explain what this means at the end of this section. First, I must briefly say what I mean when I talk about mechanisms, in this context.

In this thesis, I use the mechanism definitions adopted by the EBM+ group. One type of mechanism, described by Illari and Williamson (2012) is that a complex systems mechanism for a phenomenon consists of entities and activities organised in such a way that they are responsible for the phenomenon. The other type of mechanism they put forward is a mechanistic process. This: 'consists in a spatio-temporal pathway along which certain features are propagated from the starting point to the end point' (Parkkinen et al., 2018, 13). They attribute this view of mechanisms to Salmon (1998).

Since the original statement of the RWT, work has been done by Illari (2011) to further clarify it. Illari set out the distinction between mechanistic evidence as type and object of evidence. This means there is a distinction between meaning mechanistic evidence as evidence from mechanistic studies, and mechanistic evidence as evidence that a mechanism exists. She also suggested four different levels of demandingness we may have for evidence of mechanism as the target: '1) evidence of mechanism in detail, 2) evidence that there is a mechanism of the postulated kind, 3) postulated mechanism, based on evidence of analogous mechanisms, and 4) evidence that there is no mechanism' (P. Illari, 2011, 151). Type 1 is the most demanding, but Illari suggests that this is not a 'simple continuum of decreasing amounts of evidence of a mechanism' (2011, 151).

In response to this disambiguation between evidence of mechanism as type and object, Russo and Williamson slightly reformulated the RWT to account for ambiguity in interpretation (2011). They clarify the intended interpretation is that evidence of mechanism is the target, rather than type of evidence the RWT is concerned with. Russo and Williamson reinforce this by using the phrase 'evidence of mechanism' rather than mechanistic evidence (2011, 569). This means that, as they view it, what is important to establishing causality is evidence that there is a mechanism. It does not mean evidence that comes from mechanistic studies. In addition to this, they clarified the demandingness of evidence of mechanism; it should be evidence that a suitable mechanism exists. This better clarified explanation of the RWT is referred to as the disambiguated RWT in the wider literature. When, in this thesis, I refer to the RWT, it is the disambiguated statement of the RWT that I am referring to. In order to establish a causal relationship in medicine, the disambiguated RWT requires that the *existence* of both a mechanism and a correlation be established (Williamson, 2019, 34). This means that, according to the RWT, the size of correlation, or how the mechanism operates, do not need to be established before causality can be considered to be established. The disambiguated RWT, and the form of the RWT that I will be defending in this chapter, can be stated as follows:

> In order to establish a causal claim in medicine one normally needs to establish two things: first, that the putative cause and effect are appropriately correlated; second, that there is some mechanism which explains instances of the putative effect in terms of the putative cause and which can account for this correlation (Williamson, 2019, 33)

The fact that we need evidence of mechanism to cover the flaws in evidence of correlation, and evidence of correlation to cover the flaws in evidence of mechanism, is what provides epistemic rationale for the RWT (Williamson, 2019, section 1.2). Gillies (2019) refers to this ability of each type of evidence to cover the flaws in the other type of evidence as *strength through combining*, and attributes the idea to Illari (2011, 146). I will expand this idea in section 5.4, explaining how it relates to the idea of utilising evidence from both mechanistic and association studies, a view which has been called *reinforced reasoning* by Auker-Howlett and Wilde (2019). Allow me to briefly explain what it means for evidence of correlation, and evidence of mechanism, to each *cover the flaws* of the other type of evidence when assessing causality.

I will touch on evidence of mechanism first. If we establish the existence of a mechanism only, we are unable to establish causation because we will not know the net effect of that mechanism, or if it even has one. This is particularly the case as, often, when conducting mechanistic research, we examine mechanisms in isolation. As Illari presents it in the case of medicine, the problem is that though we may establish the existence of a mechanism, the complexity of the human system means that we may not know the extent of the effect that mechanism produces, or if another mechanism cancels it out on a more holistic scale. For example, whilst we know that there is a mechanism by which exercise can be conducive to weight loss by expending energy, it also increases appetite, which can curb weight loss by leading to an increased intake of energy. Thus, the mechanism by which exercise can cause weight loss is *masked* and, without further investigation, the overall effect of exercise on weight changes cannot be seen. Illari attributes her use of the exercise case to its discussion by Steel (2007, 68). Evidence of correlation provides evidence for the overall effect of mechanisms in the system. In this instance, evidence of correlation provides evidence for the overall effect of exercise on weight loss.

As motivated the **Excluded Explanations Argument**, observing a correlation between a proposed cause and intended effect is insufficient to establish causality, as the observed correlation may have other explanations. We must therefore have evidence for the existence of a mechanism between the two outcomes in order to rule in that a causal relationship exists between them. Evidence for the existence of a mechanism helps by telling us that there is a way by which the causal relationship in question can give rise to a correlation observed. So, where we have evidence of a mechanism, we know that there is something going on that explains why we can observe a correlation, and that we are correct in our assumption of the cause of the outcome.

4.3 Establishing

In order to go forward with my argument in this chapter, it must be seen what it is for a causal claim to be considered to be established. Given that I am advocating the adoption of the RWT in the sports sciences, I will adopt the view of establishing taken by Williamson (2019, 35-36), as that is the one with which his formulation of the RWT is meant to be taken.

The view of establishing that I adopt is as follows:

> ‘A causal claim is ‘established’ just when standards are met for treating the claim itself as evidence, to be used to help

> evaluate further claims. This requires not only high confidence in the truth of the claim itself but also high confidence in its stability, i.e. that further evidence will not call the claim into question.' (2019, 35)

This definition, in part, puts focus on what is required for a causal claim to be established on the practices of medicine and, in my case, the sports sciences. He explains this view may be taken without committing to a particular view of evidence, and whether establishing must be factive. This is useful as this question is too large for the scope of this thesis. This definition also relates back to the standards of evidence discussed in subsection 1.5.4. From this we can see that establishing a causal claim, as it requires a stable claim, requires high-quality evidence.

There are a number of views about what can be considered to be evidence (2015). A large part of this debate is whether evidence must be true. These views include evidence being characterised as knowledge, full belief, one's information, degrees of belief as informed by observation, and what one grants rationally. For instance, Williamson (2015) holds that as evidence is what one rationally grants, evidence need not be true. This can be contrasted with the view of Littlejohn (2012) who holds that evidence is factive. Littlejohn makes this claim on the grounds that evidence rests on the external, and true facts or propositions relating to this, where our beliefs can be 'non-inferentially justified' (Littlejohn, 2012, 89). This means that Littlejohn holds that for something to count as evidence, it must be true. It must relate to the way the world really *is*, instead of relating to what we think we are justified in saying about the world. Not commenting on whether evidence needs to be true means that I will leave open the question of whether establishing, which of course relies on evidence, must be factive.

Although this view of establishing allows us to be agnostic towards the factivity of evidence, in section 4.6, I argue that the main thesis of this chapter - that Williamson's inference that association studies can establish the existence of mechanisms - works whether one has a view of evidence that is factive or not.

4.4 Sports science example: creatine supplementation

It will be useful to give an example of something that is considered to be established by sports science, and how it came to be established. In this section, it will be seen that, in order to establish the effects of creatine, a mechanism and a correlation needed to be established. This case, then, provides us with a historical case of the applicability of the RWT to sports science.

Creatine is a well-known and widely used dietary supplement among many athlete populations. The effects of dietary creatine supplementation on sports performance outcomes are now considered to be well-established (Kreider, 2003, 1). The description of the effects and mechanism of creatine given in this paragraph are summarised from Woodruff, 2016 (149-151), a popular undergraduate sports nutrition textbook. The fact that it is discussed in textbooks is further evidence that it is established. Supplementing dietary creatine can cause improved athlete performance in a number of ways. Creatine is stored in the muscles and is readily available through meat consumption in a normal diet. However, supplementing creatine above what is available in a normal diet improves performance outcomes over a diet where creatine is not supplemented. Supplementing creatine into the diet allows one to consume an amount of creatine that would not be feasible using normal foods. It is established that supplementing creatine above what is normally available in a diet enhances athletes' performance in exercise that is performed for a short duration and at high intensity. Supplementing creatine above what is normally available in the diet works by increasing the pool of creatine in the body. This, in turn, allows the body to replenish stores of Adenosine Triphosphate (ATP) more quickly than it could without creatine supplementation. The body uses ATP as an energy source for high intensity, short duration exercise. Creatine supplementation also increases muscle glycogen stores, reduces muscle damage following exercise, and improves calcium re-uptake by the muscles. All of these effects can lead to improved performance in training and competition. This includes improving performance in sprinting events and weight lifting performances. Also, by improving sprinting and weight lifting performances in training, creatine supplementation can aid the training of athletes, allowing them to train harder and get fitter at a greater rate than if they did not supplement creatine. One key benefit of long-term creatine supplementation is that it aids athletes' abilities to gain fat-free mass in training, i.e. muscle that leads to performance outcomes. This allows athletes who compete in events that occur over a longer duration to still benefit from creatine supplementation, in a more holistic sense, as their training becomes more effective when creatine is supplemented. The effects of creatine supplementation on endurance sports, and sports which do not benefit from training over short a duration and at a high intensity are not well established. Creatine supplementation is also currently being investigated for its potential to improve other, wider health outcomes, particularly neurological outcomes (Salomons and Wyss, 2007). For a much more fine-grained explanation of the mechanisms by which creatine has its effects, see: Salomons and Wyss, 2007.

What is interesting about creatine supplementation as a case study for establishing in sports science is the timeline of creatine research that led to it being established as having a beneficial effect on some performance outcomes. The history of creatine research nicely highlights the importance of establishing both correlation, and the existence of a mechanism, before causal relationships can be established. This overview of the history of creatine research is paraphrased from Williams et al., 1999

(Chapter 1), which offers one of the clearest, academic, histories of creatine research available.

Creatine was first discovered in 1832 by French scientist Michel Chevreul, who extracted it from animal meats. Following this, in 1847, Justus von Leibig was able to confirm that creatine was stored in the flesh of a number of animals. Further, he found that it was stored in larger quantities in the flesh of wild animals than domesticated ones. By the 1880s, it was determined that the by-products of creatine can be found in the urine of humans, suggesting that creatine was likely also found in human muscles. Early 1900s research also suggested that, by supplementing the diet of animals with creatine, it increased the creatine content of their muscles. Chanutin et al. (1926) conducted an experiment on humans. In this experiment, humans were fed both a diet totally deficient in creatine, and later, a diet with creatine supplements. It was noted that there was a correlation between body weight of participants and a creatine rich diet. The body weight of participants increased at an unusually fast rate at the beginning of the creatine rich diet, by around 3 kg in a matter of weeks. It was also noted that creatine by-products were only found in the urine in very small amounts at the beginning of the supplementation diet. This suggested that creatine supplemented in the diet was being stored in the muscles, and that it was not being excreted much until it reached a saturation point. Chanutin was able to determine that creatine supplementation increased muscle creatine stores by between 33% and 50%. This all provides evidence that there is a mechanism by which dietary creatine is stored in the muscles. Observations of creatine amounts in the muscles and by-product excretion lead Chanutin to suggest that creatine may play a role in how protein is used by the body. This also lead Chanutin to posit that creatine may have some anabolic (muscle building) effect. Interestingly, as early as the 1960s there was also anecdotal evidence that athletes in some countries were using creatine supplements, supposing that it had beneficial performance effects.

Later, Crim et al. (1975) published a paper confirming that supplementing creatine increased creatine pools in the body by noting that creatine by-product excretion increased and decreased as dietary creatine increased and decreased. Up until this point, the majority of creatine research was mechanistic in nature. Following these discoveries, research into the effects of dietary creatine supplementation became more widespread. Research into the more fine-grained details of the mechanism by which creatine may have an anabolic effect were conducted. Further, association trials were conducted in order to try and measure the potential performance outcomes of creatine supplementation on those effects.

Williams et al. (1999) suggest that we can consider the effects of creatine on performance and muscle building to be established around the mid-1990s with the publication of the overviews: Ekblom, 1996, Newsholme and Beis, 1996, and Balsom et al., 1994. These reviews of the effects of creatine all examine both mechanistic evidence, and evidence of correlation supporting the effects of creatine's performance-enhancing effects. This means that evidence for *how* creatine may have its effects was considered alongside evidence for the net effects of creatine in the establishing of its effects.

Being faced with this case study, a hardline proponent of EBM-like methodology in EBP would likely say that only the association studies on creatine actually provided evidence of effectiveness. This would seem to ignore the fact that effectiveness was not established until both evidence of correlation and mechanism were considered together in the aforementioned reviews. However, one could counter that perhaps those overviews only considered evidence of mechanism for completeness, so an explanation for how creatine had its effects could be considered alongside the evidence from association studies that established effectiveness. This counter would be mistaken. It ignores the epistemic rationale provided for the RWT, explained in section 4.3. It was explained that, in order to determine that a correlation observed is causal in the way we think that it is, we need to know that a mechanism exists. This is because, otherwise, there are too many explanations for the observed correlation other than that the intervention is effective, to be able to rule in a causal relationship. If the EBM proponent continues to object, and insists that the association studies alone provided sufficient evidence to establish causation, then the argument of the rest of this chapter, that evidence from association studies *can* suffice to establish the existence of a mechanism comes in to play, defending the RWT.

What we can take from the history of creatine research, and the eventual establishing of its performance benefits, is that the effects of creatine supplementation required evidence of correlation and mechanism in order to be established. This is because, whilst the mechanisms by which creatine may have an anabolic effect were well known, it was not until later, when reviews considered high-quality evidence of correlation from association studies alongside it, that the effects could be considered established. This is presumably because it was not known, until these association studies were conducted, what the net effects of creatine supplementation in the body were, as mechanistic studies did not provide strong evidence for this. This case, then, nicely illustrates how the RWT applies in sports science. This, taken with the epistemic rationale for the RWT, given in section 4.2, provides strong motivation for accepting the RWT in sports science. The fact that the RWT is supported by this epistemic rationale, which does not rely on this case study, also defends it from critics who worry about case studies being used as hasty generalisations.

4.5 Key criticisms

In this section, I review five key criticisms of the RWT in general, and of the RWT as I have given it. Broadbent, Campaner, Howick, and Solomon object to the RWT in general. Gillies objects to the interpretation of the

RWT that I adopt. As well as introducing the objections, I respond to them.

4.5.1 Broadbent

Broadbent's criticism of the RWT is mostly concerned with the historical examples that inform it. Broadbent (2011), objects to the RWT as it was originally given in 2007. In 2007, it was not entirely clear if the RWT was intended to be a descriptive or normative thesis. Broadbent claims that the criticisms he levels against the RWT find problems with both interpretations. Since this, Russo and Williamson (2011) state that the thesis is intended to be both descriptive and normative. Thus, Broadbent's criticisms advance on both flanks of the RWT. Recall from earlier in this chapter, the normative element of the RWT is that it states what epistemological criteria *should* be met in order for something to be established. The descriptive element of the RWT is that it claims to describe what evidence *is* used in establishing. Broadbent claims that the RWT fails to describe practices in modern medicine. Broadbent also claims that if we take the RWT as a normative thesis, this could have had dangerous consequences in the history of medicine and, therefore, should not be accepted. I respond to Broadbent's objection to the RWT on two key grounds. Firstly, I extend Gillies' argument that the RWT *does* describe modern practice, and that the RWT is right to be informed by the historical case Broadbent takes issue with. Secondly, I argue that Broadbent's worries about the dangerous consequences of accepting the RWT are unfounded.

Broadbent's criticisms of the RWT stem from the use of the *Semmelweis* case to inform it. Gillies' (2005) examination of that case informs the description of the case that follows. In a 19th century maternity clinic in which Semmelweis worked, it was noticed that one ward had more than two and a half times the deaths of the other. It was also noticed that many of these excess deaths could be attributed to puerperal fever. Semmelweis conducted an early form of clinical trial and noted an apparent correlation between ward staff not handwashing post autopsy and rates of fever in patients examined by those non-hand-washing staff. This led him to conclude that handwashing after performing autopsies could prevent puerperal fever in patients by preventing the spread of *cadaverous particles* from cadaver to patient. This theory of disease transmission was not accepted by the medical community as a whole, however, as it was opposed by the then accepted *miasma theory.* Miasma theory, now obsolete, suggested that illness and disease were caused by noxious or bad air that was emitted from decaying organic matter. Unlike germ theory, miasma theory did not posit the existence of specific germs that caused specific diseases. Only once germ theory, and thus the mechanism of disease transfer, was established, did the wider scientific community accept Semmelweis's findings as established. Whilst it is clear that Semmelweis did fail to provide sufficient evidence to establish his causal claims in the eyes of his contemporaries, Broadbent claims that the evidence he had at the time would have been sufficient to establish those claims in the eyes of modern epidemiologists (2011, 59). This is because, given modern standards, the quality of association study conducted by Semmelweis would, Broadbent believes, have been considered sufficient to establish causation.

Broadbent argues:

- By modern standards, Semmelweis had sufficient evidence to establish a causal relationship;
- If the Semmelweis case is meant to exemplify modern standards, Semmelweis's peers should have established a causal relationship;
- Semmelweis's peers did not consider the evidence sufficient to establish a causal relationship;
- So, the Semmelweis case does not exemplify modern practice;

- Accordingly, if the RWT describes the Semmelweis case, it does not describe modern practice and, if it does describe modern practice, it does not describe the Semmelweis case (2011, 59).

As Broadbent says: ‘If the claim is merely descriptive, then we may conclude that Semmelweis's contemporaries were simply wrong to insist a mechanism be identified before they accepted a causal connection’ (2011, 59). So, Broadbent thinks that either the RWT is wrong because: either it should not be informed by Semmelweis's case because the Semmelweis case does not conform to modern standards, or, it does describe the Semmelweis's case, but as this case does not conform to modern standards, it is unsuitable as a guide for modern practice (2011, 59).

As can be seen, Broadbent's objection hinges on the idea that, if we assessed the evidence Semmelweis had, this would have been sufficient, using modern standards, to establish the causal claim. Noting this, Gillies (2019, 173-4) responds to Broadbent's objection. Although, from a modern point of view, miasma theory is incorrect, it was the dominant theory at the time. Gillies also claims that we know any dominant theory in science may be incorrect, but also that we are right to rely on the current dominant paradigm in science. Given that miasma theory was the dominant theory and that it provided evidence against the mechanism of disease transfer proposed by Semmelweis, his peers were correct not to accept his theory. There are examples to support the claim that, in contemporary medicine, we do not establish a causal claim on the basis of evidence from association studies when there is conflicting evidence of no mechanism. I argue that the practice of IARC, for instance, supports this. I give an in depth discussion of IARC practices in subsection 5.3.1, so I will be brief here. When assessing a carcinogenicity claim, IARC assess evidence of correlation and evidence of mechanism. In cases where evidence from association studies is in favour of making a carcinogenicity claim, but there also exists evidence that no mechanism exists by which an agent or exposure is a carcinogen, a carcinogenicity claim will not be established (IARC, 2019, 36). In light of this, we can see that based on modern standards, if we had the same evidence Semmelweis's peers had, we would not have established a causal claim. Just as Semmelweis's contemporaries did, we do currently use evidence of no mechanism to overturn evidence of correlation. Accordingly, the Semmelweis case does reflect modern practice. As such, by describing the Semmelweis case, the RWT does describe modern practice and Broadbent's objection to the descriptive element of the RWT does not work.

Broadbent's objection to the normative reading of the RWT is that, if taken seriously, the RWT could have negative consequences in healthcare. Broadbent argues that if we *do* require evidence of a mechanism in order to establish a causal claim, causal claims for which there is strong evidence may not be established, and this may lead to not employing useful interventions in medicine. Broadbent uses the Semmelweis case to explain his argument as follows:

- If the normative reading of the RWT is correct, we should not establish a causal claim until we have established that a mechanism exists;

- According to the normative reading of the RWT, Semmelweis had insufficient evidence to establish that a mechanism existed so, if we take this reading seriously, his peers were correct in not establishing the causal claim (2011, 59);

- If we had not developed germ theory, we would have never accepted Semmelweis's claims as we would have no evidence of underlying mechanism (2011, 59);

- If we did not develop germ theory, this would mean that we would still perform medicine dangerously, such as by performing an autopsy and then, without washing our hands, deliver a baby (2011, 59); and

- So, if the normative reading of the RWT is correct, and we did not develop germ theory, we would never have accepted Semmelweis's claims and would still be practising medicine dangerously.

He further claims that, again, this reading of the RWT goes against the modern epidemiological standard of relying on evidence from clinical studies (2011, 59). Broadbent then argues that as taking the normative reading of the RWT seriously is potentially dangerous, and it does not reflect modern practice, Russo and Williamson do not have sufficient grounds on which to argue in favour of the RWT. Broadbent thinks that we *should* want to be able to establish causal claims without knowledge of the underlying mechanism, or we will risk other similar cases in the future.

Russo and Williamson (2011, section 5), respond to this argument directly, maintaining the strong normative interpretation of the RWT. They appeal to the general fruitfulness of requiring evidence of mechanism alongside evidence of correlation in the health sciences. The high epistemic standard to which establishing ought to be upheld, on a normative reading, is important. It allows us to do things such as overturn evidence that a correlation may exist using evidence of no mechanism. Given that the spirit of EBM aims to rely on the highest standards of evidence when informing practice (Sackett et al., 1996), Russo and Williamson's defence of the normative reading of the RWT, is mostly successful here. What this means is that because modern epidemiology does hold establishing to a high standard, and does consider mechanisms in the cases like the practice of IARC, the normative reading of the RWT sets a standard that *is* met in practice.

There is one concern with Russo and Williamson's reply to Broadbent. One may still be concerned that, whilst it clears up the issues about whether the RWT conforms to modern standards, it does nothing to clear up the issue associated with practising medicine safely had we not developed germ theory. If Broadbent is right, we would still be engaging in dangerous medical practices. Although this objection is flimsy because it requires us to, questionably, consider a tricky counterfactual and imagine we had not developed germ theory, I will still respond to it. Contrary to Broadbent's objection, I maintain that even if we did not develop germ theory and an understanding of the mechanism of disease transfer, we would eventually have sufficient evidence from clinical trials to overturn miasma theory and thus accept that handwashing did something to improve safety in medical practice. I argue that this would be the case even if we did not know the details of the relevant mechanism. The first point to note here is that if handwashing does reduce disease transfer, then clinical trials would provide evidence of this. In addition to this, as evidence from association studies can establish the existence of a mechanism, (as I argue it can in section 4.6) then, if more clinical trials were conducted investigating handwashing and mortality, we would eventually have such a mountain of evidence that we would establish that there is some mechanism between handwashing and mortality. This is because the evidence from clinical studies would be sufficiently strong that we must accept that some mechanism must exist to explain our measured outcomes. We would also see that miasma theory cannot explain the observed outcomes, and so must be rejected. Accordingly, even if we did not develop germ theory, we would both reject miasma theory as providing evidence of no mechanism, and establish that some other mechanism must exist to explain the link between handwashing and mortality. So, Broadbent is wrong to suggest that if we took the RWT seriously, and if we had not developed germ theory, we would still not wash our hands. Importantly, this does not mean that Semmelweis had sufficient evidence to establish his claims. In that instance, only one trial was conducted, and the evidence produced by that one trial was insufficient to overturn the current dominant theory.

Unfortunately, I anticipate that this response will give rise to further worries. If we allow that evidence from clinical studies can, in some instances, provide sufficient evidence to overturn evidence of an absence of a mechanism, might this mean that we may end up incorrectly rejecting or accepting some causal claims? I will highlight this worry using the case of homeopathy. There are some small amounts of evidence from association studies that homeopathy is more effective than a placebo (for example: Teixeira, 2011). We also have very strong evidence that there is no mechanism by which homeopathy could be more effective than a placebo. If evidence from clinical trials *could* overturn evidence of no mechanism, is there a worry that we might incorrectly reject the strong evidence of no mechanism for homeopathy with the justification of evidence from clinical trials? I argue that this worry is unfounded. First, I must explain the evidence of no mechanism for homeopathy.

Grimes (2012) published a paper handily titled 'Proposed mechanisms for homeopathy are physically impossible'. Grimes explains that homeopaths claim that to create a homeopathic remedy we take our proposed active ingredient and create a very dilute solution of it so that the active ingredient leaves a trace of its *memory* in the solute. The more dilute the solution is, the more powerful the remedy. From chemistry, we know that one mole of a substance is the amount of a substance containing Avogadro's number of particles. This is 6.02×10^{23} particles. Homeopathic remedies are diluted such that, in 1 mole of a solution, there is less than 1 molecule of active ingredient per 6.02×10^{23} molecules of solute. As 1 mole of remedy has less than one particle of active ingredient, this means that, per mole of remedy, there is not even 1 molecule of the active ingredient present. Grimes puts this into perspective. In a common homeopathic dilution, 1 part per 10^{60} parts of water, the mass of water required for there to be one molecule of active ingredient is 15,000 times the mass of the sun, and with 28 times the radius of the sun. This means that homeopathic remedies are often simply sugar pills without an active ingredient. From our knowledge of chemistry, then, we can see that no mechanism could exist to explain how the active ingredient in homeopathy could have its effect, as the active ingredient will not be present, and there is no mechanism by which an active ingredient can leave a trace of its memory in the treatment compound. This means that, for homeopathy to work in the way that is suggested by homeopaths, some mechanism must exist to explain it that overturns the foundations of modern chemistry. Unless that mechanism does really exist, we would not want evidence from clinical trials to overturn our understanding of modern chemistry.

This worry is, however, unfounded. Just because clinical trials can provide evidence of mechanism that *could* be sufficient to overturn evidence of no mechanism, this does not mean that accepting this could lead to rejecting all we know about science on the basis of some clinical trials when we should not. Firstly, as in the case of homeopathy, the evidence of no mechanism we have is so strong that, in order to overturn it, we would need huge amounts of evidence. It is unlikely that evidence from a set of clinical trials on homeopathy would overturn all of chemistry. Further, if evidence from clinical trials did eventually provide a sufficient volume and quality of evidence to overturn a foundation of modern chemistry, it seems reasonable to suggest that there probably *is* some mechanism underlying the results of those trials that modern chemistry cannot account for. This, for instance, is what I suggest would happen in the hypothetical case where we established the effectiveness of handwashing based on evidence from clinical trials, overturning miasma theory.

As a last note, it was proposed to me that, perhaps, it may turn out that homeopathic treatments may be more effective than placebos by some other unaccounted for mechanism than the one suggested by homeopaths. However, given that homeopathic treatments are often simply sugar pills without the possibility of the presence of an active ingredient, it is unlikely that there is something special about the homeopathic process that will make this the case, except in instances where the suggestion that they are homeopathic produces a greater effect. But then, this is not some special power of the homeopathic remedy having an effect, instead it is suggestion doing the work. However, as stated above, if trials on these remedies do indicate that some unaccounted mechanism does exist, then we will eventually establish a mechanism based on that evidence if the volume and standard of evidence is high enough.

To sum up this response to Broadbent:

- As Russo and Williamson suggest, establishing should be held to a high standard;
- Holding establishing to a high epistemic standard does mirror modern practice such as that of IARC and does not set the bar too high;
- This does not mean we should worry that we would still not be washing our hands if we hadn't developed germ theory because evidence of mechanism from clinical studies would eventually override evidence of no mechanism provided by miasma theory, even if we did not have details of that mechanism; and
- Accepting that evidence from clinical studies could provide sufficient evidence to overturn evidence of no mechanism does not mean that we should worry about what else it might overturn.

My response does, of course, rely on clinical and association studies being able to establish the existence of a mechanism. I will deal with this in section 4.6.

4.5.2 Campaner

In Campaner's 2011 paper *Understanding mechanisms in the health sciences*, she opposes the RWT. In this paper, Campaner argues that we cannot think of mechanistic evidence as being separate from statistical evidence because, she claims, probabilistic knowledge is required to substantiate or construct mechanistic claims. In her own words:

> To start with, it does not seem possible to think of mechanisms separately from probabilistic evidence. The assessment that a mechanistic production system is in place cannot but start from some probabilistic relations; the hypothesis that a mechanism is in place rests, first of all, on detected correlations. (Campaner, 2011, 11)

In addition to this, she claims that:

> it is hard to see how probabilistic relations can be neglected in the construction of mechanistic knowledge. Probabilistic evidence is required to substantiate mechanistic claims (2011, 11).

Further, she claims that 'one cannot refer to mechanistic evidence as simply standing for evidence of something that is mechanistically connected' (2011, 12). This is because, on Campaner's view, mechanistic evidence is evidence that *explains*. In medicine, it is evidence with the function of explaining how to get from cause to effect. In order to fulfil this role, evidence for mechanism discovery has to, according to Campaner, come from 'probabilistic relations, previous knowledge, and interventions' (Campaner, 2011, 12). As Campaner states:

> Yet what about the assessment of newly discovered mechanistic relations? How can it be established for the first time that a mechanism is in place? Obviously, in such cases, one cannot refer to mechanistic evidence as simply standing for evidence of something that is mechanistically connected. To get from correlations to mechanistic claims one usually needs to perform some interventions (2011, 12)

and:

> it takes many years of coordinated efforts by a large number of researchers to get from the detection of a strong correlation (or, rather, a number of correlations) to mechanistic claims (2011, 12).

Now, with regard to the RWT, Campaner states that:

> I will then take issue with a particular thesis put forward by Federica Russo and Jon Williamson, viz., that two different types of evidence are at stake in the health sciences, namely, probabilistic evidence and mechanistic evidence [...] both probabilistic evidence and manipulation are essential with respect to newly discovered mechanisms. (2011, 6)

Where the RWT claims evidence must be provided for two different things – the existence of a correlation and a suitable mechanism – Campaner claims that the two types of evidence cannot be separated like this, observing those correlations is vital to understanding mechanisms. Thus, as probabilistic evidence is a key part of producing evidence that helps us to understand how mechanisms work, Campaner argues that we cannot separate the two types of evidence in the way that the RWT suggests. In sum, Campaner's objection to the RWT is built on two claims. The first is that the purpose of mechanistic evidence is to explain, not to stand for the fact that there is a mechanism. The second is that when discovering new mechanisms, the evidence that supports this type of explanatory claim is supported by evidence of probabilistic dependencies. Accordingly, or so says Campaner, we cannot distinguish mechanistic evidence and evidence of correlation in the way that the RWT suggests because, when discovering new mechanisms and using this to provide explanations, this is supported by evidence of correlation.

Proponents of the RWT would not disagree with the idea that we use mechanisms to explain; for instance, Williamson claims this (2013, section 7). However, Campaner's claim that discussing mechanisms only relates to explanations would be rejected. Further, the idea that mechanisms cannot be separated from correlations because mechanism discovery relies on detecting correlations would also be rejected. The first claim can be dealt with by bringing in the type or object distinction of mechanistic evidence (discussed in section 4.2). Campaner's main claim can be rejected for two further reasons: we do not only detect mechanisms by detecting correlations and, we are able to consider evidence of correlation and mechanism as epistemically separate precisely because that is what we do in practice.

As was explained in section 4.2, evidence that explains a mechanism is not the type of mechanistic evidence with which the RWT is concerned. The RWT is concerned with evidence that a mechanism exists. So, Campaner's argument is not an objection to claims made by the RWT. Of course, Campaner says that evidence that a mechanism exists is not mechanistic evidence. This is not a problem for the RWT. Firstly, whatever one calls 'evidence that a mechanism exists', that is what the RWT is concerned with, and this evidence may not always be evidence for how the mechanism operates. So, even if there is a disagreement in the use of terms, once the terms are defined, the disagreement between Campaner and the RWT disappears. Secondly, even if Campaner is correct that mechanistic evidence is evidence that explains, since the disambiguation of the RWT (Russo and Williamson, 2011), it has been restated as being concerned with *evidence of mechanism* instead of *mechanistic evidence.* Because of this, we do not need to worry who is using 'mechanistic evidence' correctly, because this is not what the restatement of the RWT asks for. It must be noted that, given when her paper was published, Campaner likely only had the 2007 statement of the RWT to go on, which may explain why these criticisms arise.

Campaner's main claim is still important to deal with. It seems as though, even where evidence of mechanism is not evidence that explains how a mechanism operates, evidence that a mechanism exists may still be supported by evidence of probabilistic dependencies derived from intervention trials, as Campaner suggests it must be. On this second claim, proponents of the RWT would argue that though we can derive evidence of mechanism from studies that detect associations between parts of mechanisms, this does not mean that evidence of mechanism and correlation cannot be thought of separately. We can provide evidence for details of mechanisms by conducting association studies, for instance by using association studies to provide details for the presence of some variable mediating the mechanism between A and B. As the EBM+ guidebook says: 'consider a clinical study for the claim that A is a cause of C, where C is an intermediate variable on the path from A to B—e.g., a surrogate outcome. Such a study is also a mechanistic study because it provides evidence of certain details of the mechanism from A to B' (Parkkinen et al., 2018, 14). This type of study provides evidence of an association between A and C, but this association provides evidence of mechanism between A and B by evidencing a detail of that mechanism. In this sense, then, RWT proponents accept that evidence for correlations can provide evidence of mechanism. But, what about the claim that, because of this, we cannot think of evidence of mechanism and evidence of correlation as separate? Evidence for the details of a mechanism, such as may be derived from an association study, may help us to establish that a mechanism exists, This may include evidence of a correlation between A and C, and C and B. However, this being the case does not mean that we have to understand evidence of mechanism as inseparable from evidence of correlation.

Firstly, we can also get evidence of mechanism from sources that are not only association studies and in ways where the evidence of mechanism is not supported by evidence of correlation in the same way. We may get evidence of mechanism from 'in vitro experiments, biomedical imaging, autopsy, established theory, animal experiments and simulations, for instance' (Parkkinen et al., 2018, 14). All of these methods of investigation tell us more about mechanisms than that there are correlations between parts of them. These methods of investigation tell us about the organisation of entities and activities in mechanisms more readily than detecting correlations does. So, even if Campaner was correct and the only way to talk about mechanisms is to talk about explanations, we can provide deeper explanations when we look for more than correlations. This means that, even though evidence for features of a mechanism may be supported by evidence from association studies, this is not all that supports it. This gives us reason to be able to think of evidence of mechanism and evidence of correlation separately. Accordingly, the claims that support Campaner's objection to the RWT do not hold, and the RWT still stands.

Further, in an epistemological sense, we *do* think of mechanisms and correlations separately. There is a difference between understanding what correlations a mechanism may give rise to, understanding that there is a mechanism that gives rise to those correlations, and how it does so. Because there is a difference in how we understand these things, we can separate them. As before, if understanding a mechanism hinges on understanding that there is a correlation between two things, this lacks the *ability to explain* which Campaner ascribes to mechanisms. So, if following Campaner's reasoning, detecting mechanisms means only detecting correlations, this does not allow us to explain important mechanism features such as organisation and arrangement of entities and activities.

Before moving on, it must be noted that, on a metaphysical level, Campaner may be correct. Despite how we think about mechanisms as the thing that provides explanations, perhaps mechanisms are reducible to correlations and mechanisms are really made up of correlations between parts of the mechanism. As my point is epistemic, this metaphysical point, however, can be sidestepped. As was argued above, we can and do discuss the epistemic components of causality: mechanisms and correlations, separately. As such, even if one does reduce to the other, we may still also understand them separately.

4.5.3 Howick

Contra classic-EBM, Howick does see mechanistic evidence as having an evidential role in establishing causality (Howick, 2011a). Howick considers mechanistic evidence to be useful in establishing causality when mechanistic reasoning is employed in order to provide evidence for how a cause leads to an effect. This involves describing all relevant mechanisms, and the effect of each of those mechanisms under intervention (Howick, 2011a). He claims, however, that the RWT is wrong because, historically, well controlled RCTs have been sufficient to establish causality, and mechanistic evidence, as he understands it, cannot be derived from RCTs (Howick, 2011a, 929). Howick also argues that the RWT is wrong because there are examples where mechanistic evidence alone has been sufficient to establish patient relevant outcomes.

Howick attempts to provide examples of each case: where we established causation without mechanistic evidence, and where mechanistic evidence was sufficient to establish causality. If, in instances like these, causality was established without either evidence of correlation, or evidence of mechanism, they would present counter-examples to the RWT.

As an example of a treatment that had its effectiveness established based only on mechanistic evidence, Howick proposes the use of radiotherapy for the reduction of size of large nodular goiters in order to improve breathing (Howick, 2011a, 938). Large nodular goiters obstruct airways, making sufferers' breathing difficult. Howick claims that we had knowledge of: how airway size affects one's ability to breathe, and evidence that radiotherapy can shrink these goiters. Howick claims that this knowledge provided sufficient mechanistic evidence to establish the mechanism, and the patient relevant outcome, without needing clinical studies. He does, in this instance, draw on evidence from clinical studies as supporting the claim that radiotherapy has no negative side effects on airways. He claims that further clinical studies on this intervention were unnecessary, and the mechanistic evidence alone was sufficient to establish the causal relationship.

As examples of treatments, the effectiveness of which were established without evidence of mechanism, Howick gives a long list in *The philosophy of evidence-based medicine* (2011b). This list includes the use of aspirin before the identification of relevant analgesic mechanisms, and deep brain stimulation for Parkinson's tremors, for which Howick claimed, at the time of writing, there were no known mechanisms (Howick, 2011a, 930).

I will first tackle Howick's claim that his examples show that we can establish causation without needing mechanistic evidence. Howick states that:

> there are many ... examples where treatments were widely accepted before any semblance of a mechanism was established. To name a few, Percival Pott's hypothesis that soot caused scrotum cancer (1775) was accepted years before benzpyrene was identified (1933). Edward Jenner introduced smallpox vaccines (1798) decades before anyone really understood how they worked. John Snow helped eliminate cholera with cleaner water (1849) years before the Vibrio cholerae was identified (1893), and Carlos Finlay reduced the rates of yellow fever by killing mosquitoes (1881) decades before flavivirus was identified (1927). In the last century, general anaesthesia, aspirin, and the steroids were widely used for decades before their mechanisms were understood. In this century, deep brain stimulation has been used to suppress tremors in patients with advanced Parkinson's disease, and also to cure other motor function disorders such as dystonia or Tourette's syndrome, yet researchers have not been able to

> identify its mechanism of action with any certainty. (Howick, 2011b, 131-132)

Gillies (2019, section 10.5), attempts to defend the RWT from this claim of Howick's. Gillies presents a treatment of Howick's list of causal relationships established without establishing a mechanism. He splits these examples into two camps. those that came before, and those that came after, the modern conception of medicine. In this treatment, Gillies accuses Howick's pre-modern-medicine examples of being anachronistic. This is because many of them emerged in the 17th century, a pre-statistical era. The medical community in this time did not, Gillies states, 'have the concepts of statistical evidence and evidence of mechanism which we use today, and which are involved in the RWT' (2019, 179). Accordingly, the standards for establishing were different and do not bear on modern standards. Gillies, using a bevy of historical examples, suggests that serious 'scientific medicine' was not practised until the middle of the 19th century. He exemplifies this with the fact that the Royal College of Physicians openly objected to the use of hospital statistics being used into the 1820s. According to Gillies, as the RWT is framed in terms of medical statistics, counter-examples to it cannot be taken from before medical statistics were dominantly accepted. A worry with Gillies' defence here is that the RWT is informed by historical examples, such as Semmelweis's case. However, in light of the fact that Semmelweis used exactly the type of statistics modern medicine is concerned with, employing it as an informative case cannot be considered to be anachronistic. If Gillies' argument holds, Howick's points do not stand.

One may readily accept Gillies' response to Howick's objection if they are not concerned with a potential entailment of the response I will raise in the next paragraph. The examples Howick uses are anachronistic, although they may have met the evidential requirements to be considered to be established at the time, the RWT is concerned with modern medicine and the requirements it places on establishing causal claims *now.* So, one cannot use anachronistic standards of establishing as a yardstick against which to compare the RWT.

However, one may be concerned that an entailment of Gillies' response seems to be that, if we only want to take seriously cases where establishing relied on modern statistics, we would have to consider that nothing was established by modern standards before statistics were widely and commonly used. On Gillies' view, it looks like anything that was considered to be established before the middle of the 19th century was not *really* established because modern statistics were not used. The worry is that although the standards required to establish claims have changed, one may be able to argue that, in principle, some anachronistic example did meet the modern standards required to establish a causal claim and, so, would rightly have been considered to be established before the use of modern statistics. If Gillies' response is the only defence of the RWT against Howick's points, it would be troubling. This is because we do not want to commit the RWT, by defending it in this way, to the idea that claims made pre-modern-medicine were never established if we do not have to, particularly because any counter-example where we would have to consider something established that did not rely on modern statistics could be used as a counter-example to a key defence of the RWT. Even if this type of counter-example does not exist, a better response to Howick is one that does not risk them.

A stronger defence of the RWT than Gillies' is one that does not have this entailment because it will not be susceptible to these kinds of counter-examples. In order to defend from Howick's counter-examples, without having the entailment I argue that Gillies' defence has, we need to avoid the claim that pre-statistical medicine could not establish causal claims, *and* argue that causal claims that do not rely on evidence from mechanistic studies can establish mechanisms. If we show that any genuine counter-examples Howick gives, where a causal claim was justifiably considered to be established also succeeded in establishing both correlation and mechanism, then they are not counter-examples to the RWT. This, in turn, becomes possible if we accept that we may obtain evidence of mechanism without conducting mechanistic studies. This relates back to the disambiguation work Illari did which was explained in section 4.2.

Whilst the RWT is concerned with evidence for the existence of a mechanism, Howick is concerned with evidence that can be used for mechanistic reasoning, or, being able to reason from evidence to how a mechanism (or system of mechanisms) operates. Given that Howick is concerned with mechanistic reasoning, it is not surprising that he does not see that we could not get this type of evidence without mechanistic studies, and that, as he sees it, causation has been established without it. Recall that the disambiguated RWT claims that what is important to establishing is the existence of a mechanism, not how it operates. So, when Howick discusses mechanisms in his mechanistic reasoning sense as a refutation of the RWT, he ends up talking past the RWT. Gillies (2019, section 10.5), uses one of Howick's examples to explain this. Howick claims that it was established that mosquitoes could transmit yellow fever in 1881, before the virus responsible for yellow fever was identified in 1927 (2011b, 132). Gillies argues that this case is not a counter-example to the RWT as there was support for the disease mechanism, even if the exact mechanism details were not fully understood. This is perfectly allowable within the constraints placed by the RWT. Howick, by providing examples where causation was established without mechanistic reasoning, is certainly refuting *a* claim, but not one made by the RWT. Further, if Howick were to accept the way in which Russo and Williamson use evidence of mechanism, when the disambiguated RWT is taken along with the claim about RCTs establishing the existence of mechanisms given in Williamson, 2019, this appears to clear up the problem. What this shows in defence of the RWT is that where Howick says that his examples did not rely on evidence that could explain mechanisms, and this is a counter-example to the RWT, he is wrong. He is wrong because this is not what the RWT is asking for. In order for this defence of the RWT to stand, I will need to make an argument to show how we can establish the existence of a mechanism without conducting mechanistic studies. This is the work of section 4.6.

One divergence between Gillies' statement of the RWT and the 2018 statement of the RWT given earlier in this chapter must be noted. Gillies suggests that a mechanism need only be *plausible* for causality to be established. This opposes the statement given earlier that a mechanism needs to be established in order for causality to be established. This divergence does not, however, take from this defence of the RWT. Whether the existence of a mechanism needs to be established, or only a plausible mechanism need be posited, neither require the full, very fine-grained, understanding of that mechanism with which Howick is concerned. I will give Gillies' reformulation of the RWT in subsection 4.5.5 in which I argue against it.

Howick's other claim, that effectiveness can be established based on mechanistic reasoning alone, does not unsettle the ground for the RWT either. My defence of the RWT against this claim again relies on Illari's disambiguation discussed in section 4.2. This is because, where Howick claims that causality was established using mechanistic reasoning, the evidence base on which Howick's counter-examples rest is sufficient to establish both a mechanism and a correlation.

How may we establish a correlation claim without conducting clinical trials? We may, for instance, establish a causal claim where we notice a large effect size, particularly if this evidence is taken in conjunction with evidence from observations of A giving rise to B. Both of these types of evidence can be provided by mechanistic studies and will be sufficient to establish both a mechanism and a correlation. Williamson (2019, section 2) raises this point. He introduces the argument made by Smith and Pell (2003), that we have sufficient evidence of the effectiveness of preventing death from high falls by utilising parachutes to establish their effectiveness without the need for conducting association or clinical trials. In fact, in this case, it would probably also be unethical to conduct placebo controlled trials on parachute use. We have sufficient knowledge of how parachutes work, and have seen sufficiently many people die from high falls without parachutes, and survive from high falls with them, to establish a correlation, a mechanism, and thus, that they are effective.

This can be compared to the example Howick gives of the use of radiotherapy to improve breathing. Howick claims this is supported by mechanistic reasoning alone. However, we will see that even if all the studies that support this claim are mechanistic studies, they clearly give rise to evidence of correlation. If, when we apply radiotherapy to large nodular goiters, we notice that their size is reduced, and breathing becomes easier, it is clear that there is a correlation between the radiotherapy and improved breathing. This, then, obviously produces some kind of evidence of correlation. As such, one *can* reason to the claim that using radiotherapy to shrink these goiters could improve breathing using evidence from mechanistic studies, but one is still supported by evidence of correlation, even if it is tacit. So, if Howick presented us with an intervention that had its effectiveness established based purely on mechanistic reasoning, supported by evidence from mechanistic studies alone, this would still not be a problem for the RWT. This is because the RWT allows that mechanistic studies can provide both evidence of mechanism and evidence of correlation. So, one supporting the RWT could claim that whatever mechanistic studies supported the mechanistic reasoning used must have been sufficient to establish both the existence of a correlation, and of a mechanism.

4.5.4 Solomon

Solomon (2015), is a key opponent of utilising evidence of mechanism in an evidential role in *determining* or *justifying* causality, though, she does give it a preliminary role in the inventing, or *discovering*, of proposed treatments. The distinction here being that, according to Solomon, evidence of mechanism is useful in proposing treatments, but that it is not used for the evaluation of the effectiveness of treatments. She frames how she sees the issue nicely, as it relates to the medical sciences:

> A general problem with mechanistic accounts is that they are typically incomplete, although they often give an illusion of a complete, often linear, narrative. Incompleteness is the consequence of there being mechanisms underlying mechanisms, mechanisms inserted into mechanisms, background mechanisms that can fill out the mechanistic story, and mechanisms that can hijack regular mechanisms. That is, there is a complex interaction of multiple mechanisms in a chaotic and multidimensional system. There are possible hidden mechanisms everywhere in mechanistic stories, despite an easy impression of narrative or causal completeness. Since we do not have a theory of everything, it is not possible to know in advance whether a particular mechanistic intervention will have the intended result. (Solomon, 2015, 131-132)

Solomon makes two further points: 'we can have evidence for mechanisms, but that is evidence that the mechanisms operate, not evidence that a particular proposed intervention (which depends on more than the hypothesized mechanisms, even if those mechanisms exist) will work' (Solomon, 2015, 123), and '[m]echanistic reasoning (or "mechanistic evidence") does not play a role in the process of evaluating the effectiveness of new interventions' (Solomon, 2015, 132).

To condense this point, as it relates to the issue at hand, the objection can be stated as follows: evidence of mechanism alone tells us that a mechanism operates, but not that an intervention will have its proposed effect; and we will never know if our understanding of a mechanism is complete given the potential for complexity and masking, and thus, it should not have an evidential role in the assessment of intervention efficacy.

In a recent paper, Auker-Howlett and Wilde (2019, 460) effectively argue that Solomon's points 'at best' argue that mechanistic evidence is insufficient to provide evidence of effectiveness when taken alone. What Solomon succeeds in arguing, they show, is that one cannot establish causation based only on evidence of mechanism. She argues that evidence of mechanism is insufficient to establish the efficacy of a treatment because it does not provide any evidence for the extent to which an intervention has an effect, or evidence that there is a net putative effect. This is, in fact, in line with the RWT. The reason that the RWT requires the establishing of both a suitable correlation and a mechanism is so that the overall outcome of the intervention being tested can be seen. This is because knowing the existence of a mechanism alone does not mean that one knows what, once masking mechanisms are involved, the effect size of that mechanism will be. Auker-Howlett and Wilde refer to this type of reasoning, whereby evidence of correlation and evidence of mechanism are each used to provide evidence that hides the flaws in the other type of evidence as *reinforced reasoning*. This was discussed in section 4.2. Solomon may respond that this still appears to be consistent with her thesis that it is evidence of correlation that is used to determine effectiveness. She would be wrong to say this, however. As has been previously explained, evidence of correlation alone is insufficient to establish effectiveness as, without evidence of a mechanism, we cannot know if the correlation observed is causal, or has some other explanation.

4.5.5 Gillies

Donald Gillies (2019) argues that in order to establish causality, the only evidence of mechanism we need to have is evidence that a plausible mechanism exists. This goes against the disambiguated statement of the

RWT which claims that in order to establish causality we need to *establish* the existence of a mechanism underlying a causal relationship, not merely posit a plausible one. Gillies takes the main claim given by the RWT, but holds off on the need to *establish* a mechanism. To Gillies, a plausible mechanism is one 'confirmed by background knowledge but not necessarily by particular investigations and experiments designed to test it out' (Gillies, 2019, 140). He informs his argument with the case of establishing the carcinogenicity of smoking. Here, I reject his reformulated statement of the RWT on the grounds that we can consider the case study he uses to have established the existence of a mechanism.

In 1976, Doll and Peto published results from a trial conducted on doctors in the United Kingdom which followed them for 40 years, taking note of smoking habits and cause of death. This trial measured the number of deaths in given cohorts due to lung cancer. The cohorts were arranged as smoker and non-smoker, and also by quantity smoked. The results of the trial indicated that smokers were more likely to die of lung cancer than non-smokers. Further, the results of the trial indicated that there is a dose response relationship between tobacco consumption in trial cohorts, and lung cancer death in trial cohorts. Cohorts where more cigarettes were smoked per day had more deaths from lung cancer than cohorts where fewer cigarettes were smoked per day. Doll and Peto concluded that this was sufficient evidence to claim that smoking does *cause* lung cancer. Gillies grants them that they demonstrated a 'striking' correlation, but also points out that by the standards of the RWT and his own, finding a strong correlation isn't grounds for establishing causation (2019, 137). After all, notes Gillies, alcohol consumption is correlated with lung cancer deaths (although not to the same extent as smoking), but this is not sufficient grounds alone to establish the relationship causal. In fact, lots of heavy drinkers are also heavy smokers.

Gillies suggests that in 1976 there was a plausible account of the mechanism by which smoking causes lung cancer, but no established mechanism. On Gillies' account, because of this background knowledge, Doll and Peto *did* in fact establish the causal link between smoking and lung cancer, but Gillies claims that this was done *without* establishing a mechanism, only having a plausible account for some details of a mechanism. Gillies claims that it was possible to establish the causal relationship because the evidence of correlation, along with a posited plausible mechanism, were sufficient to rule out anything other than smoking as being able to explain the correlation observed. Ruling out any other explanation, then, was key to establishing the causal claim. It is not entirely clear from Gillies writing if he thinks that having a plausible mechanism means that the existence of a mechanism is established but not its details, or if it means that we do not know if a mechanism exists, but one might, and it may look like the plausible mechanism. This means that it is not clear if Gillies thinks that Doll and Peto had established the existence of a mechanism but had not filled in the details of that mechanism when establishing the causal link between smoking and lung cancer. If Gillies thinks that they had not established the existence of a mechanism, then this view of the RWT appears to be at odds with what I argue in this chapter. However, the key argument of this chapter is that we *can* use evidence from association studies to establish the existence of a mechanism. So, if the argument from the next section of this chapter holds, and evidence from association studies and background evidence of mechanism *is* sufficient to establish the existence of a mechanism, then Doll and Peto *were* right to claim they had established a causal link between smoking and lung cancer. This is because they had established a correlation, and because the strength of evidence from the association study was also sufficient to establish the existence of the mechanism, particularly considering the strength of background knowledge and dose response relationship shown. This means that whether Gillies thinks that having a plausible account of a mechanism means that the existence of a mechanism is established but not its details, or, he thinks that having a plausible account of a mechanism means that we do not even need to establish the existence of a mechanism does not matter. This is because

Doll and Peto did have sufficient evidence to establish the existence of a mechanism and a correlation. This, then, leaves us in just the place we were after discussing the Howick objections. The RWT is safe from the objections posed against it, so long as an argument that evidence from association studies can establish the existence of a mechanism can be made.

A small tangential note can be made on Gillies' objection. There is a serious worry that allowing that a causal claim can be established using only evidence of a plausible mechanism can be dangerous. This is precisely the type of thing that the EBM+, which has its roots in the RWT, wants to avoid. If the standards for establishing causation are lowered to allow merely the plausibility of a mechanism to be required, there is a worry that people could *wheel out* plausible sounding mechanisms in order to make claims about observed correlations. Requiring that evidence be assessed in order to ensure that evidence of mechanism is established has a key role in avoiding this.

4.6 The crux: can evidence from association studies establish the existence of a mechanism?

Thus far, I have introduced the RWT, and some key criticisms of it. I have also provided defences of the RWT from these criticisms. This is important work as it represents an overview of the criticisms and defences of the RWT in one place. What I have done shows that the RWT stands on solid ground. However, what can also be seen in these defences of the RWT is that, in order for the RWT to stand on this solid ground, there is one key claim that needs to hold that I have not argued for yet. This is the claim that evidence from association studies can be sufficient to establish the existence of mechanisms. Arguing in favour of this claim is vital, as the next chapter in this thesis, which discusses the applicability of an EBM+ like framework of evidence assessment for sports science, relies on it. As such, in this section, I provide a defence of the claim.

As has been seen, key opposition positions to the RWT, which would overturn the RWT, are those which claim that effectiveness can be established in medicine without establishing a mechanism, by relying on evidence from association studies only. The rationale behind this is the belief that association studies can establish correlation and causation, but do not suffice to establish the existence of a mechanism. If interventions can have their effectiveness established without establishing the existence of a mechanism, then the RWT fails to be descriptive, and, as Broadbent suggests, arguments in favour of its normativity may be lacking. In order to defend the RWT, then, an argument must be put forward for a way in which the existence of a mechanism may be established based on evidence from association studies alone. Showing that it is possible to do so in at least one way will be sufficient to defend the RWT, without preventing alternative arguments to the same effect. For instance, I would be unwilling to commit myself to the claim that there is only one method by which evidence from association studies can be used to establish the existence of a mechanism.

It may look to some that, by granting the claim that association studies can establish the existence of a mechanism, the RWT always has a get out clause for potential overturning examples. This is, of course, contentious. However, the claim that association studies can be sufficient to establish the existence of mechanisms is not presented without argument. Williamson (2019) provides a set of criteria that, when taken together, provide an inference intended to show how we can establish the existence of a mechanism using only evidence from association studies. I quote Williamson (2019, 44), who gives them as follows:

- 'There are sufficiently many independent clinical studies

- They are of sufficient quality
- Sufficiently many studies point in the same direction
- They observe a large enough correlation
- Fishing, temporal trends and non-causal relationships are ruled out
- No other evidence suggests a lack of suitable mechanism'

According to Williamson, when these criteria are met, there is sufficient evidence that a mechanism exists to consider its existence to be established. Importantly, what it is to meet these criteria is not given by Williamson. The onus of deciding if criteria are sufficiently met is on those conducting research. Recall that according to Williamson something is established 'just when standards are met for treating the claim itself as evidence, to be used to help evaluate further claims' (Williamson, 2019, 35). Importantly, these criteria do not mean that the existence of a mechanism needs to be given explicitly. By establishing the causal claim, these criteria allow that the existence of a mechanism is established implicitly.

How does this inference show that association studies can provide sufficient evidence to establish the existence of a mechanism? I will now provide my own argument to this effect.

After completing a clinical trial, or set of association studies, where a correlation has been observed between the presence of intervention or exposure A and putative outcome B, there are a number of factors that can explain why that correlation is observable. This was a key premise in the **Excluded Explanations Argument**. These have been given in detail in chapter 2, but include: chance, bias, confounding, and that there is a mechanism by which A gives rise to B. Williamson gives us a set of criteria which, if met, are intended to show how association studies can provide evidence for the existence of a mechanism, and can even be sufficient to *establish* the existence of that mechanism. I argue that these criteria show that association studies can establish the existence of a mechanism when met because they provide criteria illustrating what must be done to provide sufficient evidence to eliminate other explanations for observed correlations, other than that a mechanism exists. This can be taken as meaning that where the criteria are met, we are able to rule in the existence of a mechanism, thus overcoming the **Excluded Explanations Argument**.

It may look like this inference is inconsistent with a claim I made in subsection 4.5.1. I claimed that evidence from association studies could, in some very rare instances, overturn evidence of no mechanism by providing sufficient evidence against that mechanism. Williamson's inference states that, in order to establish a mechanism, no other evidence can suggest a lack of a suitable mechanism. I do not think that there is an inconsistency. Firstly, it is important to note that the conditions Williamson gives are sufficient, but not necessary, for establishing a mechanism. As I have said, the inference does not commit itself to the claim that it sets out the only way by which we may establish a mechanism on the basis of evidence from association studies. Further, it is *establishing* the existence of a mechanism that the inference claims can be done where there is no other evidence of no mechanism. It is not falsifying an incorrect mechanism that needs this. Secondly, clinical studies do provide some evidence of mechanism, even where it is insufficient to establish that mechanism. Where studies provide strong evidence of a correlation, they may also raise our rational degree of belief that a mechanism may underlie that correlation, even if only a little. So, in some rare instances, evidence from association studies could combine to provide sufficient evidence of mechanism to falsify current evidence of no mechanism. This would be the case, for instance, in the hypothetical case I give in subsection 4.5.1 where I argue that we would have, eventually, overturned miasma theory based on evidence from association studies, even if we did not develop germ theory.

It is important to make some notes on the structure of the argument. As was explained in section 4.3, the view of establishing that Williamson and I adopt is allowable under both factive and non-factive establishing frameworks. If one thinks establishing is factive, we can explain the framework by which we can establish the existence of a mechanism as follows:

> When one conducts a trial, there is a set of possible explananda for observed correlations between A and B. If we rule out some possible explananda, then we must infer that the true explanations must be within those that we have not yet ruled out. If we are able to rule out every possible explanation but one, we may infer that the last explanation is the true explanation. So, by giving a set of criteria that allow us to rule out every possible explanation for an observed correlation between A and B, except that a mechanism exists, the inference explains how we can show that a mechanism must exist.

If one thinks that establishing is not factive, the inference can be seen as an inductive inference:

> When we observe a correlation, a set of explananda will exist for that correlation. When the criteria laid out by Williamson are met, there is a high probability that all explanations for the correlation are ruled out, except that a mechanism exists that gives rise to the correlation. So, where the criteria are met, we may have a sufficiently high confidence that a mechanism exists to explain the correlation that we may say that we have established the existence of that mechanism.

One way that we can explain how the criteria does this is by the following reasoning. Say that after a set of association studies we have a number of things that could explain how the observed correlation came about. Given what we know about association studies, including how different factors can raise and lower the likelihood that explanations for observed correlations are genuine, we can use that knowledge to review the trials and ascertain how likely it is that any of those explanations *are* genuine in this case. Williamson gives us a set of criteria that, using what we know about association studies, can help us do this. If met, they lower the likelihood that explanations other than that a mechanism exists are the genuine explanations for observed correlations. For instance, criteria requiring that there be a large number of independent studies, and that the observed correlation is big enough, helps to eliminate chance as an explanation for the observed correlation. If, following these criteria, we are unable to show that a mechanism does not exist, and are also able to show that other explanations are highly unlikely, we must in turn raise our confidence that a mechanism must exist that gives rise to the observed correlation between A and B. If our confidence that a mechanism exists is high enough, given how we review the trials, this can be sufficient to establish the existence of that mechanism.

Whilst it also does not rely on this view, my argument can also be explained as a kind of Popperian hypothesis testing. We take possible explanations for the observed correlation and try to disprove them using what we know about association studies. Williamson's criteria use what we know about association studies and, when met, serve to show how we are able to rule out the application of the **Excluded Explanations Argument**. For instance, chance is ruled out because it verges on impossibility that chance would, in so many trials, give rise to such a large correlation. In addition to this, confounding and insufficient blinding are ruled out because the trials are all found to be of sufficient quality to account for these adequately. Further, we also try to disprove that a mechanism exists. This involves determining that there is no evidence a suitable mechanism does not exist. Thus, the criteria give us the ability to explain how we can, using evidence from association studies and background knowledge if necessary, go about hypothesis testing all possible explanations for trial outcome, and what needs to be met in order for us to use that as evidence to establish the existence of a mechanism. Admittedly, this bears similarity to the reasoning used to support the claim that ideal RCTs can provide good evidence, but I never disputed that. I disputed the claim that, in general, RCTs in sports science were close enough to ideal that we could assume that they provide good evidence.

An important aspect of the RWT is that it is intended to be descriptive. As Williamson's inference is used to support the RWT, it must also be descriptively adequate to explain how evidence from association studies can be used to establish the existence of a mechanism. Without going through all past cases where causation was established based on evidence from association studies alone, to check to see if the existence of a mechanism was established, it may be difficult to see if this tracks real life practice. As argued before, sports science is moving into the realm of EBP, which draws from medicine. Given, as it has been seen, it is possible to establish the existence of a mechanism in medicine on the basis of evidence from association studies, it should be possible to hold those instances up as examples for practice in sports science, too. However, it is also important to give a sports science example.

Ergogenic aids are substances, equipment, or techniques that can improve performance, or the ability to do work (Holowchak, 2002). Despite some aesthetic and ethical questions regarding the use of ergogenic aids in sport, they are commonly used in many sports training and competition environments. In a philosophical overview of ergogenic aids, Holowchak suggests, for instance, that some ergogenic aids such as golf clubs that are forgiving of poor technique, or anabolic steroids that improve recovery and improve performance raise questions about their use in sport. Caffeine is one such ergogenic aid that, despite some calls for greater regulation of its use in competitive sports (Sinclair and Geiger, 2000), is widely used to improve performance. These types of questions, however, can be sidestepped where we are concerned with the quality of research that supports establishing causal relationships, not whether they can ethically be employed. Caffeine is a well-studied compound and is used widely as a sports supplement. This means that caffeine may allow us to, among other things: run faster (Glaister et al., 2008), lift heavier (Giraldez-Costas et al., 2020), and cycle for longer (Pasman et al., 1995). We can consider the ergogenic effects of caffeine to be established given that they are taught as established as part of sports science and sports nutrition curricula, and because they are listed in textbooks as established (see for example: Woodruff, 2016, 147). Further, what we know about caffeine is used to support claims and investigations into other phenomena, such as claims about the ideal timing and dosing of caffeine to maximise ergogenic effects (Woodruff, 2016, 147). Recall Williamson's claim that a key component required for something to be considered established is that it is used as evidence to support further claims. This highlights that we can say that we have established the effectiveness of caffeine. The ergogenic effects of caffeine were established before the details of the mechanism underlying those effects were properly understood (Woodruff, 2016, 147). In fact, many of the *proposed* (but not established) mechanisms to explain the correlation were either wrong or flawed (Hodgson et al., 2013, Introduction). Even now, whilst some mechanisms by which caffeine has its effects on performance outcomes have been investigated and are considered established, the full details of the mechanisms by which caffeine has its effects are not known. If we look at overviews of evidence used to support the ergogenic effects of caffeine, such as that given in the comprehensive overview in Hodgson et al., 2013, and the meta analysis Doherty and Smith, 2004, we can see that the evidence supporting these effects comes from association studies, and that it is acknowledged that mechanistic studies have not managed to explain fully the specific details of the mechanism by which caffeine has its effect. What we can draw from this is that, in the case of caffeine, we have evidence for some mechanisms by which it works, but not evidence that explains the full workings of all relevant mechanisms. What we can take from this is that, whilst mechanistic studies have been conducted, they have not been sufficient to find the details of all relevant mechanisms, or to establish the details of posited mechanisms. However, causal relationships have been established for a number of outcomes for caffeine use. Further, attempts to investigate what the mechanisms supporting this causal relationship are point to the fact that researchers clearly believe that mechanisms must exist to explain the effects. Thus, evidence from association studies must have been used to establish the existence of these mechanisms. So, to explain the ergogenic effects of caffeine, the strength of studies, effect size observed in these studies, and number of studies indicating a causal relationship, must be sufficient to allow that the existence of a mechanism to explain the observed correlation is implicitly established. An important note here, given the context of this chapter, is that there has at no point been evidence that there is no mechanism by which caffeine may have its ergogenic effects. Also, some mechanisms by which caffeine has its effects are now considered to be established, such as its ability to lower perceived exertion, improve motor recruitment, and its ability to support muscle contraction, but the full details of all mechanisms are not, and these were not established when the ergogenic effects of caffeine were originally established (Woodruff, 2016, Chapter 7).

Of course, when medical researchers talk about mechanisms, they are generally concerned with how a mechanism operates, and explanations, as Howick and Campaner describe. This means that it may be possible for medical researchers to establish the existence of a mechanism, without referring to doing so as 'mechanistic research' in discussion portions of research. They likely do follow the guidelines given by the RWT without really knowing it: they would not consider a causal relationship established if they did not meet the required criteria to establish the existence of a mechanism to account for it, but they do not think of doing that as mechanistic research. After all, causation is unlikely to be established if there was not a high standard of evidence that the intervention caused the putative effect. Implicit in this is establishing that there is a mechanism that allows A to cause B. Researchers *know* that randomization, placebo control, and confounder identification are intended to help rule out explanations other than that there is a mechanism, even if they do not acknowledge it. This does not make the RWT wrong, it simply marks a disconnect between the language of theory and of practice. From this we can see that, even without directly discussing evidence of mechanism, it is possible for researchers to establish the existence of a mechanism.

One may try to counter the claim that evidence from association studies can be sufficient to establish the existence of a mechanism with the claim that evidence from mechanistic research is necessary to set up those trials. The claim would be something along the lines of: we do not establish the existence of a mechanism without doing mechanistic research because mechanistic research plays a key role in early development of interventions and trial design. It is obvious that this *can* be the case: evidence from mechanistic trials can be used to set up and evaluate clinical trials. However, there is nothing in the inference that says that background knowledge cannot play a role in establishing based on association studies. So, this objection does not matter for the RWT. Conversely, there is nothing saying that it also must play a role. So, trials that are set up without background knowledge of mechanisms do not violate the inference. Further, this objection actually gives credence to the RWT by suggesting that evidence of mechanism plays an essential role in establishing causation, given how intertwined it is with assessing the results of association studies.

4.7 Anticipated criticisms

An opponent of the RWT, such as one with views similar to Broadbent, may raise the Semmelweis case against the idea that association studies can provide evidence sufficient to establish the existence of a mechanism. A critic may raise it to suggest that, in the Semmelweis case, a correlation was established, and that, as a result, causation *should* also have been established, based on modern standards. Thus, if association studies could establish the existence of mechanisms, a critic may claim that Semmelweis's case should, along Williamson's view, also have established the existence of a mechanism. As it did not, they may claim that the view does not hold.

For now, generously assume that the evidence collected by Semmelweis in his investigations fulfilled the criteria given by Williamson's inference. This would include the assumption that the data collected was sufficient to count as many studies, and that it ruled out all other explanations for the observed correlation other than that a mechanism existed. Would, according to Williamson's view, this not have been sufficient to establish the existence of a mechanism? It is starting to look like all the criteria Williamson lays out for establishing mechanisms were met, without establishing the existence of a mechanism. This, however, is wrong. The criteria in the inference were not all met. There was, at the time, evidence to suggest the lack of a suitable mechanism: the miasma theory. Whilst it must be noted that the evidence supporting miasma theory would not be considered to be high quality now, the evidence supporting it was, at the time, treated seriously. The evidence was that the existence of the miasma was a consensus opinion, and that historical texts such as the Hippocratic texts, which advocated the miasma theory, were considered to be good evidence (Karamanou et al., 2012, 58-59). Taking historical context seriously, the 19th century evidence of mechanism would, here, overrule the correlational evidence. This is in line with EBM+ methodology, derived from the RWT. Some may find this troubling. However, EBM, and EBM+ both aim to hold establishing to a high epistemic standard, which Semmelweis did not meet at the time, however troubling that being the case is. Therefore, despite the fact that, given current evidence for disease transmission, we can consider the causal relationship to be established, this does not mean that Semmelweis's contemporaries would have been right to do so, in light of the evidence they had at the time.

It could also be raised that the inference being discussed allows for a black-box view in disguise. Black-box views for establishing causation state that it does not matter, or that it is even preferable, if one does not know the mechanism responsible for a putative effect, when establishing causation. This is for a variety of reasons, including that basing causal inferences on mechanistic evidence can be misleading. This is a view that EBM+ is trying to move away from. However, if one establishes causation based on evidence from association studies, not utilising any evidence for how the relevant mechanism operates, this could be construed as akin to the black box view. The inference simply outlines criteria by which the existence of mechanisms can be established based on association studies alone. However, it is still explicitly concerned with the existence of mechanisms; it still leads to the RWT, and to EBM+ methodology, which call for the explicit evaluation of evidence of mechanism, as will be explained in the next chapter. It must be admitted, though, that, in some instances, by following Williamson's inference, we may not know the details of a mechanism. Whilst this is allowable under the RWT, and therefore sufficient to establish causality, it does represent an impoverished understanding of the causal relationship (to echo Gillies) when compared to an understanding that includes mechanism details. In part III of this thesis, I discuss how having evidence for the details of mechanisms should be an important concern to EBP because of the utility it affords research and practice prescription.

4.8 Conclusion

In this chapter, I began by introducing the RWT. This is the thesis that, in order to establish a causal relationship in medicine, we need to establish the existence of both a suitable correlation, and a mechanism. I also defended the RWT from five prominent criticisms. An outcome of these defences was that, for the defences to hold, it must be possible that evidence from association studies could establish the existence of a mechanism. This is essential because there are historical instances where causal relationships were established on the basis of evidence from association studies. I, then, used Williamson's inference to argue that evidence from association studies can be sufficient to establish the existence of a mechanism, thus allowing that it can establish causal relationships. I did this by arguing that the inference laid out by Williamson gives us sufficient criteria to infer that a mechanism must exist to explain observed correlations.

I also used the case of creatine supplementation to improve sports performance as a historical case to illustrate that the RWT applies in sport. Further, I also used the case of caffeine to show that evidence from association studies can establish mechanisms, and therefore causation, in sports science, as well as in medicine.

This chapter sets up the next chapter. In chapter 5, I argue that, in light of the RWT, and in order that we may justifiably establish causal relationships given the low quality of evidence from many RCTs in sports science, we may take note of the practices of IARC, and those recommended by EBM+ in sports science. This means that we should assess both evidence from mechanistic studies, as well as association studies. This, I argue, allows us to better fulfil the EBP goal of abiding by the best possible evidence by giving better grounds on which to establish causal relationships. This supports the **Better Evidence Thesis**.

Chapter 5

The importance of assessing mechanistic studies

5.1 Introduction

In the introduction to this thesis, I explained that much of sports science sees itself as moving towards an evidence-based framework, borrowing heavily from EBM. In part, this manifests itself in the privileging of evidence from RCTs. Also recall that the evidence-based framework advocated by many in sports science either diminishes the importance of evidence from mechanistic studies, or discounts it entirely. This is, in part, due to the perceived inadequacy of mechanistic studies for establishing efficacy. This is supposed to render it less useful for establishing causal claims, and therefore establishing practice. To sum up the conclusions of this thesis so far: RCTs in sports science often produce low-quality evidence as they are often susceptible to the **Excluded Explanations Argument**. Following this, informed by historical cases and the RWT, there is motivation for the claim that in sports science, evidence of correlation and of mechanism is normally necessary to establish effects. So, to establish causal claims we need evidence of correlation and of the existence of a mechanism. Further, group and N of 1 RCTs alone are probably not sufficient to establish those claims in many cases.

This is obviously a problem for sports science. If what is supposedly the gold standard method of evidence gathering cannot produce evidence of a sufficient quality to establish causation, is there anything we can do to provide a stronger evidence base for causal claims? Further, given that many sports scientists and practitioners want to employ an evidence-based framework, justifying practice on the basis of strong evidence, how can practice be justified? In this chapter, I propose a solution to the problem. Sports science could better fulfil the goal of relying on the best possible evidence if, in addition to assessing evidence from RCTs, evidence from mechanistic studies was considered when assessing causal claims. This is because, by assessing evidence from mechanistic studies as well as association studies, we can increase our rational confidence that we have excluded alternative explanations for observed outcomes, ruling in a proposed relationship as causal, thus avoiding the **Excluded Explanations Argument**. This claim is the **Better Evidence Thesis**. This solution is informed by Evidence-Based Medicine Plus (EBM+), and the International Agency for Research on Cancer (IARC). It also furthers the motivation of the claim that establishing causation in sports science generally requires evidence of mechanism and correlation.

5.2 Clarifications

In this thesis, and this chapter in particular, I use the terms and phrases: *the status of a claim*, *establishing claims*, and *general mechanistic claim* as they are used by Parkkinen et al. (2018). I do this because I am defending the view of Parkkinen et al. in their 2018 publication that espouses the EBM+ framework, and because I am arguing for the importance of that view also being adopted in sports science. Thus, it is best to use the terms and definitions as they are employed by Parkkinen et al.

The ‘status of a claim’ can refer to a causal claim, mechanistic claim, or correlation claim. As the term is used by Parkkinen et al. (2018, 27), it is ‘the status that the evidence confers on the claim under consideration’. The lowest statuses of causal claim Parkkinen et al. give are: ‘ruled out’, and ‘provisionally ruled out’. These are where high-quality evidence warrants a ‘high level of confidence in the negation of the claim’, and where ‘moderate quality evidence warrants a high level of confidence in the negation of the claim’, respectively (Parkkinen et al., 2018, 27). The scale then runs through: arguably false, speculative, arguably true, provisionally established, and established. For a claim to be established, high-quality evidence must warrant a high level of confidence in a causal claim (Parkkinen et al., 2018, 27). Further, *bolstering* a claim is how assessment of further evidence may raise the status of a causal claim. Considering the discussion of the quality of evidence from section 1.5, we can see that established and ruled out claims are supported by evidence in which our confidence is very stable.

Importantly, this expression of what it is to establish a claim is not at odds with the view of establishing given in section 4.3. The expression given previously was that establishing ‘requires not only high confidence in the truth of the claim itself but also high confidence in its stability, i.e. that further evidence will not call the claim into question’ (Williamson, 2019, 35).

Recall from the last chapter, I am using *mechanism* in the two ways employed by Parkkinen et al. (2018). EBM+ emphasise that the mechanisms related to causal claims in medicine may be social, biological, or even technological (Parkkinen et al., 2018, 5, Clarke and Russo, 2017, chapter 9). According to Parkkinen et al. (2018, 16), in order to establish efficacy, evidence for the existence of a mechanism must support the *general mechanistic claim.* This is the claim that: ‘there exists a mechanism linking the putative cause A to the putative effect B, which explains instances of B in terms of instances of A and which can account for the observed correlation between A and B'. So, in order to establish efficacy in medicine, we have to provide evidence for the general mechanistic claim. What, then, can provide evidence for this claim? In the previous chapter, it was seen that association studies can provide evidence of mechanism. Also, I argued that in some rare instances, evidence from association studies may be sufficient to establish the existence of a mechanism. Further, evidence of mechanism can be derived from mechanistic studies (Parkkinen et al., 2018, 78). Mechanistic studies often provide evidence for the features in a mechanism by which A is supposed to cause B (Parkkinen et al., 2018, 14).

5.3 Examples from medicine

In this section, I introduce two different examples that promote, or use, evidence of mechanism alongside other types of evidence in assessing causal claims. The importance of demonstrating the explicit use of evidence from mechanistic studies alongside evidence from clinical or statistical sources is important as it marks a departure from, and improvement to, the original EBM framework, on which much of sport and sports science aims to base its practices. These cases help to serve as motivation for the idea that sports science, too, should move to a more nuanced or improved version of EBP. The IARC example is an example from medical practice. The EBM+ example, whilst drawing on practice, offers epistemological rationale for considering evidence from mechanistic studies.

5.3.1 IARC

IARC is concerned with conducting and assessing research with a view to finding environmental causes of cancer in humans. RCTs to determine the carcinogenic effects of different chemicals and exposures would be unethical in most instances.[1] Because of this, IARC draws from epidemiological studies, animal studies, and laboratory studies in order to determine the likelihood that different chemicals or exposures are cancer causing, this aids in gathering evidence of mechanism (IARC, 2019, 16, Leuridan and Weber, 2011, 92-94). As is explained in the Preamble to the IARC Monographs (2019, section 6), what is generally required to establish a claim about the carcinogenicity of a chemical is that the epidemiological evidence be sufficient to justify the claim. IARC claim that 'well-conducted cohort and case-control studies provide most of the evidence of cancer in humans evaluated by working groups' (2019, 16). However, as part of the classification procedure, the quality of evidence from epidemiological studies is assessed for bias, confounding, and chance, in order to determine how good the evidence from these studies is. Epidemiological evidence is ranked as being sufficient, limited, or inadequate to determine carcinogenicity, or as suggesting a lack of carcinogenicity. For a more detailed account, see: IARC, 2019, section 6.

By being labelled as 'sufficient', epidemiological evidence alone is considered by IARC to be enough to classify an agent as carcinogenic (IARC, 2019, 37). In these instances, animal experiments and evidence from mechanistic studies are not necessary in order to justify the classification. However, in cases where epidemiological evidence is ranked as being less than sufficient to classify a carcinogen, the other sources of evidence may be used in order to make the classification (IARC, 2019, 37, Leuridan and Weber, 2011, 96-97, Birkett et al., 2019, 343, Lauby-Secretan et al., 2016, 2222, Parkkinen et al., 2018, 102). For instance, if epidemiological evidence is ranked as 'inadequate' to make a classification, sufficient animal evidence and strong evidence from mechanistic studies may be employed in conjunction in order to classify an agent as a carcinogen (IARC, 2019, 37). Interestingly, mechanistic information can also be used to upgrade the classification of an agent's probability of being a carcinogen if it is good enough. This means that evidence of a mecha-

[1] Perhaps cases like smoking cessation studies could be considered an outlier here.

nism may take the grading from, say, *possibly a carcinogen in humans* up to *probably a carcinogen in humans* (IARC, 2019, 35). Conversely, if there is strong evidence that the mechanism which causes an agent to be carcinogenic in animals does not act in humans, this can be used to rate the agent as not classifiable as to its carcinogenicity, the lowest rating (IARC, 2019, 36). For our purposes, the important takeaway from the practice of IARC is that where the ideal type of evidence is insufficient to justify a causal claim, evidence from mechanistic studies can be considered alongside it in order to provide more warrant for the claim being made, improving our confidence in the causal claim that can be made. This is in line with the claim I will be arguing for.

Seen through the lens of the RWT, **The Excluded Explanations Argument**, and the last chapter, we can look at the practice of IARC in the following way. Sometimes, IARC considers evidence from observational studies to be sufficient to establish the carcinogenicity of agents. In these instances, the observational studies were sufficient to establish the existence of a mechanism, and of a correlation. This is because the quality of evidence is sufficient to rule out alternative explanations, and rule in the chemical or exposure under investigation as a cause. However, sometimes, observational studies cannot provide evidence of sufficient quality to establish the carcinogenicity of agents. In these instances, IARC uses evidence from mechanistic studies, in addition to evidence from observational studies, to provide greater warrant for causal claims. Here, the additional evidence from mechanistic studies is needed to rule in the chemical or exposure under investigation as a cause. In addition to this, there are times when evidence of no mechanism can be used to downgrade a carcinogenicity claim. This is the case, for instance, with d-limonene. It was found by IARC that d-limonene caused cancer in rats, and a mechanism was identified by which this occurs (IARC, 1999, 322). If it was established that a similar mechanism existed in humans, this could be taken as some evidence that d-limonene was carcinogenic to humans too. However, it was established that this cancer causing mechanism that operates in rats is not found in humans. Because of this, d-limonene was concluded to be 'not classifiable as to its carcinogenicity in humans' (IARC, 1999, 322).

It must be noted that, in most evidence hierarchies, evidence from epidemiological and observational studies, such as IARC relies on, are seen to provide lower quality evidence than RCTs. So, one could argue that as sports science relies on evidence from RCTs, an evidence gathering method of perceived higher quality, it need not employ evidence from other sources to justify causal claims. This is because the evidence from an RCT that an observed correlation is genuinely causal is often seen as being less limited than evidence from observational studies. This is because RCTs are seen as better at ruling out alternative explanations, and are often seen as being sufficient to determine causality alone. This implies that the evaluation of evidence of mechanism from other sources will not be useful or necessary in supporting causal claims. If this argument holds, using the practice of IARC to motivate the explicit evaluation of evidence from mechanistic studies alongside evidence from RCTs in the sports sciences is unwarranted. However, this criticism of my motivation would miss the mark. What I am arguing for is that: RCTs in the sports sciences will often be insufficient to provide strong evidence in favour of causal claims, as argued in chapter 2; as this is the case, evidence from mechanistic studies should also be explicitly evaluated in order to help provide stronger warrant for causal claims than could be made using evidence from RCTs alone. Of course, there may be times when evidence from RCTs and evidence from mechanistic studies contradict each other. In these cases, it would be up to expert evaluation in order to assess a causal claim, and decide how strong the evidence in favour of that claim is. In other cases, it may be deemed that, once assessed, RCTs alone are sufficient to establish a causal claim; again this is for expert reviewers to determine on a case by case basis. It is just important that a causal claim isn't accepted on the basis that the evidence in favour of it happens to come from an RCT, or set of RCTs, without assessment.

5.3.2 EBM+

The importance of the explicit evaluation of evidence from mechanistic studies advocated by EBM+ is motivated by the RWT. The model of evidence assessment proposed by EBM+ is intended to be a step forwards from classic EBM, which prioritises evidence derived from RCTs, and systematic reviews and meta-analyses of them (for examples of this, see Ashcroft, 2004). This was discussed in detail in the introduction. As well as proposing what should be done in the assessment of causal claims, EBM+ suggests a set of tools to integrate the assessed quality of evidence of correlation along with the assessed quality of evidence of mechanism. This is instead of starting from a pre-determined assumption of evidence quality based on evidence gathering method, such as is suggested by some EBM evidence hierarchies. EBM+ propose that the status of a causal claim is dependent both on the quality of evidence that there is a correlation, and the quality of evidence that there is a suitable mechanism. Accordingly, where an efficacy claim is being assessed, evidence for both claims should be explicitly assessed (Parkkinen et al., 2018, 6.3, and throughout). The status of the overall causal claim is then the minimum of the status of the correlation claim and the status of the mechanism claim (Parkkinen et al., 2018, 92). Put another way, causal claims about the efficacy of an intervention or exposure cannot achieve a 'higher status than both the correlation claim and the general mechanistic claim' (Parkkinen et al., 2018, 92). For example, if a correlation is established provisionally, and a mechanism that explains that correlation is established provisionally, then an efficacy claim supported by those pieces of evidence is also established provisionally (Parkkinen et al., 2018, 92). According to EBM+, establishing causal claims in medicine can involve assessing RCTs, which is typically recommended by EBM, but also assessing mechanistic studies. EBM+ promote the idea that evidence from mechanistic studies can help to bolster or reinforce, or even undermine, the status that would have been conferred on a claim when assessing evidence from RCTs alone. *Bolstering* or *reinforcing* deserves its own section and will be discussed in section 5.4.

There are two important distinctions about evidence types and evidence gathering methods made by EBM+ which, as they state, 'do not align' (Parkkinen et al., 2018, 93). They are key to the method of evidence assessment EBM+ promote. These distinctions have already been touched on briefly in chapter 4 where I mentioned evidence relating to mechanisms as a type or token of evidence. The first distinction EBM+ makes is that we may have evidence of mechanism, and evidence of a correlation. The second is that we may conduct clinical or association studies, which repeatedly measure A and B together, and we may have mechanistic studies, which investigate the mechanism linking A and B. Both types of study may provide evidence of each type, with varying degrees of strength. What this does not mean, is that clinical and association studies are the only method of providing evidence of a correlation, or that they only provide evidence of correlation. For instance, as was explained in subsection 4.5.2, we may conduct association studies in order to find evidence for details of mechanisms between A and B by finding evidence for variables that mediate the mechanism. Further, it also means that mechanistic studies do not only provide evidence of mechanism, and it also means that they are not the only method of finding evidence of mechanism, as I argued in chapter 4.

These claims, made by the EBM+ group, and their practical implications, can be summed up as follows. Association studies, such as RCTs, provide evidence of correlation and, often to a much lesser degree, evidence for the existence of a mechanism. Mechanistic studies provide evidence for the details of a mechanism which helps to establish the existence of a mechanism, and often to a much lesser degree, evidence of a correlation. As we need to establish both correlation and mechanism claims in order to establish a claim about causality, we will often need to assess the evidence provided by both types of study in order to have strong evidence for both types of claim. This is discussed in greater detail in section 5.4.

Whilst we are here concerned with efficacy, evidence of mechanism is also important when making causal claims about external validity within the EBM+ framework. When it comes to establishing external validity with the EBM+ framework, the status of a causal claim is supported by the status of a causal claim in a study population, the similarity of the mechanism in the study and target populations, and the status of causal claims in target populations (Parkkinen et al., 2018, 5).

5.4 EBM+ and reinforcing

In section 4.2, I explained how evidence of mechanism and evidence of correlation are needed to establish causation, as they each cover the flaws of the other type of evidence. In this section, I explain reinforced reasoning. Reinforced reasoning is, most simply put, the notion that when assessing causal claims, we should assess evidence from both association studies and mechanistic studies together. As will be explained in this section, this allows the evidential basis for causal claims to be stronger than if we assessed only evidence from association studies. This is why the EBM+ group proposes evaluating evidence from mechanistic studies, as well as association studies, when assessing causal claims.

Before I give an in depth discussion of reinforced reasoning, I must explain how reinforced reasoning differs from the RWT. If this clarification is not made, one may accidentally conflate the two and this section may appear to be a repetition of the epistemic rationale for the RWT given in section 4.2 because it discusses how we may reinforce the status of a claim by utilising evidence from association studies as well as evidence from mechanistic studies. The primary thing to remember in order to see the distinction is that the RWT is concerned with what we need evidence of, whereas EBM+ and reinforced reasoning are concerned with how we obtain that evidence. It is true that reinforcing is supported by the epistemic rationale that supports the RWT, but the two should not be conflated. Once it is remembered that association studies can produce both evidence of correlation and of mechanism, and mechanistic studies produce evidence of both types too, the difference between the epistemic rationale for the RWT, and how mechanistic studies can reinforce claims made on the basis of evidence from association studies, should be more clear. I will highlight these distinctions throughout this section for clarity. Some, for instance Auker-Howlett and Wilde (2019) do not highlight this distinction. This does not put us at odds. We do not disagree. I simply, in the context of the arguments of this thesis, find the distinction important so as not to appear to contradict myself. Highlighting this distinction, and the fact that evidence of mechanism may be derived from association studies, and correlation from mechanistic studies, is important because without making and highlighting the distinction, this section may appear to be at odds with my argument from the previous chapter in favour of evidence from association studies being sufficient to establish the existence of a mechanism in some instances.

Well conducted RCTs can provide evidence for the net effect of A on B, providing evidence of a correlation between A and B. However, as I have already argued in chapter 2, RCTs will often not be able to rule in that A is a cause of B with a high degree of rational confidence. In these instances, although they may provide evidence that there is a correlation between A and B, the evidence they provide in favour of a mechanism linking A and B is often poor because they do not provide mechanism details, and many other things may explain the observed correlation. Even in instances where RCTs or other association studies do provide strong evidence of correlation (with the obvious exclusion of instances argued for in section 4.6), though they may provide some evidence that there is a mechanism between A and B, they will be unlikely to provide sufficient evidence to establish that mechanism. As such, if only evidence from association studies is assessed when assessing a causal claim, it will often be unlikely that we will establish that claim as, though the evidence of correlation may be strong, the evidence of mechanism will often be weak.

Mechanistic studies are on the flip side of this evidence balance. Evidence from mechanistic studies can help to rule out alternate explanations of observed correlations, and rule in others, such as that A has an effect on B (Parkkinen et al., 2018, 16). Evidence from mechanistic studies does this by providing evidence for features of mechanisms. Having evidence for features of mechanisms helps to: ascertain the direction of causation; work out what may act as a confounder; identify and rule out areas from which bias may arise; and determine whether observed measures may vary temporally (Williamson, 2019, 39). Evidence for this type of feature is what helps to rule in A as causing B by showing that there is some mechanism by which A *can* cause B. High quality mechanistic studies can do this by providing high-quality evidence of mechanism, which is a tall order for RCTs and other types of association study, as was seen in the previous chapter. Importantly, as per the distinction mentioned above, mechanistic studies can provide some evidence of correlation too (Williamson, 2019, 2.2). However, in most cases, the evidence of correlation provided by mechanistic studies will be weak. It will, for instance, be unlikely to tell us much about the net effect of that mechanism in the real world. The problem is that, outside of study conditions, the mechanism uncovered in mechanistic studies may be masked by another in a way that the mechanistic studies cannot show and, as such, intervening with A may have no net effect on a measured outcome B in more complex than trial settings (Williamson, 2021, section 3). So, high quality mechanistic studies alone can provide high-quality evidence of mechanism. However, they will likely not be able to establish causal claims. This is because, although they can provide some evidence of correlation, it will often not be of sufficient quality to establish that causal claim. As such, if we assessed evidence from only mechanistic studies, the causal claim, according to EBM+, will likely not be established as the claim cannot achieve a higher status than either of the mechanism and correlation claims, and the correlation claim would not be established.

I will now explain how evaluating evidence from both mechanistic studies and RCTs together can help to boost the status of a causal claim using an imaginary example. Imagine an RCT is conducted, and the data it produces establishes that a correlation exists between A and B. Also, through thorough controlling and through strength of association, that RCT may also provide some evidence that a mechanism exists that explains how A causes B, despite giving no details of the mechanism. In this imaginary example, imagine that this evidence alone may be sufficient to only provisionally establish efficacy as it does not rule out, in principle, all other possible explanations of the observed correlation, such as that A and B share a common cause. Imagine, also, a mechanistic study is conducted that establishes the existence of a mechanism by providing the details of the mechanism by which A can cause B. In addition to this, it provides evidence that instances of B are correlated with instances of A. However, this evidence of correlation has limited applicability outside the mechanistic study as the mechanistic study is conducted in a more simple system than the one in which the mechanism would operate in real life. It is therefore insufficient to fully establish the effect of the mechanism, as it does not provide evidence for how this mechanism interacts with other mechanisms in a less idealised system. If the evidence from both studies is evaluated together, there will be evidence that establishes that a mechanism exists, and also evidence that establishes the net effect of A on B in a system more complex than in the mechanistic study. In this instance, a correlation is established, and a mechanism is established, only when we assess both types of study. Either study alone would only provisionally establish causation as they do not fully establish both correlation and mechanism alone, but, together, causation is established.

The EBM+ group makes a very insightful comparison between the strength of a reinforced concrete structure, and the strength of causal claims in medicine. The following explanation is paraphrased from Parkkinen et al., 2018, section 7.1. Concrete can be subject to great compressive force whilst resisting it, but will break easily if tension is applied to it. Conversely, steel has a very high tensile strength. When steel and concrete are used together in construction, with the steel inside the concrete, the material that combines the two, *reinforced concrete*, resists compression and tension, the resistive properties of each individual material contributing to the overall strength of the composite. In the same way, then, evidence derived from clinical studies, and evidence from mechanistic studies, both have different flaws and limitations to the quality of evidence that they may produce. These flaws come from how strongly each type of study, generally, can produce evidence of each type. However, again, like reinforced concrete, together they provide stronger evidence for a claim by covering the weaknesses of the other type of study. So, whilst either clinical studies or mechanistic studies may, alone, provide some evidence of both correlation and of mechanism, explicitly evaluating the evidence from both types of study may 'boost the status of the correlation claim to established [... and] the overall status [to] established', where evidence from clinical studies alone would simply provide poor quality evidence for a causal claim (Parkkinen et al., 2018, 94). This will become important in the following section when I argue for how this model of evidence evaluation can be applied to assessing efficacy in the sports sciences.

Before moving on, I would like to address a subtlety here. Previously, it has been mentioned that EBM+ claim that the status of an efficacy claim is both the minimum of the combination of general mechanistic claim and correlation claim, but also that the efficacy claim cannot achieve a higher status than both of those claims. This may seem at first glance to be at odds with bolstering. However, it is important not to confuse the idea that we need to establish mechanism and correlation in order to establish causation which EBM+ advocate, with the idea that evidence of mechanism bolsters the status of a causal claim conferred by just evidence of correlation, which is a misinterpretation of what EBM+ advocate. This subtlety and following possible misinterpretation, comes from the fact that mechanistic studies do not only produce evidence of mechanism, and association studies do not only produce evidence of correlation. A mechanism claim does not bolster a correlation claim, or visa versa. Both need to be established to establish a causal claim, and a claim cannot have a higher status than those individual claims. Evidence from mechanistic studies in addition to evidence from RCTs can bolster the status of a claim over what could be conferred on that claim by evidence from one type of study only. This is because, by utilising evidence from both association studies and from mechanistic studies, we improve the quality of evidence we have supporting both mechanistic and correlational claims.

5.5 Calls for evidence of mechanism in the sports sciences

The RWT provides motivation for the necessity of evaluating mechanism claims in order to establish causality. The work of the EBM+ group, and IARC, motivate the assessment of evidence from mechanistic studies, in addition to evidence from association studies. In this section, I will briefly examine two cases. They help to add weight to my argument that explicitly examining evidence of mechanisms in the sports sciences is important in assessing causal claims being made, as it allows us to rule in and out alternative explanations to observed correlations. Thus, giving us a way around the **Excluded Explanations Argument**. It will also serve as a reminder of the potential complexities of placebo controls in the sports sciences that can, if not understood, confound results.

In the last chapter, I gave an example from sports science where evidence of mechanism was assessed before a causal relationship was established. This was the case of creatine. However, reading sports science literature, it can be difficult to see explicit evaluation of evidence from mechanistic studies for proposed mechanisms underlying observed correlations. This is particularly noticeable in publications about clinical studies, which may often purport to show effects, but do not give overviews of evidence for how these effects may arise. This makes it difficult to find large numbers of case studies to highlight how evidence from mechanistic studies and evidence of mechanism together can be useful in assessing causal claims in sports science. However, something not currently being widespread practice is no strong argument that it is not potentially a beneficial practice, especially given that the practice is gaining traction in medicine. After all, sports science only shifted to an evidence-based paradigm relatively recently compared to medicine, so some amount of playing catch-up can be expected, and this is to be expected given that it has modelled itself on EBM, and not EBM+.

Fortunately, sporadic examples do exist indicating that the practice is *possible* in sports science. For instance, Saunders et al. (2017, 664) assess both evidence of mechanism and correlation together in their systematic review and meta-analysis on the effects of β-alanine supplementation on exercise capacity. This was necessary because supplementing β-alanine was associated with improved performance in exercise with durations of between 1 and 10 minutes, but not other durations. β-alanine is an amino acid that can be obtained through a normal diet or supplements. It is a precursor to carnosine, and a rate-limiting factor in the synthesis of carnosine, which in turn means that supplementing it may improve certain types of performance in sports that are affected by carnosine concentrations in skeletal muscle (Culbertson et al., 2010, 80). So, if you supplement β-alanine above what is normally available in your diet, you will be able to perform better in activities which are normally limited by amounts of carnosine in the body, as it helps to produce more. Saunders et al. utilise evidence of mechanism when investigating the effects of β-alanine supplementation on exercise capacity to rule in the effects of the amino-acid as an explanation for differences in observed effect size seen across groups that exercised for different durations when taking β-alanine supplements. There is evidence for the existence of a mechanism by which carnosine concentrations in skeletal muscle limits performance for exercise performed at some intensities or for some durations. A mechanism exists that can explain the correlation in the 1 to 10 minute exercise ranges, but not outside this. This is because, exercise performed within these ranges normally use different systems within the body to perform exercise than exercise outside these ranges. Systems within these ranges are more likely to have carnosine saturation as a limiting factor in performance. This evidence allowed researchers to explain why β-alanine supplementation was correlated most greatly with improved performance in exercise within the 1 to 10 minute range, but was not correlated with improved performance outside these ranges. Thus, evidence that a mechanism exists that can explain observed correlations between β-alanine supplementation and improved exercise performances between 1 and 10 minutes explains that the observed correlations are causal.

What is also emerging, is literature that promotes the importance of, not just having evidence that a mechanism exists, regardless of knowing its details, but evidence that leads to understanding those mechanisms and being able to provide their details. It is the work of the third part of this thesis to argue in favour of the benefits that providing mechanism details confer on sports science, but the following example is given here as it reinforces the point made by this chapter. Perhaps one of the most interesting of these is a call made by Beedie et al. (2020) for sports science researchers to incorporate findings from neuroscience, including knowledge and evidence of relevant mechanisms, when conducting research on placebo and nocebo effects. Beedie et al. do this with a view to 'help explain variability to treatments, and in doing so ... allow researchers to better understand the conditions in which treatments are likely to be most effective' (2020, 322). The article is motivated by their claim that '[w]hilst much research in sport describes positive effects on performance following a placebo treatment, most studies do not identify mechanisms' (Beedie et al., 2020, 318). According to Beedie et al., and in agreement with what has already been said in this thesis, this can lead to an inability to eliminate alternate explanations for observed correlations seen in data. One such example is a real-world phenomen seen in athletes being investigated in running trials. Depending on what wing of a trial an athlete believes they are in, they can often be observed adopting different pacing strategies in trials with outcome measures related to speed, distance, or time. This, of course, impacts observed outcomes in unintended ways. Beedie et al. explain that what this means is that if athletes believe themselves to be de-blinded, they will pace how they run differently in trials investigating the effects of interventions on running performance. Those who believe they are in a placebo wing of a trial will approach pacing differently to those who believe they are in a test wing of the trial. Whilst randomization could ensure that the belief one is de-blinded is split evenly across the trial, it can still provide an explanation, other than that the intervention under investigation is effective, for observed correlations. This means that placebos can be inadequate to properly control trials. Further, Beedie et al. claim that without investigating placebo mechanisms, research can also not explain a number of things. These include: the variation in response to placebo affects seen in trials, how the presentation of a placebo effects trial results, and the impact of placebo or treatment conditioning (what someone comes to associate the method of placebo administration with) on outcome measures in trials (Beedie et al., 2020, 318). This can, of course, lead to under and overestimating treatment effects. In sum, not knowing the mechanism by which a placebo or intervention works can lead to wrongly inferring causal relationships where it is not clear if what is intended to be used as a placebo has its effect via a placebo mechanism.

Beedie et al. (2020, 321) recommend, for instance, as an example of a time where evidence of mechanism is useful when investigating placebos in sport, the direct assessment of placebo mechanisms in trials using functional Magnetic Resonance Imaging (fMRI). This type of mechanistic study has been useful in the ongoing evaluation of glucose rinsing in sports science to determine if the observed effects associated with glucose rinsing are due to the placebo effect, or if there is another causal mechanism at play. Glucose rinsing involves rinsing the mouth with a glucose solution but not swallowing or intentionally ingesting any. This is useful, for instance, when it is used with the intention of it being a placebo. In trials testing the efficacy of ingesting glucose sports drinks on performance, rinsing is used on the assumption that it has none of the characteristic effects of glucose ingestion. Beedie et al. explain that, by using fMRI to examine the mechanism by which glucose rinsing has an effect on the body, we can determine if glucose rinsing has this effect via the placebo effect, or some other pathway. If it has its effect via some other pathway, and is then used as a placebo, it can lead to misestimations of effect sizes in trials. This could lead to either: glucose rinsing being employed as a performance enhancing practice over other, more beneficial practices, such as actual glucose ingestion, or misestimations of the effect size of glucose ingestion. Mechanistic investigations into glucose rinsing help to illustrate how regulatory processes in the body can be 'deceived' into responding to a predictable cue, like the sweet taste in the mouth when rinsing. This can be seen in that glucose entering the mouth, which normally indicates to the body that glucose will be available in the intestine, triggers a similar allocation of the body's resources to the intestine that actual glucose ingestion does. Using fMRI provides evidence that after glucose rinsing, information passes through the medulla and thalamus and *projects* onto other parts of the brain, triggering behavioural, emotional, and cognitive responses that can improve performance (Beedie et al., 2020, 321). The way in which these changes occur can be attributed, according to Beedie et al. (2020), to one of the same functions that is seen in instances of placebo effects. Whilst it is not currently settled whether glucose rinsing improves performance via the placebo effect or not, this case helps to show the importance of having evidence of this mechanism. If we understand how a placebo has its placebo effect, we can better justify its use as a placebo.

What this case helps to illustrate, in a more general sense, is that evidence of mechanism, and explicitly evaluating relevant mechanisms, is starting to be taken seriously in sports science. It must be noted, however, that this concern in sports science is an emerging concern, so this case cannot be considered to represent the current attitude of all of sports science.

5.6 Reinforcing in the sports sciences: special considerations

In this section I argue that reinforced reasoning, as it applies in medicine, can also apply to sports science. Further, I will argue for some special considerations necessary for reinforced reasoning and an EBM+-like methodology in the sports sciences that arise from the differences between the sports and medical sciences. It is important to discuss and make explicit differences between medicine and sports science. This is because not taking account of these differences may lead to the adoption of inappropriate methodologies. I previously argued, for instance, in section 2.4, that this is the case with the adoption of GRADE in sports science.

Recall that, in medicine, if one buys the reinforced concrete analogy, in individual instances where evidence from one type of study is insufficient to warrant a strong causal claim, or is insufficient to establish a causal claim, evidence from another type of study can help to establish or provide warrant for that claim. Then, if we allow that evidence from mechanistic studies can help to rule in causal explanations by providing high-quality evidence of mechanism, it will be useful to assess evidence from them in order to boost the status of claims being made in sports science, thus helping to justifiably inform practice. If we can boost the status of a claim, we are, of course, relying on better evidence, motivating the **Better Evidence Thesis**. This requires, of course, that the arguments made by EBM+, relating to medicine, carry through from medicine to sports science, and considerations about evidence, specific to sports science, that set it apart from medicine, need to be made.

In general, RCTs in sports science produce evidence that is of a low quality, and which falls foul of the **Excluded Explanations Argument** by failing to rule in a causal relationship and rule out confounders. Further, just as is the case in medicine, establishing causal claims in sports science generally requires establishing the existence of a mechanism and a correlation. Of course, if we then assess evidence from mechanistic studies in addition to evidence from association studies, we may bolster the status of the causal claim we are able to make, just as EBM+ propose. Particularly, this rests on the greater ability for evidence from mechanistic studies to provide evidence that rules in the existence of a mechanism, ruling out confounders. We will have more confidence in our claims of correlation and mechanism than if we assessed evidence from only one source. Clearly, if we have more confidence in a claim because our evidential base is stronger, we are relying on better evidence, motivating the **Better Evidence Thesis**. As such, EBP should follow in the path of EBM+ and take seriously the notion that evidence assessment should include evidence from mechanistic studies as well as evidence from association studies. However, there are important concerns with applying this reasoning, stemming from the practical and fundamental differences between sports science and medicine.

Why, then, may the arguments in favour of evaluating evidence from mechanistic studies in medicine *not* carry through to sports science? The primary concerns one may have arise from differences between sports science and medicine, and the particular difficulties associated with establishing causal claims in sports science.These are: 1) the quality of evidence provided by RCTs in sports science is often worse than the quality of evidence provided by RCTs and clinical studies in medicine so, even with the addition of evidence from mechanistic studies, we cannot establish a causal claim as we will be unable to establish a correlation claim. 2) A special case of this previous concern, particularly important in sports science is: where a sports science RCT is, in all other respects, close to ideal, it may, as they often do, have a sample size too small to generate observable small to moderate effect sizes. This becomes a problem when, if no effect size is seen, no evidence of correlation can be generated. One cannot know if larger trials would, or would not, have showed an effect size associated with the intervention. Finally, 3), unlike much of clinical science, which is largely concerned with biological responses to interventions and exposures, sports science, being concerned with a social pursuit, may often require, in addition to evidence of biological/physiological mechanisms, an increased amount of evidence pertaining to social/psychological mechanisms than medicine will. I will argue that these concerns do not provide motivation against assessing evidence from mechanistic studies; in most cases, they motivate the evaluation of evidence from mechanistic studies in addition to evidence from RCTs in sports science.

5.6.1 Where evidence is poorer than in medicine

Where evidence from RCTs is poor, high-quality evidence from mechanistic studies may be needed to establish a causal claim. Recall from chapter 2, evidence from RCTs will often be particularly poor in sports science due to the often unavoidable limitations to trial quality. This means that an evidence gathering tool that can, in some instances, be relied upon to provide high-quality evidence of correlation, and even of mechanism, may not be able to do so in many areas of sports science. Compare sports science RCTs to, for instance, RCTs in medicine. RCTs in medicine often have much more funding and much larger sample sizes. They will often be easier to placebo control and blind, as was argued in chapter 2. This means that RCTs in medicine will likely, often, provide higher quality evidence of both correlation, and of mechanism, than those in sports science. This means that to reach the same status of causal claim in sport as medicine, the strength of evidence from mechanistic studies necessary to establish the status of a claim may need to be higher than it would need to be in medicine.

For instance, evidence from a typical RCT in medicine may be very strong and only require in addition to it, say, a small amount of evidence from mechanistic studies in order to rule out alternative explanations for observed outcomes, thus establishing correlation and the existence of a mechanism and an efficacy claim. However, given the likelihood that RCTs in the sports sciences may have limitations to their ability to produce high-quality evidence, in order to secure evidence sufficient to establish an efficacy claim, researchers may require that evidence from mechanistic studies be of very high quality in order to determine causality. For instance, mechanistic studies may need to provisionally establish, or even fully establish, the general mechanistic claim alone, in order to warrant an efficacy claim where evidence from RCTs is particularly weak.

Whilst this marks a departure from medicine, it should not be considered to be a reason why an EBM+-like methodology cannot be carried through to sports science. The difficulty of establishing causal claims reflects on sports science, rather than my thesis, and the methodologies I argue for. The fact that, even though we may adopt an EBM+-like methodology, it may still be hard to gather sufficient evidence to establish causal claims does not mean that adopting this methodology does not improve the evidential basis for claims. Clearly, reinforcing the status of causal claims by additionally assessing evidence from mechanistic studies will improve the evidential basis on which those claims rest over simply assessing evidence from RCTs. In fact, the potential need for stronger evidence from mechanistic studies in sports science than medicine, if anything, stresses the importance of conducting mechanistic studies. It also stresses the importance of explicitly evaluating the combined evidence of mechanism and correlation from both RCTs and mechanistic studies in sports science, when it comes to informing practice and establishing causation. This being contrary to current evidence-based practice guidelines, which, just as in the case of EBM, dismiss or diminish the importance of evidence from mechanistic studies.

5.6.2 Sample sizes

As was seen in chapter 2, sample sizes are a particular trouble in sports science research. Evidence from mechanistic studies may also need to be very heavily relied on to determine causality in situations where trial sizes are small. This means that, where sample sizes are necessarily small, such as in elite athletes, evidence from mechanistic studies may be necessary to help provide the extra evidence necessary to satisfy claims about causality. We may compare common sample sizes in sports science trials to those in medicine. As was seen in chapter 2, a sample size of 10 is not uncommon in sports science (as is suggested in Pyne et al., 2010). In medicine, for example, a US regulation body, the FDA, recommends a sample size of 300 to 3,000 in trials to determine the effectiveness and side effects of drugs (US Food and Drug Administration, 2016). Recalling the **Harm Profile Thesis**, this number may be higher than is necessary for sports science, given the potential risks of medical drugs and need to find side effects, but note must be made of the disparity between sample sizes and how recommendations are made.

Where the sample size in a trial is too small to see small to moderate effect sizes, a trial will not be able to provide evidence of correlation for interventions with small to medium effect sizes. It will therefore not provide evidence to rule in the claim that A is a cause of B in a trial population. However, a study being insufficient in size to provide evidence for a claim is not evidence *against* that claim. If, then, every RCT on an intervention in sports science is too small to provide high-quality evidence of correlation, even when a meta-analysis is performed, such as may be the case with studies on elite athletes, those RCTs will neither help nor hinder the status of a causal claim. Again, it should be noted that this is an issue for sports science itself, rather than the methodologies advanced in this chapter, and actually motivates adopting an EBM+-like methodology.

In instances like this, it may be necessary to rely on evidence from mechanistic studies to bolster causal claims. In some instances, the evidence of correlation provided by mechanistic studies may even help to improve the status of correlation claims. Consider the following type of case. If all the group RCTs on some intervention, or in some research area are, unavoidably, insufficient in sample size to provide sufficient evidence to support a causal claim in sports science, a practice may not be sufficiently justified to be employed using evidence provided RCTs. Instead, other methods of gathering evidence must be used. For instance, whilst these RCTs may not be able to provide high-quality evidence of correlation alone, the evidence from small sample size RCTs *and* evidence from mechanistic studies together may be assessed and provide some level of evidence of correlation. The evidence from these studies, taken together, improve the status of a causal claim, even if it must still remain weak. Here, the evidence a mechanistic study can provide is relied upon, to improve the status of a causal claim. Even if, due to limitations in sports science, the status of the claim must remain low, the status is still improved over what may be achieved assessing RCTs alone.

5.6.3 Mixed mechanisms

As well as investigating interventions that are, largely, biological or physiological in nature, the sports sciences will often be concerned with interventions or exposures that have major social or psychological components. This is, in part, due to the social nature of much of sport. The prevalence and importance of the social elements of research in the sports sciences raises the importance of conducting mechanistic research in the sports sciences, as I shall argue here. The types of social mechanisms sport is concerned with also helps to illustrate some key differences between the sports sciences and medical and clinical sciences. For instance, a key subset of research in medicine, namely, whether an intervention is efficacious, does not generally involve social mechanisms. The importance of social mechanisms to sports science represents a reason why adopting EBM methodologies to sports science without considering differences between the sciences was misguided. This further indicates that research into evidence evaluation in medicine cannot be blindly applied in sports science contexts. Importantly, this means that the adoption of an EBM+-like evidence assessment methodology in sports science will often need to involve the explicit assessment of social mechanisms as well as physical and biological mechanisms, which is not as often the case in medicine. Constrained by the length of this thesis, I will not discuss any of the ongoing debates about the use of mixed mechanisms in science; for an in-depth examination, see: Russo, 2008.

The importance of assessing mechanistic studies for social mechanisms in the sports sciences can be illustrated with an example. Research into pacing strategies in a running race is an interesting example of research in the sports sciences that involves mixed mechanisms. Pacing strategy involves managing the balance of chemicals used in the muscles for movement, such that physiological failure is not reached before the end of the race, but also so that there is not an excess of energy reserves at the end of a race that could have been used to run faster (Thiel et al., 2012, 1107). Pacing in this way can be effective when the only competitor is the clock, and is often used to break world records. Deviations from a set, maximally efficient, pace tend to be minimal in world record-breaking races (Thiel et al., 2012, 1107). However, in competitive scenarios where other athletes are involved, or where podium places are the key interest, athletes may run at a pace that is less than efficient (above or below the most efficient pace at different times in the race), in order to try and gain a psychological or competition advantage by breaking away from the competition early, or having energy left for a fast finish (Thiel et al., 2012, 1107-1108). This deviation from maximally efficient pacing strategies is seen in high-level competition, such as at the Beijing Olympics, where ‘microvariations’ in pace occur even within laps, such that reporting individual lap times does not fully capture the variation in pace (Thiel et al., 2012, 1110). Mean speeds of Olympic finalists in a number of events can be seen in Figure 5.1. On the plots is also the mean lap speeds of the same events, but taken from world record races. These graphs help to illustrate how variable pace can be in podium-oriented races, compared to world record attempts. This deviation from the maximally efficient pacing strategy illustrates the importance of social and psychological elements in research into pacing strategies. If one only concentrated on physiological mechanisms, it would be difficult to explain why runners paced races differently in different racing scenarios.

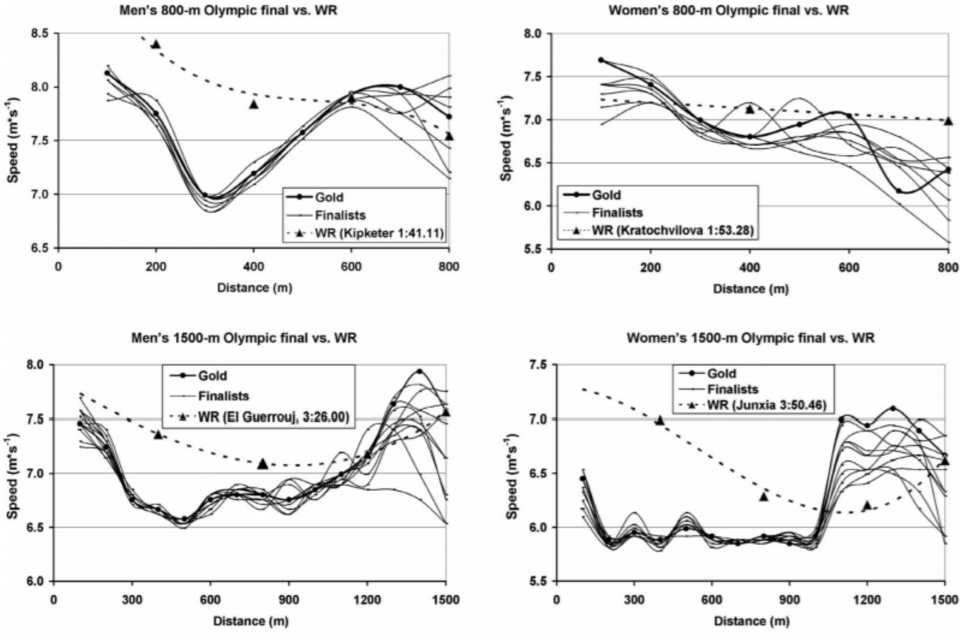


Figure 5.1: Mean speeds of Olympic finalists in 4 different track events at the Beijing Olympics, and the mean speeds of world records at the time, reproduced from Thiel et al., 2012.

In this instance, mechanistic studies concerned with social and psychological mechanisms can help to shed light into how different pacing strategies affect podium positions, whilst RCTs may be used to determine the net effects of different pacing strategies on podium places and finish times. Conducting mechanistic studies concerned with only physiological and biological mechanisms, in this instance, as in much of sport, would mean that it would be difficult or impossible to explain differences in performance strategies between podium focused and record focused events. Further, understanding relevant social mechanisms is important to help us understand how and when to employ different interventions, like pacing strategies, in practice.

A note must be made here about the importance of social mechanisms in medicine. I do not want to falsely claim that social mechanisms are of no importance to medicine. Social mechanisms can be relevant to causes of illness, adherence to treatment, and recovery from illness, for instance. The point I want to make here is not that social mechanisms are not important to medicine, it is that they are also important to sports science, and given the social nature of sport, their importance may be relevant more often than in medicine.

5.6.4 A brief digression on informing practice

It may be the case where, due to limitations to evidence gathering in the sports sciences, no means of gathering evidence are sufficient to establish some types of causal claims. For instance, it may be difficult to establish the claim that one intervention is more effective than another in some instances, such as the one I will give below. In these instances, we cannot provide good justification for our practices. This is because, along EBP guidelines, we should ideally only practice based on established causal claims. This may be the case, for instance, where sample sizes are small. Adherence to the RWT does potentially put incredibly high standards on the quality of evidence required to establish causality, standards that may often be difficult to meet. This is a problem if, as sports science has practical ends, we cannot engage in many practices as the available evidence is often insufficient to establish efficacy. But, surely, we should want the standards of evidence required to establish causality to be high, as was explained in the previous chapter.

In these instances, it may seem that practice may never be justified reliably. However, given, as has been argued in chapter 3, the relatively small size of, and lack of harm caused by side effect of sports interventions, practice in some instances may be possible to be informed with slightly lower quality evidence. It may be acceptable, in some instances, to conduct practice in sport that is not established as causal. Considering the **Harm Profile Thesis**, this is because there is normally less risk if you get it wrong. In addition to this, there may be cases where it is better to employ some practice that is not particularly well justified, rather than employing no practice at all. One sports intervention may be slightly less effective than a different intervention, or another may be ineffective but not harmful to health or performance. This is unlike medicine where interventions, and their harm profile, can be outright dangerous. The negative consequences of losing an amateur football game because a team's stretching routine is poorly evidenced and happens to be slightly less effective than a different option is much different to losing a life because a slightly less effective antiviral is used. Of course, this is not ideal. Athletes may be wasting their time performing useless interventions, but at least it is unlikely that they are also dangerous.

This is, of course, a slightly hasty generalisation and is not always the case; for instance, elite athletes may risk more by utilising ineffective or less effective interventions, such as jeopardising their one shot at an Olympic title. A case I raised in the introduction to this thesis also provides an example of the dangers of improperly informing practice in the sports sciences. The promotion of over-hydration lead to a number of deaths. Practising based on less-than-established causal claims may also present a high risk in extreme sports, such as if the evidence supporting braking technology in racing cars is not sufficient to know they will stop a car when they have to. But, in many instances, we can probably get away with informing practice with a slightly less-than-established causal claim in sport than medicine if we have to, where the risks associated with getting it wrong are minimal. Interventions with a poor evidence base are actually often used, even by high-level athletes, in sport. For instance Olympic Swimmer Michael Phelps famously employed 'cupping' as part of his training. Cupping is the practice of creating a vacuum seal between the skin and a hollow 'cup' so that the skin reddens inside the cup as blood vessels expand, and the skin starts to be sucked up into the cup. It is a practice that has little to no evidential support, and is often considered to have no effect on performance (Beedie et al., 2018, 817-818). I am not, of course, advocating that it is good to practice based on poor evidence. Perhaps Phelps would have been even faster if he had not utilised cupping, but we can certainly see the risks are minimal compared to medicine. Of course, in all cases, strong evidence should be striven for when informing practice, but this may be setting the goal posts impossibly high in some instances.

This line of reasoning may also be used to reach a complimentary conclusion. It may also be acceptable, in sport, to employ a practice justified by low-quality evidence that one intervention is more effective than another. Imagine there are two practices, α and β, with strong evidence supporting a beneficial effect on the same outcome which crosses some minimum effectiveness threshold, but an athlete may engage in only one. Imagine also, that in addition to this strong evidence that each practice has a similar size of effect, there is some weak evidence that α is actually more effective than β at improving this outcome. Given that we have strong evidence that α is at least as effective as β, and some evidence to suggest that α may, in fact, be better than β, would it not be sensible to adopt α based on this weaker evidence?

Examples of this type of line of reasoning can be found in the literature. This is exactly how the reasoning plays out in practice with regard to suggested protein requirements of male strength athletes, for instance. There is strong evidence that a moderate consumption of protein is sufficient to meet the average dietary protein requirements of these athletes to maximise improvement in performance. However, there is also some weaker evidence that a much higher dietary protein intake may slightly improve performance over a moderate protein intake (Bandegan et al., 2017). As a high dietary protein intake will likely be at least as effective at improving performance as a moderate dietary protein intake, and there is some evidence to suggest it may be slightly more effective, it is suggested by some, for instance Bandegan et al. (2017), that the recommended protein intake for strength athletes should be high, rather than moderate.

5.7 Conclusion

In this chapter, I argued that, in order to best justify causal claims in sports science, we should generally assess evidence from mechanistic studies, as well as association studies, as it will better help us to fulfil the goal of relying on the best possible evidence. This is the case because evidence from RCTs in the sports sciences is often of low quality, but evidence from mechanistic studies can help to cover the limitations to evidence in RCTs. Primarily, evidence from mechanistic studies, by providing higher quality evidence of mechanism than can usually be derived from RCTs helps to rule in, or out, exposures or interventions under investigation as causal. From this, we can clearly see that, as the joint assessment of evidence from both methods of gathering evidence can provide us with greater warrant for causal claims, it better fulfils the goal of relying on better evidence, motivating the **Better Evidence Thesis**.

The idea that we should assess evidence from mechanistic studies in order to bolster causal claims was informed by the practice of IARC and EBM+ in medicine, and recent research in sports science calling for the assessment of evidence of mechanisms when investigating interventions. In addition to this, I discussed some special cases in sports science where assessment of evidence from mechanistic studies may be particularly useful. This included cases where sample sizes are unavoidably small, and cases where evidence from RCTs is particularly poor.

This brings an end to Part II of the thesis, and the bulk of the argument that assessing evidence from mechanistic studies is important, in the sports sciences, as a means to establish causality motivating the **Better Evidence Thesis**. Part III of this thesis is concerned more with the importance of providing details of mechanisms as it leads to understanding mechanisms, not just establishing their existence.

Part III

Giving mechanism details a sporting chance

Introduction to Part III

In Part II of this thesis, the main thrust of my argument was that evidence that a mechanism exists is essential to establishing causation in sports science, and that a solution to the problem whereby RCTs in the sports sciences often produce evidence of insufficient quality to establish causation is to assess evidence from mechanistic studies in addition to association studies. In Part II of this thesis, I also defended the RWT, and the claim that evidence from RCTs can be sufficient to establish the existence of a mechanism. An implication of these claims is that it is, in theory, possible to establish the existence of a mechanism without knowing the details of that mechanism. For instance, in chapter 4, I gave the example of caffeine, the ergogenic effects of which were established before the details of the mechanism giving rise to these effects were known.

In Part III of this thesis, I argue for a further claim that may be considered even more controversial in the eyes of some EBP proponents than the claim that we should assess evidence from mechanistic studies when establishing causality. In Part III, I argue that, not only should we assess evidence from mechanistic studies in order to better assess causal claims, but, in many instances, we should also seek to provide details of mechanisms underlying causal relationships. This is so that we can better understand the mechanisms that explain those causal relationships. Being able to explain causal relationships is, of course, an important part of science broadly: deepening our theories explaining causal relationships. Evidence-based methodologies, however, often disregard the importance of understanding mechanisms as less important than evidence that establishes a causal claim. Largely, this is because evidence-based methodologies are concerned with informing practice, and are far less concerned with being able to explain *why* that practice has its proposed effect. Opposed to this way of thinking, I will argue that the utility we gain by providing details of mechanisms goes beyond helping to establish causation, and should be motivation for EBP to take seeking to provide details of mechanisms seriously. For instance, by helping to explain why interventions are ineffective, using our understanding of these mechanisms, we can better research and develop new, effective interventions. As such, the arguments of Part III of the thesis continue the thread begun in Part II of this thesis, that EBP should take the assessment of evidence from mechanistic studies seriously.

In order to make this argument, I will present two in-depth case studies. One case study looks at a football injury prevention programme, and its successes and failures. This case study mainly motivates the importance of understanding the biological and physical mechanisms by which an intervention works. In this chapter, I proceed by arguing that without providing details of intervention relevant mechanisms, we cannot adequately explain why the intervention is effective in some groups and not others. I further motivate this claim by arguing that if we understand the relevant mechanisms we can improve, adapt, or change the intervention to be effective in different populations. The second case study looks at exercise interventions for obesity. This case study focuses more on the importance of understanding social and psychological mechanisms in sports science. Presenting the case studies not only supports the importance of understanding mechanisms, it also goes some way towards motivating the feasibility of conducting this kind of work.

In both of these case studies, I rely on evidence both from mechanistic and association studies. This is done, first, by examining association studies and their findings on the intervention for each case study. Upon reading the case studies at these points, it may appear that I am relying solely on evidence from association studies like RCTs for my case studies, something I argued against in chapter 2. In the background and exposition sections of these chapters, I do mainly rely on publications of this type, but this is because it is publications of this type that contain the bulk of the discussion of these interventions. However, as the chapters progress, I also support causal claims I make by examining relevant mechanisms, and explaining that we can consider them to be established.

In chapter 8, I argue that we can extend the lessons from chapter 6 and chapter 7 naturally to the general case. This is not a repetition of the arguments from the previous two chapters: rather, I show that we can extrapolate the arguments from the previous chapters, and their lessons. I do this by drawing on the works of Gillies, and Anjum and Mumford, who provide motivation for the importance of understanding how causal relationships arise, and also discuss the benefits of understanding the mechanisms underlying them. I finish the chapter by giving some broad practical benefits of providing details of mechanisms underlying causal relationships in sports science, and arguing that as these benefits align with goals expressed by EBP, that EBP ought to take providing details of mechanisms seriously.

Understanding and providing the details of a mechanism

In this part of the thesis, I am concerned with the idea that we should seek to understand mechanisms. By this I mean we should provide *details* of mechanisms. That is, identifying components of a mechanism that interact together in order to give rise to the causal relationship we are interested in. I am therefore really concerned with the importance of deepening our understanding of relevant mechanisms, not providing a full explanation of those mechanisms. In line with this, if we are able to identify key causal factors relevant to some causal relationship, on my view that counts as providing the details of mechanisms as it deepens our understanding of how a cause gives rise to an effect. As such, objections to the view that we should understand mechanisms that attack it on the grounds that we cannot know if we fully understand a mechanism will not hold. This also allows me to sidestep the issues of reducibility and emergence when it comes to discussion of mechanisms. For the purpose of this thesis, and the types of use I argue that providing details of mechanisms elicits, it is not important to know the fundamental physics underlying a mechanism, or if a social mechanism is emergent or reducible to physical properties.

Understanding mechanisms at what may be seen as a coarse grained level may give rise to worries about invoking mechanism descriptions that are otherwise masked in a way we do not know in a larger system. I discuss this objection in section 8.6 where I once again invoke the RWT as a solution to this type of worry.

Chapter 6

Case study I: the FIFA 11+ and the importance of understanding mechanisms

6.1 Introduction

Football is the world's most played sport. In the most recent worldwide survey of football participation, it was found that over 240 million people play football in Fédération Internationale de Football Association (FIFA) registered countries (FIFA, 2007). As football has a very high injury rate, FIFA has promoted an Injury Prevention Programme (IPP) with the aim of helping to reduce this injury rate: the FIFA 11+ (Bizzini and Dvorak, 2015). Evidence from a number of high quality RCTs suggests that the FIFA 11+ is highly effective at reducing injury rates in some populations. Because of this, it is promoted by many FIFA member nations (Bizzini and Dvorak, 2015). However, despite being highly promoted and endorsed, and being supported by RCT evidence, it fails to significantly reduce injury rates in some key populations (Bizzini and Dvorak, 2015; Gatterer et al., 2012; Hammes et al., 2015; van Beijsterveldt et al., 2012). Previously, in 2000, FIFA promoted simply 'The 11'. The 11+ is a revised version. The proposed intervention mechanisms are similar enough that I will always assume the use of the revised version. This was necessary as the surrounding literature often doesn't make clear which version they are discussing, or even that there are two versions of the intervention. The key groups I examine however, veteran males, skilled males, and young females were all tested using the 11+ scheme.

In this chapter, I argue that providing details of the mechanisms relevant to the FIFA 11+ is important because: 1) it goes some way towards helping to explain why the intervention is ineffective for some populations, 2) it helps us to understand where we can, and cannot, apply an intervention, and 3) it can help to guide future research by helping us to see how we could adapt the intervention into a new intervention that would be effective in the populations in which the original was not. I will do this by highlighting how physical and social mechanisms relevant to the 11+ can differ between effective and ineffective groups, and explaining how they affect the effectiveness of the IPP and how they can be used to determine how to improve it. I will use examples of physical mechanisms: strength and neuromuscular ability. I will also discuss some social mechanisms relevant to the intervention's success: adherence and motivation. This motivates the importance and utility of being able to provide details of mechanisms.

6.2 Background

Football has one of the highest injury rates for any sport (Gatterer et al., 2012; van Beijsterveldt et al., 2012; Wong and Hong, 2005), with rates as high as 45 injuries per 1000 hours played being reported, most of these in 'informal play' (Scanlan and MacKay, 2001, 145). To put this into perspective, a three-year study of acute and overuse injuries in high-level US intercollegiate play in multiple sports found football (in their wording soccer) to have an injury rate of over 230 incidences per 10,000 athlete exposures, which puts it in the highest injury rate bracket with wrestling ($\sim$190 injuries per 10,000 exposures) and field hockey ($\sim$210), where other sports such as woman's swimming and diving ($\sim$16), rowing ($\sim$50), and woman's cross-country and track and field ($\sim$25) have significantly lower injury rates (Yang et al., 2012).

Both the popularity of football and high instance of injuries per hour in football make it a prime candidate for the research into, and development of, IPPs. In a recent literature overview of IPPs for football, it was found that around 85% of these were exercise-based (Bricca et al., 2018). An exercise-based IPP being one where exercises are performed in order to train the athlete physically to avoid injuries. This can be compared to, for instance, an equipment-based IPP that could encourage the use of helmets or ankle braces. One such exercise-based IPP, introduced in association with FIFA, is the FIFA 11+. The 11+ is a warm-up routine intended to reduce the injury rates in amateur players (Bizzini and Dvorak, 2015). Evidence from RCTs was considered sufficient to establish that the intervention was effective in significantly reducing injury rates for young female players. Later, further large scale RCTs were deemed sufficient to establish that it is effective for some other male and female player groups (Bizzini and Dvorak, 2015; van Beijsterveldt et al., 2012). The method by which the FIFA 11+ seems to work is by training strength, functional balance, and neuromuscular control in the lower extremities through exercises performed during a set warm-up routine (Bizzini and Dvorak, 2015).

Unfortunately, despite evidence from RCTs that the 11+ is effective for some groups, trials show either no significant correlation between adopting the 11+ and reduced injury rates for older male soccer players in the 'veteran' age category (Bizzini and Dvorak, 2015; Gatterer et al., 2012; Hammes et al., 2015; van Beijsterveldt et al., 2012).[1] In addition to

[1] This is not a fixed age category, but often begins around age 35 and over.

this, there appears to be no significant correlation between adopting the 11+ and reducing injury rates for male players at an ‘intermediate’ and above skill level (Gatterer et al., 2012). When we consider that a good number of RCTs do not find significant correlations between intervention and reduced injury rates in these populations, and the fact that mechanistic evidence exists (which I explain in section 6.3) that can explain why the 11+ is not effective in these populations, we can consider the evidence base that the 11+ is ineffective in these populations to be relatively strong. It is important that adult men seem to not have reduced injury rates when the 11+ is adopted, as they make up the largest at risk group for injuries in football (van Beijsterveldt et al., 2012)[2] and despite its lack of effectiveness, the 11+ is still promoted to these groups. The problem of properly applying sports injury research from one group to a different one, and even just from research into effective practice, does not only exist in football, it is a current major concern in sports science and injury prevention research as a whole (for examples see: Bahr and Krosshaug, 2005; Chalmers, 2002; Finch, 2006, 2011; Hanson et al., 2014; Hanson et al., 2012). Highlighting this problem, in a 2010 analysis it was found that only 492 of 12,000 manuscripts published regarding sports injury actually examined the effectiveness of IPPs. This means that the effectiveness of most IPPs devised is unknown, and even worse than this, only 162 articles addressed their proper effective implementation (Klügl et al., 2010).

It can be considered to be established that both strength, and neuromuscular control and skill, which the 11+ intends to intervene on, are causal factors in injury rates in sports in general given that the links are discussed in sports injury textbooks used to teach in universities (see for example: Fu and Stone, 2001, 704). Further, much of the IPP literature regarding football specifically suggests that injury rates can be

[2] The adult male population is the largest at risk group as there are more players in this group than others. This comment does not invalidate that the injury rate is higher in female players.

decreased with the application of strength, neuromuscular, and proprioceptive training in the warm-up and pre-season (see for instance: Croisier, 2004; Croisier et al., 2008; Heidt et al., 2000; Hübscher et al., 2010; Soligard et al., 2008), even by as much as one-third (Lauersen et al., 2014). Further, knowledge that these things *can* improve injury rates is used to devise and test IPPs. What we can take this to mean then is that we can take it as established that if we improve strength, and neuromuscular and proprioceptive control, we can reduce injury rates. What this also tells us is that where the 11+ is effective in reducing injury rates, it is effective in virtue of improving these factors. I will explain this claim in greater depth in section 6.3. Given this, what I will argue is that in this case, understanding mechanisms helps us to understand and explain why the 11+ is effective in some populations and not in others.

Before moving on, I need to make a number of clarifications. I need to adopt a definition of injury and explain injury mechanisms. I also need to explain how we measure injuries and how we can compare injury rates.

6.2.1 Mechanisms of injury

In order to be able to measure injury rates and to attempt to reduce them, we need to know what an injury is. I am limiting my discussion of the cause of injuries to the biomechanical pathways within the body that explain the capacity for injuries to occur. Another possibility would be to take a holistic view of all causal factors, including the sociological, physiological, and psychological causes of injuries. However, these give rise to many multi-causal and non-linear relationships that require complex causal systems to explain how injuries occur (Bittencourt et al., 2016; Philippe and Mansi, 1998). Understanding injuries in terms of these complex relationships both beyond the scope of this chapter, and beyond what the 11+ attempts to intervene on.

Defining sports injuries can be contentious. I will adopt a stance given broadly in the relevant literature. A sports injury is very commonly given as an incident in play or training that causes a player to be unable to fully partake in some future play or training (Keller et al., 1987; Petersen and Hölmich, 2005; Soomro et al., 2015). It is useful to be able to compare how long an injury causes one to be unable to play or train. This means that we can measure injuries and their severity indirectly and quantitatively based on time missed from play, rather than having to analyse injuries in and of themselves. This includes injuries such as head injuries that would require an athlete to miss perhaps the end of a game, and severe bone breaks that may mean missing an entire season. There is obviously a deficiency here in that some injuries may make one miss a short amount of play, whilst being more serious than others. A concussion (or repeated concussions) may be worse, but require less time off, than a badly twisted ankle, for instance. However, a benefit of this definition over one that includes things like small grazes and blisters is that it allows us to focus on injuries that cause a loss of time in play or that impact sport-related performance and health rather than superficial injuries which are more of an annoyance (Keller et al., 1987). There is another problem with this definition. It does not include types of injury that would seem in ordinary circumstances to be substantial but would not stop play, a fracture of the finger for instance. However, as much of the IPP literature uses this definition, and key 11+ research conducted by Soligard et al. (2008) uses it, I will adopt it as well despite this problem.

These time-loss injuries are caused when damage is done to the body. Biomechanically speaking, this damage occurs when the loading on the body is higher than its tolerance (Bahr and Krosshaug, 2005; Croisier et al., 2008; McIntosh, 2005; Wong and Hong, 2005). Bahr and Krosshaug, in their review paper on understanding injury (2005), encourage us to use the commonly accepted injury understanding of C.F. Fung, the father of modern biomechanics. On this view we should treat injury as being 'equivalent to the failure of a machine or a structure' (2005, 325 Fung, 2015). This view takes into account both the properties of tissues, and characteristics of loading on them. This helps to explain how the transfer of energy causes injury. On this view, injury occurs when the ultimate strength of tissues in the body is exceeded when they undergo stress and strain, causing damage, just like in a machine (Bahr and Krosshaug, 2005). There are external and internal mechanisms that can cause the body to undergo loading above tolerance. These can be put into seven broad categories: (Bahr and Krosshaug, 2005, 326)

1. Impact
2. Overuse
3. Structural Weakness
4. Lack of Flexibility
5. Overload in movement
6. Imbalances in strength between muscles
7. Fast growth.

To give an example: in football, a common mechanism by which the hamstring is injured is that during the leg swing (such as occurs whilst kicking a ball) there is an increase of hamstring tension whilst lengthening the muscle. If this increased load on the hamstring is greater than its tolerance, injuries like strain may occur (Croisier et al., 2008; Petersen and Hölmich, 2005). This injury-causing overload can be due to, for instance, strength imbalances between the quadriceps and the hamstring where muscles used in knee extension (straightening of the leg) for the swing produce more force than the hamstring can tolerate (Croisier et al., 2008).

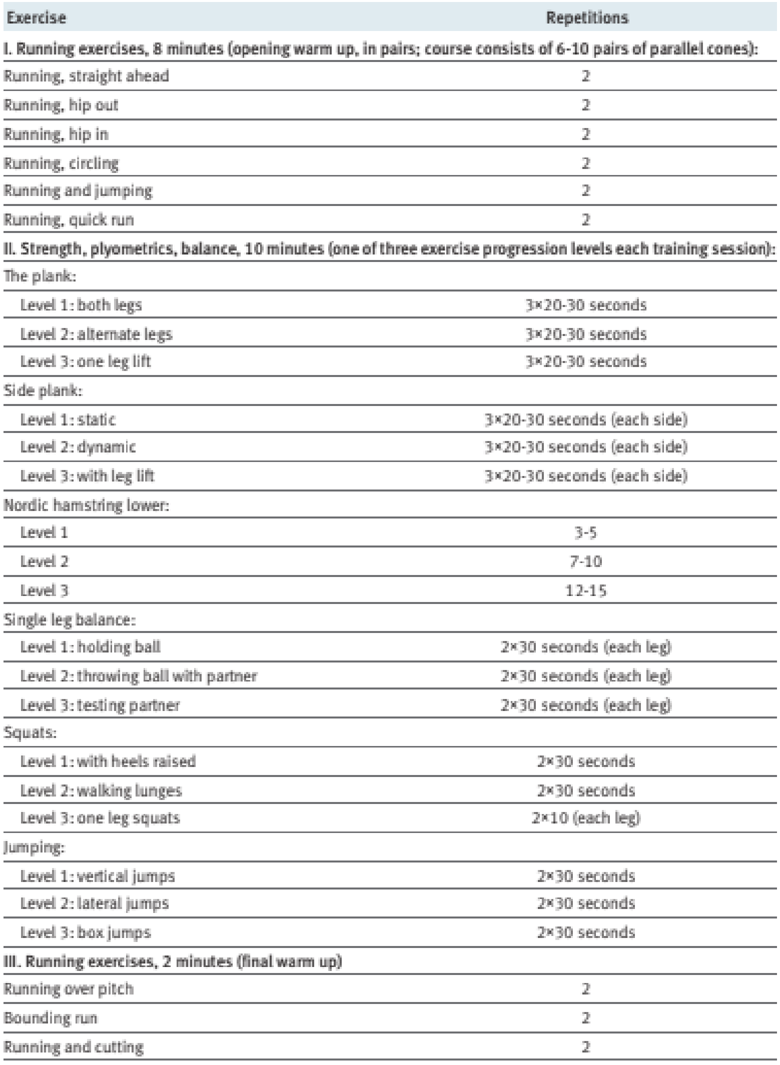
As an IPP that aims to reduce injury rates, it makes sense, given the mechanism of injury, that the mechanical function of an IPP is to increase the body's load tolerance, and to reduce the amount of potentially deleterious loading that the body has to undergo. Key elements of the 11+ are that it aims to train: strength, athletes to move in ways that reduce injurious forces on the body, and the body's ability to undergo sudden directional changes (Soligard et al., 2008). This could, for instance, counteract overload in movement and strength imbalance injury mechanisms. Assuming the 11+ is able to do those things, there is a biomechanical basis to its proposed effectiveness.

6.2.2 Injury rates

In order to make inter-sport and intra-sport injury rate comparisons, we need a way of measuring how frequently injuries occur. Two of the most common are by comparing either injuries per given number of player appearances (e.g. 10,000 exposures), or injuries in a given number of hours (typically 1000) (Phillips, 2000). It is also possible to measure the severity of injuries. This is normally measured by the amount of time missed from training or play as a result of injury (Phillips, 2000). Unsurprisingly, when measuring injuries in football it is found that most football injuries received are to the lower extremities like the legs, feet, and ankles (Scanlan and MacKay, 2001; van Beijsterveldt et al., 2012; Wong and Hong, 2005). Interestingly, female players have been observed to have a higher injury rate than male players. However, there are no observed differences between injury rate in training for adolescent and professional players regardless of gender (Wong and Hong, 2005), although this is not to say that causes of injuries are the same in these groups.

6.2.3 The 11+ in some depth

The exercises in the 11+ focus on balance, stability, and hamstring strength. The running exercises are intended to warm athletes up and improve control of the knees and core, whilst also performing landing and cutting movements which are common in football (Soligard et al., 2008, 2-3). The hamstring and core strength exercises include an increasing difficulty progression, so intensity may increase with ability to some extent. All the exercises included are based on previous intervention studies, which suggest improvements in injury rates if they are performed (Soligard et al., 2008, 8). Athletes are encouraged to focus on quality of movement, stability of the core, and maintaining alignment and control of the hip whilst performing the 11+ (Soligard et al., 2008, 3).



Exercise	Repetitions
I. Running exercises, 8 minutes (opening warm up, in pairs; course consists of 6-10 pairs of parallel cones):	
Running, straight ahead	2
Running, hip out	2
Running, hip in	2
Running, circling	2
Running and jumping	2
Running, quick run	2
II. Strength, plyometrics, balance, 10 minutes (one of three exercise progression levels each training session):	
The plank:	
Level 1: both legs	3×20-30 seconds
Level 2: alternate legs	3×20-30 seconds
Level 3: one leg lift	3×20-30 seconds
Side plank:	
Level 1: static	3×20-30 seconds (each side)
Level 2: dynamic	3×20-30 seconds (each side)
Level 3: with leg lift	3×20-30 seconds (each side)
Nordic hamstring lower:	
Level 1	3-5
Level 2	7-10
Level 3	12-15
Single leg balance:	
Level 1: holding ball	2×30 seconds (each leg)
Level 2: throwing ball with partner	2×30 seconds (each leg)
Level 3: testing partner	2×30 seconds (each leg)
Squats:	
Level 1: with heels raised	2×30 seconds
Level 2: walking lunges	2×30 seconds
Level 3: one leg squats	2×10 (each leg)
Jumping:	
Level 1: vertical jumps	2×30 seconds
Level 2: lateral jumps	2×30 seconds
Level 3: box jumps	2×30 seconds
III. Running exercises, 2 minutes (final warm up)	
Running over pitch	2
Bounding run	2
Running and cutting	2

Figure 6.1: The exercises and number of repetitions of each exercise included in the 11+ warm up scheme, reproduced from (Soligard et al., 2008, 3).

The primary measured outcome of the 11+ in early trials was any injury to the lower extremities sustained after the first performance of the IPP. Researchers also considered injury rates in other body parts secondary outcomes. A relevant injury was defined as one that occurred during a match or training session which caused a player to 'be unable to fully take part in the next match or training session' (Soligard et al., 2008, 6). In groups with the highest level of compliance, an initial RCT observed that the rate of injuries was 35% lower for those prescribed the 11+ than in groups where no warm-up routine comparable to the 11+ was performed (Soligard et al., 2008, 5). The exercises chosen for the 11+ were supported by evidence from previous association studies, and the rationale for choosing them is supported by an understanding of how these injuries come about (Soligard et al., 2008, 7-8). However, it is notable that the original RCT paper on the 11+ admits that:

> Our prevention programme is multifaceted and addresses many factors that could be related to the risk of injury ... it is not possible to determine exactly which exercises or factors might have been responsible for the observed effects. (Soligard et al., 2008, 8)

6.2.3.1 Why examine the 11+

It is worth mentioning a previous, and similar, case study to my examination of the FIFA 11+ as some interesting comparisons can be drawn. In Cartwright and Hardie's book 'Evidence-Based Policy: A Practical Guide to Doing it Better' (2012), they examine a case study: the Bangladesh Integrated Nutrition Policy (BINP), a policy which was intended to reduce child malnourishment rates in Bangladesh.

A nutritional policy employed in Tamil Nadu dramatically reduced rates of malnourishment in children. A very similar policy, when employed in Bangladesh, had little to no effect on the levels of nourishment in children. The nutrition policy was adopted in Bangladesh based on the evidence produced in Tamil Nadu. Its effectiveness in that population was taken as evidence it would be effective elsewhere (Cartwright and Hardie, 2012). This is similar to what occurred with the 11+. The fact that it was seen to be effective in some populations led to it being prescribed for use outside those populations. Again, like the 11+, when the intervention was applied to a different population it was found to be ineffective. Cartwright and Hardie argue that just because a policy works in one population, it does not mean that it will necessarily work in another, the BINP being their example of this. The claim I am interested in, which Cartwright and Hardie also push, is not just that the same intervention may not be effective in different populations, it is that understanding relevant mechanisms can help us explain why an intervention may be effective in one population and not in a different one.

Both myself, with my examination of the 11+, and Cartwrght and Hardie with the BINP, are doing very similar work. We examine some causal differences between populations where interventions had varying levels of success and use these differences to explain why there were varying levels of success. Cartwright and Hardie examine the factors that differ between child-rearing in Tamil Nadu and Bangladesh, which lead to the differences in effectiveness of such similar policies in changing child nutrition levels (Cartwright and Hardie, 2012). I compare differences in factors between older and skilled males, and adolescent female soccer players and how these differences influence changes in injury rates as a result of the 11+.

The case studies have subtly different thrusts to their arguments. In their work, Cartwright and Hardie compare the differences in social structure between Tamil Nadu and Bangladesh, which lead to the BINP not increasing the level of nutrition of children. For example, they examined differences in the social factors that lead to how food distribution is organised in families in Tamil Nadu and Bangladesh, such as who in the family is in charge of meals. They did not focus on the actual physiological mechanisms of malnutrition. However, this is not a fault of the investigation. It is reasonably well established that if children eat more, they will be better nourished. My case study of the 11+ deals with both physiological and social mechanisms related to the intervention's proposed effect and the way that they are interrelated and interact with each other.

The aim of the BINP case study is to examine policymaking decisions and how to do this correctly in a broadly social context. My examination of the FIFA 11+, however, is more concerned with the explanatory importance of mechanisms than it is with policy (although, understanding mechanisms may influence sports policy decisions). For this reason it was important to examine a case study that included both physiological and social/psychological mechanisms, as both are key to sports and sports science. This includes examining things such as how motivation (a broadly social/psychological mechanism) will affect exercise quality and intensity (a primarily physiological mechanism). This interrelatedness of social and physiological mechanisms makes the 11+, not just a better case for examining mechanistic evidence in sports science, but it also makes for a more interesting case than the BINP.

6.3 Understanding mechanisms

Now that I have illuminated the mechanism of injury, how injury rates are used, and the theory behind the 11+, I can go on to argue in favour of the importance of understanding mechanisms in sports science. Using the 11+ as motivation for my stance, I present three positions:

1. Understanding the mechanisms behind the 11+, and understanding how these differ between different populations can help us understand why, despite RCT evidence vouching for its effectiveness in trial populations, it can be ineffective at reducing injury rates in

others.

2. Using this knowledge, and by comparing mechanisms across populations, we can also identify which populations the application of the 11+ would likely be effective in, and which it would likely not.

3. Further, understanding the mechanisms by which the 11+ does work in some populations, and why it does not work in others, can help us adapt and improve a new version of the 11+ for groups where the original was ineffective. This can help guide future research. We can see what changes we may want to make to an intervention before testing, so we have some idea it may work, and some idea why. This can potentially reduce the number of studies that need to be conducted in order to find an effective intervention.

In this section, I will detail some injury-preventing mechanisms in the 11+. Using an understanding of relevant mechanisms, I will suggest how we can use knowledge of these mechanisms to explain why the 11+ was ineffective in some populations. I will use this as a basis for arguing that understanding mechanisms can help us explain why an intervention is ineffective in some groups. Then, in section 6.4, I will give empirical examples of how we can use our understanding of mechanisms to adapt an existing intervention into a new intervention, which could be tested for effectiveness in populations in which the 11+ is not effective. These two sections both provide motivation for the claim that understanding mechanisms by being able to provide their details can benefit more than mere understanding.

A complete understanding of the mechanisms and factors that lead to sports injury is difficult because they have a "multifactorial nature" (Bahr and Krosshaug, 2005, 324). Multifactorial in that many complex social, psychological, physical, and biological factors are needed to give a complete explanation of the causes of sports injury. Whilst a full explanation of all relevant mechanisms is impossible here, I have included examples of both physical and social factors that could be responsible for the 11+ being able to significantly reduce injury rates. This will be useful when, in chapter 8, I take the 11+, and obesity intervention examples, and extend the lessons we draw from them to the general case because they give a wide range of the types of important mechanisms in sports science.

Allow me to posit three key factors which were present in successful test groups where the injury rate was reduced:

1. Athletes who performed the FIFA 11+ improved strength in thigh muscles and improved core stability over those who did not (van Beijsterveldt et al., 2012, 1117-1118).

2. Improved motor patterns, proprioceptive skill, and dynamic stability were seen in athletes performing the 11+ (van Beijsterveldt et al., 2012, 1117).

3. Players adhered to the IPP regime, with coaches helping to ensure compliance (Soligard et al., 2008).

This is a non-exhaustive list of factors that, when improved (or met, as in the case of adherence) by the 11+, improved injury rates. Here, I will address the factors in turn. I will examine the mechanism by which they can be improved by the 11+. I will, then, suggest how understanding these mechanisms, and how they differ between populations, can be used to explain how outcomes as a result of the 11+ differ between populations.

6.3.1 Strength

Sufficient evidence has suggested that improvements in strength can reduce injury rates in sports-people that the claim is now a key part of injury prevention research, and, as such, much IPP research investigates how to utilise this claim by researching methods of improving strength (Croisier et al., 2008; Heidt et al., 2000; Lauersen et al., 2014). As part of the program, the FIFA 11+ includes hamstring strengthening exercises as this is a regularly injured area. Notably, the Nordic hamstring curl is employed, which is an eccentric hamstring strengthening exercise, a diagram of which can be seen in Figure 6.2 (van Beijsterveldt et al., 2012). One set of three to fifteen repetitions is completed during the 11+ warm-up programme, based on ability level (Soligard et al., 2008). There is evidence for the 11+ effectively improving strength in some groups: a review found that proxy measures for strength like improved agility skills and jumping, and knee strength ratios, are seen with the adoption of the 11+ (Bizzini and Dvorak, 2015).

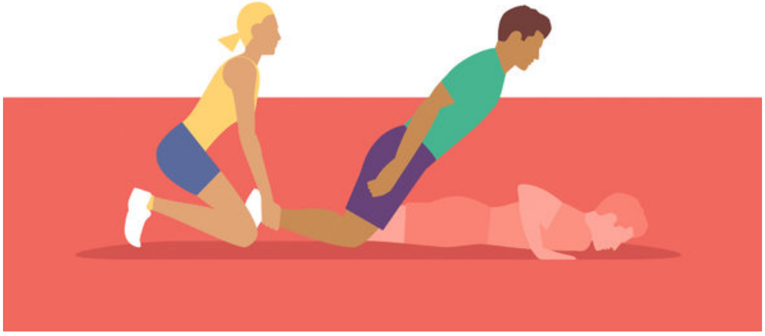


Figure 6.2: A diagram of the Nordic hamstring curl exercise reproduced from https://well.blogs.nytimes.com/2014/06/05/the-great-hamstring-saver/ illustration by Ben Wiseman

As mentioned earlier, a common injury in football players is hamstring injury. This regularly occurs when the leg is swung and the muscles that straighten the leg produce more force on the hamstring than it can tolerate. This can lead to a strain injury (Croisier et al., 2008; Petersen and Hölmich, 2005). There is a good base of evidence that suggests that improving hamstring strength can reduce injury rates (Arnason et al., 2008; Mjølsnes et al., 2004; Petersen et al., 2011). The claim is supported by evidence from RCTs, and also evidence of mechanism. Heiderscheit et al.

(2010) explain the mechanism by which this works. If muscular imbalances that lead to tension on the hamstring reaching or nearing tolerance levels can be reduced by increasing the strength of the hamstring, this is because the hamstring's peak strength and longer muscle lengths from improved strength and size helps to 'offset the concentric action of the quadriceps' (Heiderscheit et al., 2010, 78). In simple terms, this means that if we can sufficiently strengthen the hamstring in relation to the muscles that straighten the leg, the leg straightening muscles will not be able to produce enough force that it can damage the hamstring by pulling it to its longest length too hard and fast. The fact that there is statistical and mechanical evidence to suggest that improving hamstring strength can reduce injury rate suggests that if the 11+ is successful in intervening on hamstring strength, it will also intervene on injury rates.

Strength training works by stimulating muscles in three ways: applying tension mechanically, damaging muscle fibres, and causing metabolic stress to the muscles. A muscle undergoing these kinds of stimulation will then be caused to have a *hypertrophic* response. This is a response where the muscle is caused to grow in order to be better able to cope with these stresses in the future (Kraemer et al., 1998; National Strength & Conditioning Association, 2015; Schoenfeld, 2010; Zatsiorsky and Kraemer, 2006). Intensity of exercise (the percentage of weight loaded on an exercise relative to the maximum weight one could lift for a single repetition) has a significant impact on the level of hypertrophic response (Ahtianinen et al., 2003). Schoenfeld, one of the most highly cited resistance training researchers worldwide[3], argues that exercise intensity is 'the most important exercise variable for stimulating muscle growth' (2010, 2863). The importance of intensity is an issue for the 11+: Beijsterveldt et al. (2012) suggest that the lack of reduction in injury rates in some adult male player groups could be the product of insufficient exercise intensity to generate sufficient strength gains to reduce injury rates. The athletes were already too strong for the 11+ to increase their

[3] For a ranked list, see: https://www.expertscape.com/ex/resistance+training

strength further. If the 11+ is not intense enough to increase strength, then there will not be a reduction in injury rates from an increase in strength.

6.3.1.1 Intensity: a mechanical aside

I believe that the mechanism behind hypertrophy, and how it is affected by intensity, is important to this discussion. However, it is on the technical side and therefore deserves its own subsection.

A number of pathways exist explaining why intensity is a key to maximising muscle growth responses from exercise (Zatsiorsky and Kraemer, 2006). I will detail the most important mechanism that explains this, drawing largely on Zatsiorsky and Kraemer, 2006 (chapter 4). In order to best achieve a hypertrophic effect, we need to raise amino acid uptake above resting levels. This allows for a greater synthesis of contractile proteins, leading to muscular growth and strength increases. This is achieved when we force, with heavy resistance training, the muscle to supercompensate the amino acid uptake (National Strength & Conditioning Association, 2015; Zatsiorsky and Kraemer, 2006). Supercompensation is where, after undergoing stress (such as performing resistance exercises), the body attempts to return to normal working conditions by adapting to stress by 'compensating' for it by doing things such as improving muscle size, strength, and energy stores. If this is managed correctly, the body instead 'supercompensates' for the stress and further improves athlete ability. The graph in Figure 6.3 shows how expending energy during exercise and reducing muscle growth and repair temporarily will lead to an increase of muscle size over time due to this supercompensation effect. In the graph, one can see that the level of protein synthesis is higher after exercise than before exercise, which means that muscle growth is increased after the exercise. This happens because, by forcing a muscle to use more of the fixed amount of energy it has at any one time for mechanical work and not protein synthesis during heavy resistance training, the muscle catabolizes (breaks down) more muscle protein than it synthesises. After exercise, the muscle compensates for this by increasing amino uptake levels and synthesising more protein in order to be able to perform movements again in the future (National Strength & Conditioning Association, 2015; Zatsiorsky and Kraemer, 2006). This is explained graphically in Figure 6.4. This increased protein synthesis of course leads to muscle growth: hypertrophy (Zatsiorsky and Kraemer, 2006).

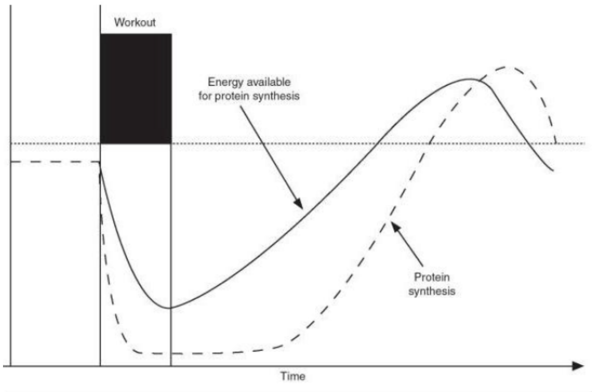


Figure 6.3: The rate of muscle anabolism after a training bout as a function of time (reproduced from Zatsiorsky and Kraemer, 2006, 53).

Intensity plays a role as, in order to maximise muscle growth, we need to find the correct balance between muscle breakdown and energy for synthesis, whilst performing work to maximise amino acid uptake. With a relatively small resistance, lots of energy is used for the mechanical work of the muscle and there is little muscle breakdown. Conversely, if the resistance is relatively high, the total energy used is lower whilst the catabolism increases. ‘The total amount of degraded protein ... is a function of both the rate of protein catabolism and the mechanical work performed’ (Zatsiorsky and Kraemer, 2006, 73). We want to maximise the amount of degraded protein in order to increase supercompensation and the hypertrophic response (Zatsiorsky and Kraemer, 2006, 71). Because of this, intensity of resistance exercises should be moderate. As is stated in a key masters level textbook used in strength and conditioning courses Zatsiorsky and Kraemer, 2006, ‘as a rule of thumb, no more than

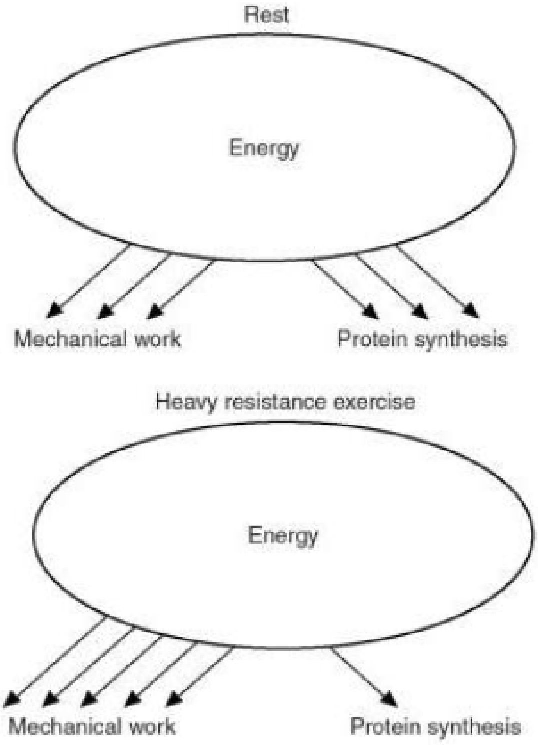


Figure 6.4: A pictorial explanation of why increasing the amount of energy used for mechanical work will reduce the amount available for protein synthesis (reproduced from Zatsiorsky and Kraemer, 2006, 52).

10-12RM should be used for muscular strength development' (71). 10-12RM, being an exercise performed with the maximal weight one could move for between 10 and 12 repetitions.

If we take this mechanism into account when discussing the 11+, it becomes apparent why the Nordic Hamstring exercise may not improve the strength of some athletes. When an athlete already has strong hamstrings (for instance through years of high-level play) there is a high likelihood, as suggested by Beijsterveldt et al. (2012), that the exercise is not intense enough to promote strength gains. This is because the repetitions required will be easily performed. If the intensity isn't sufficient to induce a hypertrophic response and improve strength, the exercise will not contribute to a reduction in injury rates.

This lack of intensity in the 11+ becomes more apparent when compared to an example of a trial where the Nordic curl *was* seen to improve strength, and was determined to be a causal factor in reduced injury rates in skilled, and highly trained, football players. It included performing almost double the repetitions prescribed by the FIFA 11+ over multiple sets, as well as increasing the load on the hamstrings once the number of repetitions was insufficient to improve strength (Mjølsnes et al., 2004). An increased number of repetitions as well as loading the Nordic curl beyond body-weight clearly amounts to a much greater level of intensity and volume.

What I have done here is to illustrate that these mechanisms are established and well understood. I have, then, used the understanding of these mechanisms, and how intervening on them effectively may require different intensities and volumes of exercise, to explain why the FIFA 11+ is ineffective in some groups. What this does is provide motivation for the claim that understanding mechanisms, not simply establishing their existence, is very important in sports science research. I will advance this line of argument in the following subsections, giving different examples of mechanisms where our understanding of them helps to explain the effectiveness gap.

6.3.2 Neuromuscular training

Part of the way that the 11+ is proposed to work is by improving neuromuscular skill, proprioceptive ability, and functional balance (Soligard et al., 2008; van Beijsterveldt et al., 2012). It is theorised that young players may benefit greatly from this as they likely haven't fully established movement patterns yet (Soligard et al., 2008; van Beijsterveldt et al., 2012), and are often less skilled (Keller et al., 1987). This means that it will help them train non-injurious movement patterns. Improved functional balance, muscular activation, and quicker stabilisation times have been found, in a review (Bizzini and Dvorak, 2015), to be seen in some populations who adopted the 11+ for whom it was effective at reducing injury rates. This suggests that, as these outcomes can be used to measure neuromuscular adaptions, the 11+ *can* train neuromuscular and proprioceptive ability in some populations.

Neuromuscular training works by changing and enhancing motor responses that are performed unconsciously (National Strength & Conditioning Association, 2015, Risberg et al., 2001). The idea is that with this kind of training, people can improve: their ability to fire muscles optimally; stability; and muscle responses to joint forces (Cerulli et al., 2001; Risberg et al., 2001). One way in which this happens is that athletes can be trained to use different strategies of movement in order to change how stress is applied to the body (Cerulli et al., 2001; Risberg et al., 2001). As stated earlier: it is above-tolerance stress that causes injuries. So, to be effective at reducing injury rates as a result of neuromuscular training, an IPP must be able to train the body to change how it moves under load from a more to a less deleterious movement pattern, and to cope with more stress, increasing maximal tolerance.

The more efficient one becomes in a neuromuscular sense, the fewer motor units need to be recruited to perform a movement. As this efficiency increases, a greater load is required to tax and therefore cause an improvement in the system (Bompa and Buzzichelli, 2015). This improvement in efficiency explains why the FIFA 11+ has a much smaller effect on reducing injury rate in adult male players competing at an intermediate or above skill level than those at lower skill levels (Gatterer et al., 2012). People with higher skill levels have better proprioceptive abilities (Hrysomallis, 2007; Paillard et al., 2006; Paillard and Noé, 2006), and a superior athlete will have better muscle coordination and control as a result of this neural adaptation (Zatsiorsky and Kraemer, 2006). As younger players are likely to be less skilled (Keller et al., 1987), and more skilled players are likely to have better proprioceptive abilities, there are two considerations with regard to injury rate that the mechanism of neuromuscular ability can help us explain:

- Skilled athletes may be sufficiently trained in neuromuscular ability that the 11+ is insufficiently taxing to improve this ability where it is sufficient in unskilled and younger players.
- A reason for the 11+ not significantly reducing injury rates in skilled players is that the initial injury rate is already conditional on a current high level of proprioceptive ability, unlike in young and less skilled players.

The greater someone's proprioceptive ability, the more challenging training will need to be to improve it. Therefore, the ideal training suitable for people with low levels of proprioceptive ability is likely different to that suitable for people with higher levels of proprioceptive ability. So, where skilled players do not have their proprioceptive ability improved, they will not see a reduction in injury rates. This is because their injury rate is already conditional on high levels of proprioceptive ability. This is because a skilled player's muscles will already react well to stress and have high tolerance for the IPP exercises.

6.3.3 Compliance

In the introduction to this thesis, I discussed, briefly, the difference between effectiveness and efficacy. When we move from establishing the efficacy of an intervention to establishing if an intervention is effective, we need to address a range of key contextual factors (Hanson et al., 2014). Having discussed two physical factors, I will now examine a key social contextual factor for any intervention: compliance, or adherence. If the target population does not actually follow the intervention, then it cannot be effective. This seems obvious, but adherence should be a key concern when promoting an intervention, particularly an IPP, if we want it to translate from research to public use (Hammes et al., 2015; Lauersen et al., 2014).

In the RCTs where the 11+ was shown to have significant effects, such as when trialled on adolescent females, the warm-up was performed two to three times a week (Hammes et al., 2015) and with around 77% of training and match sessions using the warm-up (Soligard et al., 2008). Compliance was helped in the original young female based RCT by encouraging coaches to ensure the 11+ was carried out (Soligard et al., 2008). This is a useful tactic as players may wish to start playing sooner, omitting the warm-up entirely. In comparison to the level of compliance in RCT groups where the 11+ IPP was successful, the veteran male football players studied only trained and employed the IPP once per week (Hammes et al., 2015). This means that the 11+ was performed fewer than half of the amount of times it was performed in RCT groups where it significantly reduced injury rates. Veteran male players also showed decline in motivation to perform the warm-up as time passed (Hammes et al., 2015). As may be expected, these players showed no significant change in injury rate (Hammes et al., 2015). Clearly, the level of compliance is a factor here. These players are not to be confused with the population of players of intermediate and above skill level discussed in Gatterer et al., 2012.

There are two ways in which we can be interested with how compliance affects an intervention:

1. Assuming people comply, will the intervention have an effect?
2. Assuming the intervention has an effect, will people comply to it?

For an intervention to have an effect outside of trial conditions, both factors must be met. If an intervention has no effect, it does not matter how much people do it; further, it does not matter if an intervention has a huge potential effect if no one adheres to it. We have evidence that the 11+ is effective in some populations who complied. We also have some evidence the intervention was ineffective in some groups who did comply. However, in the case of veteran male players, we do not have evidence of whether the intervention would have been effective if they had complied. What we do know is that they did not comply regularly to the intervention. To even get evidence about whether an IPP will be effective in a population, given compliance, it is crucial that the population will comply to it. It is particularly worrying that they did not comply in a trial. It is generally understood that adherence is better in trials than out of trials (Osterberg and Blaschke, 2005, 487).

Compliance is a key social factor affecting the capacity for the 11+, once promoted, to reduce injury rates. If the IPP were effective if athletes complied, and athletes comply, there will be a reduction in injury rates. There are many social factors that influence compliance. These factors will make up part of the complex system mechanism that gives rise to compliance. This system will include things such as motivation, time, energy, danger if the intervention is not complied with, whether a coach enforces the intervention, and perceived usefulness. It will obviously differ between individuals. Some may find perceived usefulness more concerning than the fact that it takes time to do. Some may only engage in the intervention if a coach makes them.

Understanding some key causal factors that make up this mechanism can help us explain why the intervention is adhered to well in some groups, and not in others. Which, of course, explains some reasons why the intervention was effective in some groups and not others. Hammes et al. (2015) suggest some of these factors that can impact an older player's ability and motivation to perform an IPP. A key one that is unlikely to be present in adolescent female players is job-related commitments. Clearly if *real life* commitments outweigh, in the minds of players, the importance of performing the IPP, then 11+ compliance will be low. If compliance is low, it will not be regularly performed and cannot cause the strength or neuromuscular improvements, which in turn cause a reduction in injury rates. A job being a priority for players is an example of a factor that plays a role in the complex system that influences compliance. Intuitively, we prioritise different things based on their perceived importance to us. For a player to perform the 11+, or any IPP, they must be motivated to do it, and they must prioritise it enough that it will be performed sufficiently to have an effect.

6.3.4 The importance of explanations

In the previous subsection, I have given examples of three different mechanisms relevant to the effectiveness of the 11+. Two physiological, and one social. Through my examination of these mechanisms, I have illustrated how, by looking at differences between populations, and using our understanding of intervention relevant mechanisms, we can explain why the 11+ was highly effective in some populations, and yet ineffective in others. We can see that the difference in effectiveness between populations is largely explained by whether the 11+ was actually able to intervene on key factors responsible for injury rates. It appears to have effectively intervened in some populations, and not others.

In this chapter, I have given concrete examples of the usefulness of understanding mechanisms. Without understanding the mechanisms behind hypertrophy, neuromuscular training, and compliance, we would not have a clear idea why something deemed to be so effective at reducing injury rates in one population should be ineffective in others. By taking the mechanisms out of their RCT black-box, and shedding some light on them, I have been able to give reasons why the IPP was ineffective in some populations, rather than simply having to say that it is ineffective in those populations. If we relied solely on evidence from RCTs, even if they were sufficient to establish a causal relationship in some populations, we would not be able to say *why* there was no causal relationship in others.

Debates relating to mechanisms based extrapolation in the philosophy of medicine, and in the philosophy of public policy, are clearly important here. This case provides an example where providing details of intervention-relevant mechanisms is useful because it helps us understand where an intervention may or may not be effective and why. It also helps us to differentiate between populations that may be viewed as homogeneous (veteran aged men, and skilled adult men) in their lack of response to the 11+, but which fail to have their injury rates reduced for different reasons. I discuss this in more detail in section 8.5.

Some may object that we don't need to *understand* mechanisms to tell us that an intervention won't be effective outside a trial population. They may say that we could simply conduct many high quality RCTs in that population to determine if it is effective. We could then conduct a number for each population in order to tell us where to apply it, and where to not apply it. This is true. In cases where this is possible, RCTs could give us evidence that a mechanism exists, without details, and evidence of correlation, establishing causation. This is in line with Part II of my thesis. However, if one has an idea that a mechanism differs sufficiently between populations that an RCT must be conducted in a new population to determine if an intervention will be effective, it would be useful to also make that mechanism, and the evidence for it, explicit. I will explain the use this gives us, beyond the importance of understanding, in the next section before expanding on the argument in chapter 8.

6.4 Mechanisms and adapting interventions

In the previous section, I gave examples of some factors that are part of the complex network that can reduce injury rates. In this section it will be seen that, similarly to my previous argument in favour of providing details of mechanisms, if we rely purely on evidence from RCTs and association studies, and shun evidence that helps us explain a phenomenon mechanistically, we would miss useful evidence that can help us adapt and improve intervention. There is nothing in evidence from association studies of this type which tells us how, given an intervention's ineffectiveness, we should attempt to improve it. This work is done by evidence from mechanistic studies. This helps to motivate again the importance of understanding mechanisms. I discuss this again in the next chapter, as it relates to exercise intervention for obesity, and again in section 8.6, making a more general case.

6.4.1 Strength

We may know that hamstring training is correlated with reduced injury rates because of RCT data, and we know that there is an established mechanism by which increasing hamstring strength can reduce those injury rates, but when looking at IPPs, it is useful to use our knowledge of the mechanisms of strength training to determine if, and where, IPPs will be effective. If we want to improve strength in order to reduce injury rates, we must make sure that the way we want to intervene on strength (the 11+) *can* improve strength. As the hypertrophic response of exercise is heavily influenced by the intensity of muscle stimulation, we must ensure that the intensity of the exercises in an IPP are sufficient for the groups which it is applied to. One of the ways in which we may attempt this would be to tailor strength components of an IPP to the strength of the participants. In the case of the 11+, this could include a greater external load during the Nordic curl exercise to increase relative intensity.

If the 11+ can be adapted to be sufficiently intense in terms of hamstring strength training (such as to a level suggested to be effective mentioned earlier in the chapter (Mjølsnes et al., 2004)) to promote a hypertrophic response for each group that adopts the scheme, this will be beneficial for the strength of the hamstrings. This change in intensity will then cause the hamstring to be able to tolerate greater loads during training and play. If the hamstring can tolerate greater loads, the likelihood of injury is reduced and the injury rate should decrease as a result of increased intensity of hamstring training. Whether the ways in which the intervention is changed have a net beneficial effect would still need to be tested to see if a correlation exists, perhaps with further RCTs, but without understanding these mechanisms, we would not know what about the 11+ could be changed to potentially improve injury rates in ineffective populations.

6.4.2 Neuromuscular training

As neuromuscular adaptions occur, fewer motor units are needed to perform a movement as a greater level of efficiency is reached (Bompa and Buzzichelli, 2015). Therefore, to increase performance with respect to ‘ever-increasing system efficiency’ loads need to be increased (Bompa and Buzzichelli, 2015, 30). If we look back to the background section at Figure 6.1, which lists the exercises in the 11+, we see very minimal load increasing, especially not in exercises aimed at improving the neuromuscular system and proprioception (notably the balance exercises and running exercises). The result of this is that players with a greater training age and or skill level with good proprioceptive skills will already be efficient at the proprioceptive exercises, thus reducing their training effect. In order to be more effective at challenging neuromuscular ability, an examination of the mechanisms relating to the interventions that are intended to tax this system could be undertaken, with the view to adding more levels of complexity, thus increasing proprioception in already skilled players and, therefore, causing a reduction in injury rates in that population.

6.4.3 Compliance

It is easy to blame the players for not performing the IPP, but it is wrong to place the blame completely on the players. As suggested earlier, athletes may have good reasons to prioritise other things in their life over the 11+. A good IPP will have methods of ensuring or encouraging participation. Successful IPPs often have compulsory participation or teaching. For instance, the RugbySmart programme in New Zealand, which reduced spinal injury rates from the scrum in rugby to 11% of the expected value, requires coaches to complete the course yearly. It has a reach of almost 100% (Quarrie et al., 2007). Compared to this, outside of trials, the 11+ has limited reach. In 2013 only 25% of FIFA member associations promoted the 11+, and only about 5000 coaches worldwide had been instructed in how to perform it (Bizzini et al., 2013). Given that football is played by over 200 million FIFA registered people, there is clearly a deficiency in IPP dissemination, here.

Hammes et al. suggest other ways to increase participation in the 11+ which can improve compliance (2015). Looking at the social mechanisms at play with the 11+ and considering that they are largely time related, in their paper they propose a modified programme that can be performed at home so that it can be done in spare time rather than dedicated time. They also suggested adding to the 11+ ball handling and individual skills in order to improve the lack of variation that leads to low enthusiasm levels.

6.5 Conclusion

In this chapter, I introduced and explained the FIFA 11+ IPP. I explained that this IPP is highly effective at reducing injury rates in some populations, but not in others. I argued that if we were to rely on only evidence from RCTs, we cannot explain the difference in effectiveness between these populations. Further, I argued that examining details of the relevant mechanisms allowed us to explain these differences. I, then, argued that we may use this information in other ways in addition to merely explaining differences in outcomes. We may use this information to help adapt and develop interventions that may be effective in groups which the original intervention was not. In the next chapter, I make a similar argument, using the case of exercise interventions for obesity. Importantly, this next case study places greater emphasis on the importance of understanding social and psychological, as well as physical, mechanisms in the sports sciences. This is valuable given the importance of social and psychological mechanisms to sports science. Finally, in chapter 8, I extend the argument of the two case studies to make the argument for the importance of understanding mechanisms more general.

Chapter 7

Case study II: exercise interventions for obesity

7.1 Introduction

In the last chapter, I employed the case of research into the FIFA 11+ injury prevention programme to begin to motivate the importance of not just knowing the existence of mechanisms relevant to causal relationships in sports science, but providing the details of those mechanisms. This chapter further motivates the importance of providing mechanism details in sports science, but with an increased focus on the importance of social and psychological mechanisms. In this chapter, I introduce the case of exercise interventions for obesity. In the case of these types of interventions, there is often a significant difference in effect size between supervised and unsupervised interventions. Often, unsupervised interventions are insufficient as tools to manage obesity, where supervised interventions are sufficient. Just like the last chapter, I argue that we cannot provide a good explanation for these differences in effect size without examining details of intervention relevant mechanisms. I further argue that if we do not consider social and psychological mechanisms, alongside key physical mechanisms, when attempting to explain observed differences in effect size, any explanation we provide will miss important aspects key to that explanation. I do this by arguing that exploring relevant physical mechanisms cannot fully explain the observed difference in effect size. I, then, argue that if we examine the details of relevant social and psychological mechanisms, this can help to explain and inform our understanding of the observed difference in effect size. I argue that if we examine the two types of intervention (supervised and unsupervised), and determine that supervision is a key factor that drives differences in effect size, we can determine that supervision does this by ensuring adherence to the intervention and effort used when partaking in the intervention. Further, I go on to argue that we can use details of social and psychological mechanisms to help aid intervention design. This is particularly important in the case of obesity interventions, as will be seen. This is because the problem of obesity is widespread, and what looks to be the solution to ineffective interventions, supervision, is not a practical one given the scale of the problem. However, an understanding of why supervision is beneficial to interventions can help us develop other, more practical, interventions to ensure adherence and effort in exercise.

The following question may be raised: why am I considering obesity interventions to be under the remit of sports science when obesity looks to be a public health concern? In this instance, it is fair to consider these obesity interventions as part of sports science for two reasons. Firstly, the focus of this chapter is on *exercise* interventions for obesity specifically. As exercise and training are a key part of sport and improving performance, and because sports can be used as a way to facilitate exercise, it is clear why the link can be made to sports science. Secondly, as explained in the introductory chapter to this thesis, I interpret sports science in a broad sense that includes exercise science. A key part of exercise science is the use of exercise in the treatment and management of physical and mental illnesses and diseases, like obesity. From this definition, it can easily be seen why interventions which utilise exercise in order to combat obesity fit under a broad interpretation of sports sciences.

There are important questions associated with how we understand mechanisms in a social context that should be raised, even though there is insufficient space in this thesis to address them thoroughly. Notably, Beach (2021, 1) suggests that adopting different views of social mechanisms can render enquiry that uses evidence of mechanism alongside evidence of correlation 'easy but superficial, very productive but challenging, or almost impossible'. At the very least, he suggests, this type of project is more difficult in the social than the natural sciences. This difficulty is related to both what degree we 'unpack' mechanisms, and whether we are realist about the social structure underlying these mechanisms or whether we simply evaluate how well theories correspond to the observable world (Beach, 2021). Beach's points arise from the fact that evidence of mechanisms is not uniformly characterised in the social sciences (Beach, 2021, 3).

Beach claims that there are two important distinctions we must make when discussing how we should understand social mechanisms:

> The first distinction relates to whether mechanisms should be decomposed into their constituent parts in order to evidence mechanistic claims, or whether the theoretical mechanism can be analytically grey-boxed, focusing thereby almost solely on the epistemic question relating to how a linkage can be evidenced (2021, 3).

The second distinction Beach makes is as follows:

> The second distinction relates to the degree to which theories of mechanisms—if unpacked at all—take seriously the particular nature of social phenomena and the epistemological consequences that flow from this, as in realist approaches

> to the study of mechanisms, or whether more neopositivist-based foundational assumptions are adopted that result in a focus on theorizing the more directly observable aspects of mechanisms without attempting to tap into the distinct social dimension of human interactions within causal processes (2021, 3).

Beach claims that where we stand on these distinctions impacts what the analytical benefits of studying mechanisms are. What Beach identifies is that the more productive we wish our understanding of a mechanism to be, the more in-depth our understanding of mechanistic claims needs to be, and the more challenging this will be for researchers. Considering Part II of this thesis, and what was argued is needed to establish a causal claim, if all we wanted to do was assess a causal claim, *grey-boxing* seems to suffice those ends. If we can provide sufficient evidence that a linkage exists (in addition to the necessary evidence of correlation), we can establish a causal claim. Part III of this thesis, of course, diverges from Part II, and is, instead, concerned with how understanding mechanisms can be useful in sports science. As such, in line with Beach, it does not seem sufficient to grey-box mechanisms. So, to what extent should we seek to provide the details of mechanisms when, as in the case of EBP, we have practical goals? To this, I say we should seek to provide the details of intervention relevant mechanisms insofar as it is useful to the outcomes we are concerned with. In some instances, we may achieve greater practical outcomes if we seek to provide the details of mechanisms in greater detail; in others, we may achieve our practical ends with a much less fine-grained understanding of intervention relevant mechanisms. As such, as the answer is so often in this thesis, the question of how far we ought to investigate and provide the details of a social mechanism in order to achieve our practical ends is left for experts to determine on a case by case basis.

On Beach's second point, the question of whether social and psychological mechanisms are actually reducible to physical mechanisms, or whether there is something distinctly social about them, is not necessary to discuss in the context of this thesis. Whether social and psychological mechanisms can be decomposed into physical mechanisms, we are able to fruitfully discuss those mechanisms at the social or psychological level without making reference to, say, fundamental physics. That this can be fruitful is shown throughout Part III of this thesis. What is important, then, is whether we should, in our assessment of mechanisms, treat social and psychological mechanisms as reducible to physical mechanisms, or, if we should consider them at the directly observable social level. As was the case with Beach's last point, the level of decomposition at which we find our understanding of a mechanism most fruitful will differ on a case by case basis. In the case of EBP, what is important is providing the details of a relevant mechanism at the suitable level in order to best achieve practical ends that EBP is concerned with. If we have reason to believe that an intervention may be useful on the basis of understanding a mechanism at a broadly social level, it may be fruitless to examine that mechanism at the level of brain chemistry. Hypothetically, for instance, we may notice that a team of athletes performs better when they receive positive rather than negative reinforcement. We may also be able to identify some mechanism that explains this at a social level, perhaps that teams who receive positive reinforcement are more likely to work together and are less likely to play as if they are the only person on their team. In this instance, although it may be possible to understand this mechanism at the level of brain chemistry, the benefit of doing so may not outweigh the cost of attempting to do so if it delays or prevents the application of a positive reinforcement intervention. This of course does not preclude that we may possibly understand these mechanisms at a more fine-grained level. Alternatively, there may be instances where our understanding of a broadly social mechanism becomes more fruitful when, either, we consider it at a more fine-grained level, or, seek to develop a fine-grained account of that mechanism in addition to the high-level understanding.

In terms of Beach's view, what these responses mean is that how we choose to understand social mechanisms does vary the difficulty of understanding those mechanisms, and the analytical benefit of understanding those mechanisms. However, at least in the case of EBP, deciding where to stand on the distinctions he raises can be assessed on a case by case basis. This is, if anything, in line with what Beach notes, that social mechanisms are not uniformly characterised in the social sciences.

7.2 Background

In this section, I discuss why using the case study of obesity interventions is important, why BMI is used as an indicator of obesity, and the importance of adherence.

7.2.1 Why an obesity case study?

In this section, I detail the extent of the health concerns caused by obesity in individuals, and the cost to the public as a result of obesity. According to the World Health Organisation, obesity is the accumulation of excess body-fat that may impair health (2021). Tackling obesity is also one of the main public health priorities of the NHS Long Term Plan in the UK (2019). This puts tackling obesity in the same category of importance as smoking, cancer, and antibiotic resistance. Obesity is now considered to be an epidemic disease. The World Health Organisation has called it 'one of the most serious public health challenges of the 21st century' (World Health Organisation, 2015). By school year 6 (ages 10 to 11), one in three children in the UK is considered to be overweight or obese (The Association of UK Dieticians, 2018). Once obesity is established in a child, it tends to track into adulthood (The Association of UK Dieticians, 2018). In 2011, in the UK, the adult obesity rate was 23% with an additional 38% being categorised as overweight, rendering the UK one of the most obese nations in Europe (Department of Health and Social Care, 2011b). Unless the obesity trend is curbed, UK government projections estimate adulthood obesity will reach 60% in men and 50% in women by 2050, with an additional 35% of adults being overweight (Department of Health and Social Care, 2011a). Obesity costs the National Health Service of the UK £6.1 billion a year, and it is estimated that the overall costs to the wider economy reach £27 billion yearly, although it is not explained how the cost to the wider economy is calculated (Public Health England, n.d.). Public health England also reports that obesity is projected to cost the NHS over £9 billion yearly by 2050.

As a result of the excess adiposity (fat tissue) that characterises obesity, an obese individual is part of a population with an increased risk of health complications over those outside of obese populations (Khaodhiar et al., 1999, 17). Aside from a general increase in mortality, comorbidities associated with obesity include: insulin resistance, type 2 diabetes, cardiovascular disease, sleep apnea, stroke, and osteoarthritis (Khaodhiar et al., 1999, 17). Whilst obesity is an excess of body-fat, this is hard to measure. To aid measurement of this, the Body Mass Index (BMI) scale is often used to help categorise obesity. Weight classes are often stratified according to BMI, which is used as a predictor of body-fat levels. It has the unit of body mass per metre squared of body. So, body-fat levels can be used to identify obesity, which is particularly useful in the case of individuals. However, BMI is more widely used than actual body-fat percentage for practical reasons, which will be discussed. Overweight is classified as a BMI of between 25 and 29.9 and obesity is categorised as a BMI of at least 30 (McArdle et al., 2008, 451). One may take umbrage with the idea that someone with a BMI of 30 is considered to be obese, and one with a BMI of 29.9 is simply overweight, as this seems to arbitrarily put one in a more at risk population with a body weight difference that may be tiny. The same seeming arbitrariness occurs when stratifying according to body-fat percentages, too. A man with 25% body-fat is obese, and a man with 24.9% body-fat is not. However, it is not the object of this chapter to argue for or against how obesity is categorised.

It is not uncommonly argued that part of the reason that obesity is seen to be a problem is because of a general predilection for society to moralise bodies (Greener et al., 2010, 1043). Discussions regarding intervening on obesity in this chapter are not intended to be taken with a moral stance. The discussion here, instead, pertains to types of evidence used when conducting research in the sports sciences, with obesity acting as a useful example case, given the current level of interest paid to interventions by the World Health Organisation and public policymakers. What this chapter is concerned with is the evidence that supports different practices which aim at intervening on being *overfat* where intervening on being overfat can improve health outcomes. Being 'overfat', needs to be intervened on, for instance, as it is a cause of, or contributes to, any of a large set of comorbidities including: insulin resistance, glucose intolerance, hypertension, higher levels of visceral fat, etc. (McArdle et al., 2008, 456).

The medical implications of obesity are common, and increasing per number of population. In a philosophical review of obesity studies, Federica Russo found that diseases that can be caused by the excess adipose tissue have drastically increased since the obesity epidemic began (Russo, 2012, 141). Additionally, in a 2016 press release, Cancer Research UK reported that by 2035 obesity and overweight[1] could be the cause of 440,000 cases of diseases such as coronary heart disease, type 2 diabetes, and cancer, per year (Cancer Research UK, 2016). It is noted, however, that it is possible to be considered to be overweight or 'overfat' and not exhibit any of these comorbidities associated with obese syndrome (McArdle et al., 2008, 456). Just as not all smokers will get lung cancer, being overfat simply puts one in a population where there is a high population level increase in comorbidities and mortality. Within the context of this chap-

[1] In the literature, one can be described as 'being' or 'having' overweight or 'overfat', which leads to odd sounding grammatical constructions. I have followed the standard usage of the literature.

ter, when the need to intervene on obesity is discussed, or people are referred to as obese or overweight, the intention is to focus on interventions for people suffering from obese syndrome and morbidity caused by this, not simply larger but healthy people, even if such terms are often used interchangeably.

7.2.2 A note on BMI

In day to day conversation, even with philosophers, one will often be challenged if they bring up the use of BMI as an indicator of obesity.[2] In the medical sphere, there is also some concern about the use of BMI (see for instance: Garn et al., 1988; Smalley et al., 1990; Yusuf et al., 2005). As such, many claim it should not be used. As this is the case, I will briefly defend the use of BMI. The thought of opponents of BMI is that people will likely be diagnosed as obese when they are simply heavy for their height in a way that does not risk their health. For instance, people worry that some may be categorised as obese without having excess adiposity, or whilst having excess adiposity but being otherwise healthy. There is some evidence to back up the claim that people may be erroneously diagnosed as obese using BMI as a measure. A Nature article (Rothman, 2008, S57), for instance, claims that around 8% of people will receive a false positive obesity diagnosis on the basis of BMI. BMI also has an error in estimating body fat levels in individuals of between 3 and 5% when corrected for age and gender. This compares poorly to other body fat testing methods such as water or air displacement, which give an error of closer to 1% (Deurenberg and Yap, 1999, 3, 8). As obesity is defined as an excess in body fat rather than body weight, these errors do seem troubling. However, despite it often being claimed in day to day conversation that BMI will incorrectly classify people as obese, it is far more problematic in the other direction. In one study of people from five different European countries, 41% of men and 32% of women

[2] Every time I have presented this chapter, someone has questioned the use of BMI.

get false negative obesity diagnosis using BMI as an indicator of body fat (Deurenberg et al., 2001, 977). This means that 41% of men in the study population who had a BMI indicating that they were not obese did in fact have levels of body fat sufficient to classify them as obese. Given that this study was completed in Europe, it may not be wholly indicative of the accuracy of BMI in other populations.

If these problems with using BMI as a classifier of obesity exist, why is it used? If you peruse government obesity policy guidelines, World Health Organisation press releases, or obesity papers published in medical journals, you will almost always see reference to BMI as the key measure used to indicate obesity. The first point to note is that when age and sex are taken into account alongside BMI, it becomes a much better indicator of body-fat on a population level (Gallagher et al., 1996, 238). It is also a better indicator of obesity when specific populations such as bodybuilders, pregnant women, or ethnic groups with high bone densities are accounted for (Deurenberg and Yap, 1999, 2). A study across five European nations also found that once age had been accounted for, BMI was a good population level indicator of obesity, although on an individual level, biases could be high (Deurenberg et al., 2001). These biases include, for instance, lack of, or abundance of muscle compared to the average, and extremes in height. Any of these can give a BMI that is not a good indicator of body-fat levels. In a summary of the use of BMI aimed at the general public by the Harvard School of Public Health, it is explained that the reason BMI is used is, in part, due to practical constraints. BMI is used as it is easy to measure, has a long-standing history of use, and is still a very good predictor of disease on a population level (Harvard School of Public Health, n.d.). We can see that the use of BMI as an indicator of obesity comes down to the practicality of measuring obesity on a population level, and determining the risk of obesity related diseases on a population level.

As a final note: often, the effects of obesity interventions are measured in change in weight in kilograms rather than as a change in participant BMI, although this is not universal. I have been unable to find the reasoning behind this in the literature, however, as weight is intended to change as a result of these interventions but sex and height are not, this may explain why change in weight is used as a measure.

7.2.3 Adherence

As will be seen as this chapter progresses, adherence plays a key role in the effectiveness of exercise interventions for obesity. How well a patient can frequently and optimally follow a treatment plan is called the level of *adherence* (World Health Organization, 2003, XIV). Many factors can complicate adherence, for instance: the type of therapy, the system administering the therapy, and economic and social factors (World Health Organization, 2003, XIV). Adherence is a key to moving from clinical efficacy to real-world effectiveness, and improving adherence with low-cost methods is consistently found to significantly reduce overall spending on treatments (World Health Organization, 2003, 22-23). In general, treatments that need lifestyle changes have particularly low levels of adherence (World Health Organization, 2003, 35). A prescribed treatment with established efficacy will of course be more effective if a patient actually engages in it than if they do not. Unfortunately, in a recent Cochrane systematic review of interventions to improve adherence, it was found that there were no common characteristics shared by the few interventions that did improve adherence (Nieuwlaat et al., 2014, 2-3). It should also be noted that clinical trials can have remarkably high levels of adherence when compared to real-world treatments; this is put down to the attention received by patients and the selection of specific patients (Osterberg and Blaschke, 2005, 487).

7.3 Obesity interventions

Given the severity and expense of obesity, both in number of cases and potential health outcomes, many governments seek to intervene on the problem with public health initiatives. Public health initiatives are potentially useful in the case of obesity given that it is difficult to provide individualised interventions for the large number of people who suffer from obesity. Public health initiatives, such as those outlined by the NHS Long Term Plan, are intended to be helpful, cost-effective measures, that can also reduce pressure on public services (Alderwick and Dixon, 2019). Given that this chapter discusses evidence in the sports sciences, I will constrain discussion to exercise interventions for obesity. I will also constrain my discussion to how exercise interventions are intended to intervene on excess adiposity and obesity. We may reasonably suggest that even in cases where exercise does not reduce adiposity, it may improve more broad health outcomes, but this is beyond the scope of this chapter. I will however make one small point regarding this. As one of the key things that explains the difference in effectiveness between supervised and unsupervised interventions is adherence, it may well be that unsupervised exercise interventions are also less effective at intervening on broader health outcomes than supervised interventions, just as they are for obesity specifically.

An increased caloric intake is often blamed for the obesity epidemic. However, as can be seen in the example of the United States, the increase in per capita energy intake is not alone sufficient to account for the rise in the average body mass (McArdle et al., 2008, 471). A decrease in energy expenditure must also be present to account for this increase in body mass. As such, many interventions for obesity often employ, at least in part, activity-increasing or exercise-focused components. Unfortunately, despite there being a strong inverse correlation between physical activity levels and weight (Sport England, 2017, 27), exercise-focused interventions that seek to reduce obesity often have low or insignificant effect sizes, as shall be explained.

In 2019 in the UK, the NHS recommends taking at least 30 minutes of exercise a day and up to 90 minutes per day for the already obese for 5 days in the week in order to tackle obesity (National Health Service, 2019). However, an overview of a large number of studies found that outside of supervised settings, an exercise prescription of 30 minutes of exercise a day for 5 days a week was insufficient to reduce body weight in participants (Donnelly et al., 2009, 461). In the UK particularly, these initiatives tend to focus on school-aged children as, once established, obesity will tend to track throughout the lifecourse of an individual. In the UK, many such initiatives have been trialled such as MEND, and HeLP (Lloyd et al., 2018; Sacher et al., 2010). Both of these are interventions for childhood obesity which include exercise as a key component. These types of interventions aim to teach young people about exercise and nutrition in a way that helps them bring those practices into the rest of their lives. MEND, for instance, includes twice weekly physical activity and education sessions attended by both parents and children. At the completion of the MEND programme, families are given 12-week free swimming vouchers. The aim of this type of scheme is not just to intervene on obesity and overweight throughout the course of the programme, but also to give children and families the tools they need to manage their weight throughout the course of their lives (Sacher et al., 2010, S63). However, systematic reviews of school-aged obesity interventions have found studies are largely contradictory or inconclusive with regard to obesity reduction (Summerbell et al., 2003; Waters et al., 2011).

In a review that makes up the ACSM position stand on obesity interventions by Donnelly et al., it was found that unsupervised, exercise only interventions rarely reported an average weight decrease of more than 3% in participants (2009). Similar findings are also supported by other, more recent retrospectives (Hagobian and Evero, 2013; Petridou et al., 2019). Whilst this may be beneficial for health, it is insufficient to dramatically improve health outcomes for those who need substantial weight loss to achieve healthier levels of body fat (Donnelly et al., 2009, 461). This seems to be the case regardless of how the prescribed exercise is structured. For instance, results were similar when many shorter medium intensity bouts of exercise are used, and when fewer but longer bouts of medium intensity exercise are employed (Donnelly et al., 2009, 461). Garrow and Summerbell (1995), and Wing (1999) also found that of the few physical activity interventions that did manage to reduce participant weight, typically only a weight loss of 2 to 3 kg was seen in each participant.[3] Interestingly, in comparison to unsupervised exercise interventions, when employed under supervision, exercise interventions can be effective at reducing participant weight (see for example: Craighead and Blum, 1989; Donnelly et al., 2003; Ross et al., 1995). Just as in the last chapter, it will become clear that we are justified in inferring that this relationship is causal from the evidence available. This is because we have evidence of correlation from a number of RCTs, and also established mechanisms, which are discussed in the next section.

One could counter that, perhaps, trials that found unsupervised exercise interventions to be insufficient to dramatically improve health outcomes may all have been poorly conducted and that they wrongly showed no effectiveness. However, given the number and breadth of the analyses discussed above, we are reasonably justified in ruling out this idea. Particularly, we are justified in ruling out this idea when the following explanation is provided for why effectiveness appeared to be low in these trials. Reviews and retrospectives do mention that adherence, and reported adherence to the interventions in these trials may have been low, so the trials do not provide evidence for the effects of confirmed amounts

[3] To put these weight loss intervention findings into perspective, I will use myself as an example. At the time of writing, I am 6'2 and around 108 kg. My 30.5 BMI places me firmly in the obese category. To reach what is considered to be a healthy weight, I would need to lose around 20 kg. This is far more than it is expected that I will be able to lose with exercise interventions, according to these systematic reviews and meta-analyses.

of exercise (Petridou et al., 2019, 165, Donnelly et al., 2009). Discussion of this is similar to the discussion of the 11+ in subsection 6.3.3. We can measure the effectiveness of an intervention if it is adhered to, or we can measure the effects of an intervention if it is prescribed. What this means is that in trials with low adherence, *prescription* of the intervention did not lead to sufficient weight loss to improve health outcomes. This means that those trials do still provide evidence that being prescribed an exercise intervention is likely insufficient for the management of obesity.

In addition to the evidence that supports the lack of effectiveness of unsupervised exercise interventions to reduce weight in obese participants, is the generally accepted claim that participants who do experience weight loss will often be unable to maintain that weight loss (Donnelly et al., 2009, 462). A systematic review of weight regain trials found that even if exercise was reported as continued after intervention, participants would regain 0.28 kg of body weight per month (although this is lower than the 0.33 kg per month gained back by non-exercise participants) (Fogelholm and Kukkonen-Harjula, 2000). The review found that the amount of energy expenditure through exercise required to maintain new weight in participants was between 1500 and 2000 kilocalories per week, which far exceeded what was achieved (Fogelholm and Kukkonen-Harjula, 2000). From these findings together, we can infer that after a weight loss of 3 kg in one of the few cases where an exercise intervention was able to produce some weight loss, participants will likely reach their starting weight again within the year. What we can take from this is that when actually put to the test in non-supervised settings, typical exercise recommendations are inadequate to reduce participant weight substantially and to prevent weight regain after the intervention is ended.

The enormity of obesity as a public health problem, and the lack of effectiveness of many interventions for obesity makes it a prime candidate for a discussion of evidence and understanding. This disconnect between proposed effectiveness and actual effectiveness becomes increasingly concerning when we consider that the amount of exercise recommended by public health interventions, such as those recommended by the NHS, does not appear to be effective for patient weight management, at least not when prescribed as it is. Surely we want to be able to explain why these interventions do not work, particularly given that they are endorsed by major public organisations. That this is the case will become more clear in the next section, where I explain that physiological mechanisms indicate that exercise interventions for obesity *should* work. However, I argue that research into the topic that adopts a purely physical framework cannot explain fully why they do not work.

7.3.1 The mechanisms that explain why obesity interventions ***should*** work

The mechanisms explaining how exercise *can* be used to intervene on weight are established. We know this because they are taught in many undergraduate and graduate level textbooks (such as: Braun and Miller 2008; McArdle et al., 2008; National Strength & Conditioning Association, 2015). We also know that they are established because research into obesity interventions uses them as evidence when investigating and developing interventions. I will use this section to introduce the mechanisms that explain how excess consumption can cause obesity, and how exercise can be used to intervene on this.

In the most basic terms, energy enters the body via diet and is expended via a mixture of physical activity and the energy used to sustain the body's basic functions at rest. Changes in the amount of energy stored in the body are mainly marked by changes in the level of fat stored in the body. This means that, generally, expending more energy than one intakes will cause a loss in weight, and using less will cause an increase. Food that enters the body that is not used for energy straight away is stored chemically ready to be used for mechanical work when needed. When required for work, these bonds are broken down into adenosine triphosphate (ATP), the chemical muscles utilise in order to contract. The fat, carbohydrates, and protein found in food can all be metabolised for use as energy in the human body. For a more detailed description of this mechanism, see Braun and Miller, 2008 pages 2 and 3.

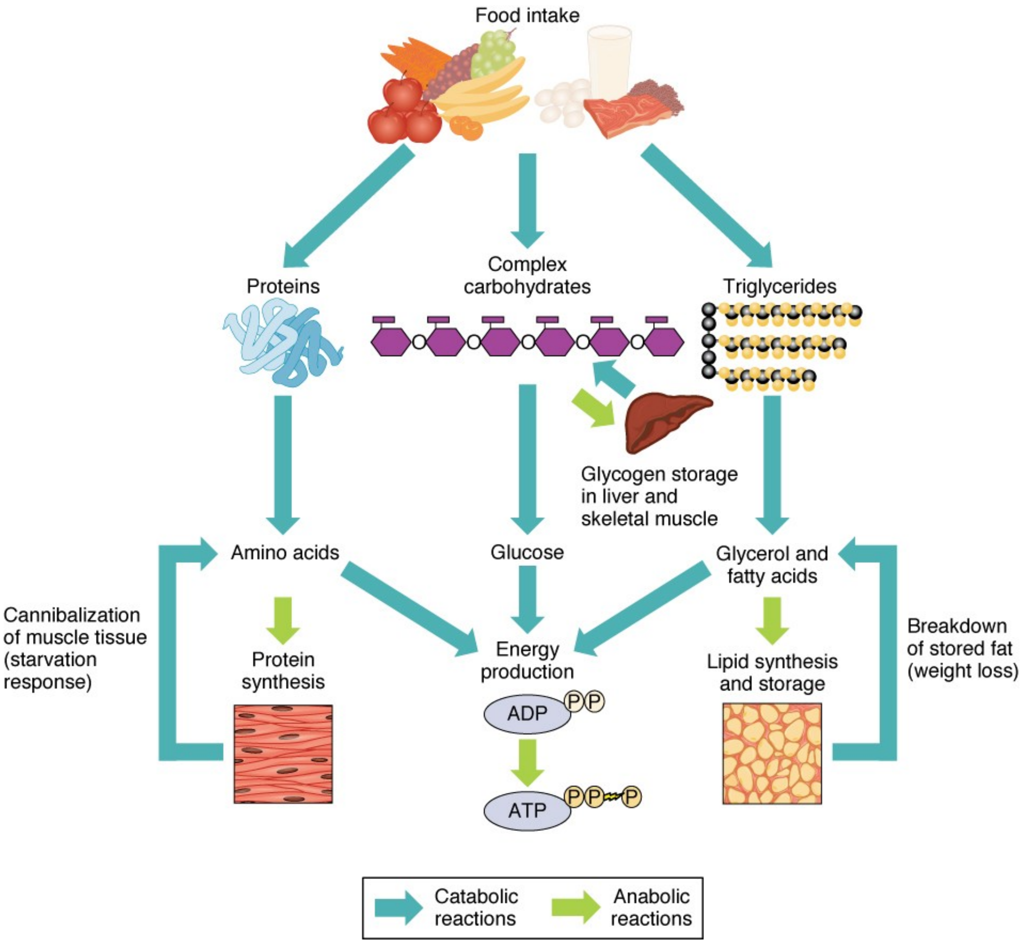


Figure 7.1: A diagram showing the mechanism by which food ingested is stored as muscle, fat, and can be used to produce ATP. Reproduced from OpenStax, n.d.

This mechanism is shown in greater detail in Figure 7.1. From the diagram it can be seen that food taken in is broken down, and is then eventually stored as fats or muscle, or is used to create ATP which is then used as energy. From the diagram it can also be seen that through the breakdown of muscle and fat tissue, amino acids, glycerol, and fatty acids can also be produced in order to create ATP for use as fuel by the body.

As has been explained previously, in order to maintain body weight, energy input must be balanced with energy output (McArdle et al., 2008). A positive energy imbalance can be contributed to by inactivity and overconsumption of energy. Obesity interventions seek to work by managing or tipping the energy balance in the other direction, meaning more energy is used than is consumed, causing weight loss or management. The mechanisms that explain how inactivity and overconsumption of energy themselves arise are not dealt with in this section, as they are multifactorial and complex, but include such complex systems as satiety hormones not suppressing appetite adequately to avoid excess energy intake, and time management skills leading to a perceived lack of time to exercise. Insidiously, one does not have to consume many calories above the amount that balances energy input and output in order to gain weight. The calorie equivalent of just one of each of around 50 grams of roasted peanuts, or three slices of bread, or slightly more than one standard chocolate bar, consumed every day above the amount of food needed to sustain body weight will, theoretically, lead to an increase in mass of 7.3 kg in just one year (McArdle et al., 2008, 451). This would amount to roughly 300 kilocalories or 1255 kilojoules of energy consumed in excess of what is needed to maintain body weight per day.

7.3.2 Mechanisms explaining why larger individuals use more energy to perform the same activities

So far, it has been explained how the balance of energy consumed and expended can be reflected in body weight. There are some interesting physical mechanisms that can affect this balance that differ between obese and non-obese individuals. In obese individuals, the output of the heart and the respiratory system, which brings oxygen into the body, often needs to be higher to perform the same activity than it does in smaller individuals. This marks a higher amount of energy used to perform the same action. This is because pulmonary function is impaired as a heavier body mass must be moved. This is, in part, due to the greater load placed on the body requiring a greater recruitment of muscle fibres to produce the same movement. There are also other mechanical inefficiencies that can be seen in obese individuals that contribute to a larger energy expenditure in obese individuals than non-obese individuals during the same movements. This includes things such as the increased friction between limbs which requires more energy to overcome (for a more detailed understanding of these mechanisms see: Sothern, 2001 page 998).

This difference in energy expenditure can be seen in the results of a treadmill exercise test. In the test, 'normal-weight', overweight, obese, and severely obese children walked at the same pace until they decide to give up. In this study, whilst moving at the same speed, 'normal-weight' participants exercised at 38% of their maximum oxygen uptake, overweight participants at 54%, obese children at 72%, and severely obese subjects exercise at 84% of their maximum oxygen uptake capacity (Sothern et al., 1997). This need to intake oxygen at a greater percentage of maximal capacity as subjects reach a higher body mass indicates a greater metabolic response (Sothern, 2001, 999). It illustrates clearly the point made earlier, that larger individuals will, in general, require a greater amount of energy to perform the same action as a smaller individual. These mechanisms may, in part, contribute to the finding of a meta-analysis that larger individuals lose adipose tissue at a greater rate than less fat individuals (Ballor and Keesey, 1991).

7.3.3 How increased activity can cause weight loss

During certain kinds of activity, particularly low-intensity exercise, or prolonged work performed at a sub-maximal effort, the oxidative system in the human body uses energy stores to provide ATP, the chemical used for work by the muscles (National Strength & Conditioning Association, 2015, 52). This is done by using fat and carbohydrate in the body as the substrate to create ATP (National Strength & Conditioning Association, 2015, 52). This explains how energy that is stored in the body is used to perform mechanical work. Energy expenditure as a result of physical activity generally accounts for around 10% of total energy expenditure. The goal of exercise based interventions is to increase this energy expenditure to create an energy deficit in participants. This means that the body will use the energy that is stored in the body, ultimately, leading to weight loss (Jakicic and Otto, 2006, S58).

As mentioned earlier in this chapter, it would theoretically take eating only three slices of bread per day above body weight maintenance intake to gain over 7 kg in a year. This seems like a large gain in weight for a small excess in energy consumption. However, theoretically, the increased activity required of a person, assuming balanced energy consumption, is also surprisingly low. Walking 7 miles a week above normally performed activity for a year, the equivalent of one mile a day every day for a year, should use enough energy to amount to a weight loss of 4.5 kg in that year, assuming that the exercise does not lead to an increased energy intake (McArdle et al., 2008, 474). Employing exercise to increase the energy use of the body is, then, a potential method of shifting the energy balance in favour of reducing or maintaining body weight.

7.4 Supervised interventions

In this section, I will introduce an uncommon type of exercise intervention for obesity, one in which the exercise is supervised. The combination of strong mechanistic evidence for how exercise should help intervene on participant weight, and evidence from trials investigating supervised exercise interventions show that, in some cases, if the exercise intervention is supervised, weight management is possible.

In order to compare supervised interventions with the standard intervention, consider the exercise advice given by the National Health Service of the UK on their website (National Health Service, 2019) as an example of the typical unsupervised exercise intervention for obesity. The recommendation given is that a minimum of 150 minutes of moderate-intensity activity be performed a week, amounting to 30 minutes of activity 5 times a week. This can include activity such as brisk walking, cycling, swimming or dancing. For those who are already obese or who need to prevent the regain of weight, the recommendation for exercise raises to between 45 and 90 minutes of activity per day for five days of every week, leading to between 225 and 450 minutes of exercise per week. These recommendations are typical of those used in exercise-based obesity interventions and as advice given to the public by health institutions such as by the American College of Sports Medicine (ACSM) (Donnelly et al., 2009).

Allow me to now give an example of a typical supervised intervention, for comparison. In a 16-month study of the effects of supervised exercise on weight loss in obese and overweight participants, Donnelly et al. (2003) found that 225 minutes per week of moderate intensity *supervised* exercise per week for the course of the study was sufficient to reduce male participant weight significantly (on average 5.2 kg), and to stop weight gain in female participants. This is particularly notable when compared to the nearly 3 kg average weight gain in female controls, who were asked to maintain normal activity levels for the 16 months of the trial (2003, 1348-1349).[4] Whilst this is clearly more effective than unsupervised exercise interventions, it is also notable compared to the most successful unsupervised obesity interventions. When weight loss is averaged over a year, the intervention carried out in this trial is over a third more effective than the most beneficial unsupervised interventions. Interestingly, this amount of exercise is on the lower end of that recommended for already obese individuals by the NHS. From this we can infer that this trial did not become effective only because massive amounts of exercise were prescribed. Another concern about what may influence the effectiveness of an exercise intervention is how it may interfere with diet. However, in this study, energy intake was measured and compared between control and test groups who were all encouraged to eat ad libitum, or at their pleasure. No significant difference in energy intake was found between or within groups (Donnelly et al., 2003, 1348). So, supervision did not affect diet. This illustrates that exercise participants did not shift their energy balance in favour of weight loss by dieting and, by that means, cause the intervention to be successful instead of through the exercise. The negative energy balance appears to be as a result of exercise. Thus, we can see that the intervention is not different to other exercise interventions which do not prescribe an altered diet, except in that it was supervised.

The supervision in this trial was intended to ensure that the trial measured the effects of a validated amount of exercise and involved observation, and not encouragement. It involved all exercise being conducted with direct, in person, supervision. Heart rate monitors were provided in order to allow the research assistants to ensure the intensity of exercise was appropriate. Energy expenditure was also periodically measured by exercise supervisors. (Donnelly et al., 2003, 1344)

[4] It is suggested that as the level of difficulty of exercise was gradually ramped up throughout the study, and that it took women longer to reach the desired activity levels (5 months longer than men), that this may explain why they did not lose weight.

7.5 The problem with a purely physical framework

So far, it has been explained why interventions that tackle obesity are important, and how physical mechanisms explain why they *should* work. It has also been shown that, in many instances, unsupervised interventions do not work, whilst supervised interventions do. I will now argue that we cannot provide a good explanation for the observed difference in effectiveness between supervised and unsupervised obesity interventions without considering details of social and psychological mechanisms relevant to the interventions, alongside physical mechanisms.

When we take into account the fact that a) unsupervised exercise interventions for obesity very rarely reach effectiveness, and b), that the mechanisms that explain why they should be effective are considered to be established, it is important to find out what explains this in order to be able to devise effective interventions that are also practically scaleable given the size of the obesity epidemic. We are reasonably justified in inferring that a good explanation for the observed difference in effect sizes is that the established physical mechanisms are being masked, or even not instantiated, because of other mechanisms that are not being considered. By comparing the difference in effectiveness between supervised and unsupervised interventions, we may also infer that supervision of those interventions is a causal factor. However, I contend that simply knowing that supervision is a causal factor is not enough. When the goal is results driven, greater benefits can be seen if we explore further to determine how supervision impacts intervention effectiveness. These include benefits like deepening our theory of the causal relationship, and using this to aid practical intervention development (this is discussed in greater depth in chapter 8). Physical mechanisms alone cannot give the full picture in this case.

In order to see that it is not only physical mechanisms that mask, or prevent, the effect of this type of intervention, we must first dispel some common mechanisms proposed to explain the lack of effectiveness of exercise as an obesity intervention. A widely and popularly used physical mechanism given to explain why exercise interventions are not effective is that obese people may simply have a *slower metabolism* and are, thus, able to eat less but gain more weight than other smaller individuals. The grounding for this proposed mechanism being that the amount of energy used simply by being alive, to breathe, and create body heat, known as the Basal Energy Expenditure (BEE) of these individuals, is far lower than in smaller individuals. This is often blamed for high levels of adipose tissue (Flatt, 2007, 2546). Whilst a compelling story of a mechanism, the evidence in favour of this is lacking. In fact, there is no correlation (which would be needed to establish a causal relationship) between body fat and deviations from predicted BEE (Flatt, 2007, 2548).[5] Indeed, the mechanisms given earlier, detailing how obese individuals also need to use a greater amount of energy to perform the same movements as smaller individuals, do work towards explaining how larger individuals may, when physically active (an amount of energy not taken into account in BEE calculations) use more energy. Additionally, some colloquially suggest that their BEE is low because it has declined with age. Contrary to this, recent high-quality research suggests that BEE may be relatively invariant between 20 and 60 years of age (Pontzer et al., 2021).

As is explained by Russo (2015, 845), it is also held by some that as exercising can increase appetite, exercising may, in fact, cause weight gain as the amount of food that is consumed may be higher. Increased hunger would then mask the mechanism by which exercise is intended to reduce weight. Whilst this may occur in some instances, it was seen in Donnelly et al. (Donnelly et al., 2003), this is not always the case and, thus, does not give us reason to think that every case where effectiveness differs

[5] The R^2, a measure of how well data fits a statistical regression model, being less than 0.0002 (Flatt, 2007, 2548). This indicates that it is incredibly unlikely that there is a population level relationship between body fat and BEE.

can be fully explained by differences in diet, rather than supervision. This trial was a supervised exercise intervention where control and trial participants were allowed to eat as they wished. At the end of the trial, those participants who performed supervised exercise better controlled their weight than those who did not. It was noted by Donnelley et al. that all trial participants consumed a similar amount of energy. This can be taken as evidence that not all exercise interventions will cause participants to eat more.

Of course, dismissing two proposed physical mechanisms that could be suggested to account for observed differences in effect size, and then assuming that no other physical mechanisms mask, or prevent, the lack of effectiveness is a poor assumption and looks like a hasty generalisation. The mechanisms relevant to obesity are highly complex and likely include more than these two. These were, instead, intended to be an example of commonly proposed mechanisms that may be suggested as being able to explain the difference, neither of which have much traction. It will become clearer however, once details of the social and psychological mechanisms related to supervision are provided, that physical mechanisms alone cannot explain the difference in effectiveness between these types of intervention. This is because key mechanisms preventing the effect of the proposed physical one does, in fact, have determinants that can only be adequately explained once details of social or psychological mechanisms are taken into consideration. This mechanism is one which can cause adherence to, and motivation for, interventions to differ. This will be explained further in section 7.4.

In sum, all the evidence of physiological mechanisms and the fact that supervised exercise interventions are associated with weight loss could, and does, lead us to infer that exercise *should* be appropriate when used as an intervention for obesity. We can tell that it does lead people to infer that this type of intervention should be effective because many public health organisations, such as the NHS, promote them. This does not match up with the evidence from trials, however. Evidence from trials seems to show that exercise often does not make a difference to participant weight, unless it is supervised. The reviews previously mentioned all find little success for participants aiming to manage weight using exercise in the most common type of intervention. Also, a brief foray into some physical mechanisms that may be invoked to explain the lack of effectiveness of these interventions turns us up short, neither explaining why unsupervised interventions do not work, nor by explaining the difference in effectiveness between supervised and unsupervised interventions. As such, in the next sections, I will provide details of some social and psychological mechanisms that explain why they are more effective than unsupervised ones. This motivates the idea that providing details of these mechanisms is important and useful in the sports sciences.

7.6 Supervised and unsupervised: the difference

As, in the example I give (Donnelly et al., 2003), the amount of exercise prescribed in the supervised intervention was typical of an obesity intervention, the diet was ad libitum, and the minimum intensity of exercise was controlled for and reached the level prescribed by typical obesity interventions, it can easily be identified that the key difference in the intervention being tested by this trial and a typical exercise intervention was the supervision. We can also tell that the supervision is considered a key causal difference between the two types of studies because, in the literature, ‘supervised’ exercise interventions are mentioned independently of unsupervised interventions (see for example Donnelly et al., 2009; Ross et al., 1995). What this means is that supervision of exercise is not standard for obesity interventions, and is unusual enough that if it appears in an intervention it will likely be mentioned in the title of the paper discussing it. Other than the factor of supervision, this intervention is typical of exercise-based obesity interventions. However, an outcome of this trial that separates it from the typical unsupervised intervention, along with significant weight loss and management results, is a high level of adherence, particularly when compared to unsupervised interventions. The adherence related findings of Donnelly et al. (2003) are themselves typical of obesity interventions that include supervised exercise (see for example: Chang et al., 2008; Ross et al., 1995). Whilst this trial, and others like it, do highlight that supervision can increase adherence and can increase success, I will next argue that without the examination of social and psychological mechanisms we cannot provide a particularly good explanation of why this is the case.

As was argued in section 7.5, if research is conducted within a purely physical framework, even if evidence of physical mechanisms and evidence of correlation are used, it is impossible to explain fully the gap between effectiveness of supervised and unsupervised exercise interventions. It was suggested that there must be a non-physical mechanism masking, or even preventing the instantiating of, the proposed physical mechanisms that explain why exercise should be able to be used as a tool to manage participant weight but does not. Evidence of correlation allows us to see the net effect of a mechanism (Clarke et al., 2014, 351). Using this knowledge, we can compare the results of supervised and unsupervised obesity interventions and see that the interventions have a different net effect. By doing this, and comparing differences and similarities between the interventions themselves, as I have done, we can see that at least one important difference *is* the supervision. Using this reasoning, one may think that as we have identified supervision as a causal factor, it can therefore be employed in all instances with this intervention. Following this, exercise would then become a sufficient tool to manage obesity. However, as I discuss here, and in subsection 7.7.1, it is not practical to supervise the exercise of everyone who requires obesity intervention. Simply seeing that the presence of supervision in an exercise intervention for obesity can have an effect on trial outcomes is also not a particularly good explanation for why this difference in outcomes occurs because it is not an explanation that we can do much with.

It may be contested here that if the evidence from trials were sufficient to a) uncover a mechanism that masked or prevented weight management occurring, and b) provided grounding for the fact that there is a mechanism between supervision and effectiveness, it should be enough to simply always employ this type of intervention under supervision. For instance, the ACSM states in their position stand on exercise as an obesity intervention that exercise is a more effective treatment when supervised (Donnelly et al., 2009, 466-467). In this same position stand the ACSM provides no evidence of, or reasoning by mechanism, as to why this may be the case. Although, perhaps, this is because adherence is not often considered in mechanistic terms currently. In fact, it may look like reasoning based on the physical mechanism alone led to ineffective interventions, and that comparing results from association studies can point to a reason why these interventions are often ineffective. This may even be used as an argument against the use of evidence of mechanism in order to establish causal claims. Critics of mechanistic reasoning, for instance, may think that as reasoning mechanistically here led to an ineffective treatment, this adds this type of intervention to the list of treatments proposed on mechanistic grounds which showed little effectiveness, this list being an argument against the use of evidence of mechanism when establishing causal claims. I discuss this criticism in more depth in the next chapter.

Even disregarding what types of evidence are required in order to establish a causal claim, not being able to explain differences in effectiveness between intervention types becomes a problem when we consider the enormity of the problem posed by obesity. If obesity rates reach the predicted levels of 60% in men by 2050 in the UK, how will it be possible to supervise the exercise of over half of a nation in order that the exercise becomes an effective intervention? As a practical and results driven field, sports science seeks to find actionable solutions to problems. When we cannot provide good explanations for why differences in effect sizes occur, we lack understanding that may benefit us by helping us to improve obesity interventions. When we explore these interventions in a framework that excludes or ignores the importance of social and psychological mechanisms, this is where we may find ourselves, as evidenced by the case of obesity interventions. There may be, for instance, something that is caused by supervision which increases the effectiveness of these interventions that is easier and more practical to roll out in a wide scale intervention than it is to supervise 60% of the population. There may also be something that instantiates a similar adherence causing mechanism to supervision that it is easier to roll out. For instance, as I discuss in subsection 7.7.1, group meetings may improve adherence in a similar way to supervision, and be far easier to include in interventions. This applies, of course, outside of exercise interventions for obesity and to research in the sports sciences as a whole. If we cannot fully explain how an intervention, treatment, or type of training has its effect, there may be undiscovered or unconsidered social or psychological mechanisms impairing the net effect that, if considered, may be taken into account and improved upon. This potential ability of providing of details of psychological or social mechanisms to improve and generate interventions and solutions to problems is what really motivates taking social and psychological mechanisms seriously in sports sciences. Part of this, of course, involves taking evidence from mechanistic studies more seriously than current evidence hierarchies suggest that we ought to.

In the obesity literature, both sport and psychological, I have been unable to find fine-grained accounts of the mechanisms relevant to supervision, and how it plays a role in the effectiveness of exercise interventions for obesity. This is the case despite numerous accounts stating the importance of supervision. However, key causal factors relevant to supervision are related in some instances. For example, in their overview of research on exercise in the management of obesity, Petridou et al. (2019) discuss the importance of supervision and its relation to adherence. As I explained in the introduction to Part III, discussion at this level is in line with what I am considering improving our providing of details of relevant mechanisms to be. In the next section, I will examine how supervision may play a role in the effectiveness of exercise interventions for obesity.

7.7 Adherence and effort

Now we must turn to examining how supervising exercise interventions causes the amount of energy used to be higher than that seen with unsupervised interventions, leading to more effective weight loss. This involves leaving physiology. Just as was explained in the case of the FIFA 11+, when we look at the effectiveness of an intervention or public policy, we can examine it in two ways. We can determine if an intervention would be effective if it were followed, and we can determine if an intervention will be followed given that it is effective. However, as we see from the example of these obesity interventions, an intervention that may be effective in theory but is not in practice, is a poor intervention in a field where the research is meant to be applied. As may be easily surmised, one of the reasons that supervision improved the effectiveness of exercise interventions for obesity is that it meant that people who had been prescribed the intervention actually followed it, allowing the relevant physical mechanisms to be instantiated, allowing them to cause a negative energy balance. Although rarely mentioned in the literature, this point is conjectured briefly in a 1989 study on the effects of supervision on exercise for obesity (Craighead and Blum, 1989, 49).

This is poor news for exercise interventions. As a type of therapy, it can be readily assumed to be more difficult to ensure adherence to exercise interventions than many medical treatments. It involves much more will power and less passivity to exercise five days a week than it does to take pills daily. Unsurprisingly, reviews find that unsupervised interventions for obesity have a very low adherence rate (Burgess et al., 2017, 123). Specific to exercise interventions for obesity, barriers to adherence include a wide range of factors such as: motivation, lack of time, enjoyment (Burgess et al., 2017), and self-image (Dalle et al., 2010, 6, Billings et al., 2010, 76). Individuals also often face distraction when exercising, or are often concerned with neither health outcomes nor competitive outcomes as a result of exercise (Sothern, 2001, 996-7), meaning that meaningful engagement can be limited. Meaningful engagement can be taken to mean whether someone actually engages with the exercise, or simply goes through the motions. For instance, meaningful engagement on an exercise bike likely involves an elevated heart rate, and sweating, where one could engage non-meaningfully and simply spin their legs slowly in a low gear, thus not contributing adequately to an energy imbalance. Although, lack of meaningful engagement when characterised like this could be, in some instances, claimed to be a lack of adherence. In addition to this, the exercise recommendations made in interventions and public policy largely promote 'moderate intensity' exercise, but this level of exercise and the specific exercise modalities recommended may not be engaging or even safe (Sothern, 2001, 997), contributing to reduced adherence and engagement. A myriad of other social and psychological factors can contribute to the unwillingness of people to participate in exercise, even when not couched as an obesity intervention. These can include peer pressure, social status, and embarrassment (Billings et al., 2010). Of course, if adherence to exercise interventions is low, it will be difficult for patients to create a negative energy balance and to manage weight. The fact that adherence was particularly high in supervised interventions when compared to unsupervised interventions helps to explain why the quality of weight management is better in these studies. It also helps to identify adherence as a key part of this explanation.

As is indicated by a brief foray into the adherence literature, and by listing the types of barriers to adherence that participants in exercise interventions for obesity face, it becomes increasingly clear that physical mechanisms alone cannot explain what factors affect adherence, or how to move from efficacy to effectiveness. Many of these barriers are psychological or social barriers without direct physiological determinants. The understanding of physical mechanisms alone, also, therefore, cannot be used to improve adherence and thus effectiveness. In order to understand the link between supervision and adherence, we must once again take a step that requires the examination of details of social and psychological mechanisms. We have correlational evidence that supervised exercise leads to better adherence, and good evidence that there is a mechanism for this (experimenters ensuring that participants exercise, and at the correct intensity), as indicated by Donnelly et al. (2003). Because of this, we can reasonably assert that supervision improved adherence. We can, then, infer that adherence improves the effectiveness in this type of intervention by increasing the amount of exercise done, especially as doing the prescribed exercise *is* adhering to the intervention. Whilst these factors play a role in the physiological mechanism of weight loss, physical mechanisms alone cannot explain how supervision leads to adherence, meaning that social and psychological mechanisms are required to fill this explanatory gap.

Despite the fact that I have highlighted adherence as a causal factor in the effectiveness of exercise interventions, it is also unlikely to be the case that supervision only intervenes on the mechanism by which exercise interventions work in that one way. For a true expert team or individual, it may also be possible to further examine details of social and psychological mechanisms relating to supervision and energy used for exercise. Someone trained in sports psychology may, for instance, be able to discover whether supervision may improve not just adherence to exercise, but the quality of that exercise. For instance, there is a reasonably large amount of evidence in sports psychology which suggests that being watched whilst exercising will increase participant performance in sport by raising effort, even without the athlete noticing the increased effort (see for example: Baker et al., 2011; Lamarche et al., 2011; Sheridan et al., 2019). If examined by experts, this factor could perhaps be shown to help further explain why supervised exercise interventions are more effective at weight management than unsupervised interventions. Even if interventions are adhered to strictly, participants will likely exert themselves more when observed during prescribed exercise. This greater level of exertion will contribute to the negative energy balance. If it is a causal factor, how this effect occurs in obesity interventions cannot be wholly explained with purely physiological mechanisms.

7.7.1 Benefits to providing details of social and psychological mechanisms: practical applications

If, when conducting research, we can identify social and psychological causes and explain how they affect outcomes as I have done, this may be beneficial to intervention or policy design. I will now illustrate potential practical applications of providing of details of relevant social and psychological mechanisms by introducing some methods that have been employed in real, although not exercise related, obesity interventions that utilise some forms of supervision. The idea here being that, once relevant causal factors are understood, we can use this understanding to help develop new and improved interventions for testing.

In the case of obesity, there is a good amount of evidence to suggest that supervision improves the effect of exercise based weight loss interventions. However, it is unlikely to be viable to observe the exercise of every one of the large number of obese people who may be prescribed exercise based interventions or interventions that include an exercise component. Investigating this mechanism and seeing how supervision improves the effect of these interventions can help us, here. I have postulated some causal factors by which supervision may affect energy used in exercise: by improving adherence to exercise and increasing effort in exercise. Effort in exercise and adherence to exercise are not necessarily only caused by supervision, obviously. However, this still gives us real options for intervention development. Those developing exercise interventions for obesity may well benefit from looking at non-supervision methods of improving adherence and effort in exercise, alongside the relevant physical mechanisms, in order to improve real-world efficacy.

Once identified as a causal factor in weight management through exercise, adherence can be addressed in order to develop, improve, or adapt interventions. A potential solution regarding the problem of adherence could be weekly group meetings. In this case, people involved in exercise based interventions could be accountable to a group that they also know understands their outlook. Adopting perhaps a similar method to other types of group therapy designed to offer support, accountability, and education. A similar approach is often taken in other types of obesity interventions, such as diet alteration. Evidence from a number of studies highlights that group therapy is often more effective than individual therapy or interventions where there are no meetings (see for example: Befort et al., 2010; Donnelly et al., 2007; Orth et al., 2008; Perri et al., 2001; Renjilian et al., 2001). A meta-analysis of group cohesion as an indicator of individual sports performance also indicates that being part of a group can indeed improve performance (C. R. Evans and Dion, 2012). A good example of this type of intervention is one in which African American women attended weight loss events at a local church (Sbrocco et al., 2005). This provided social support, ensured commitment, and increased participation over other methods (Sbrocco et al., 2005, 248).

Understanding the importance of effort, and adherence as key causal factors in the effectiveness of obesity interventions may also be harnessed to improve the outcomes of these interventions in other ways. A good amount of research, particularly in older people, suggests that introducing community aspects into exercise can greatly improve adherence (see for example: Beauchamp et al., 2018; Farrance et al., 2016; Osuka et al., 2017). The progress of consumer grade technology, available for the home, has brought supervised exercise classes and the community aspects of these classes into our houses. Whilst they probably cannot yet be considered to be affordable, the cheapest complete system of these costing around £1,700, Zwift, Peloton, and Hydrow are recent technologies that do this. They allow users to run, cycle, and row in their own homes, but against competitors worldwide, and often with either live or recorded feedback from instructors. Peloton, for instance, is an at home exercise bike that allows access to live and on-demand exercise bike classes. All types of class allow you to compare your output (for instance speed and distance) with other people taking the class, and live classes even allow real-time feedback from instructors. Peloton is a new technology, but some very early psychological research (Richardson, 2020, 16) suggests that, with the average user in mid 2020 completing around 4 workouts per week, it supports 'initiating and sustaining engagement as a pathway to positive behaviour change'. Although, it should probably be noted that the type of person to buy at home exercise equipment may confound the results of adherence to Peloton exercise, as they are likely people who already want to engage in exercise, which may not be the case for some people who are prescribed exercise interventions for obesity. Regardless, these types of technologies allow some level of external supervision, via live instructor feedback, and the option to share records and classes with friends as motivation to adhere to, and put effort into, exercise interventions. The at home aspect may also be useful to improve adherence in those who do not feel they have time to visit a gym, or those who feel embarrassed exercising in public. Given the cost of obesity to the economy and the NHS, the costs of developing and providing some type of nationalised version of these services, if they are found to be effective, may be justified.

We are able to help explain how supervision has an effect on weight loss by invoking factors such as adherence and effort. These factors make up part of the complex mechanism between intervention prescription and weight loss. As has been seen, we can also use this understanding to suggest methods by which we may improve the intervention for further testing without groping in the dark. If we didn't provide details of these mechanisms, we may have to rely on guesswork when developing new interventions before submitting them to testing. Or, if we could isolate supervision as a key factor, we may not know why it is effective, meaning we would be stuck simply trying to develop interventions where huge portions of the population are supervised. When we provide these details of mechanisms, we have a good idea of what may well work. In sports science, particularly given the limited number of trials and funding, reducing the number of trials that need to be conducted is very useful to intervention development.

7.8 Conclusion

In this chapter, I introduced exercise interventions for the treatment and management of obesity. I explained that when we compare results from RCTs examining these types of intervention, we can see that supervised interventions are much more effective than supervised interventions. The mechanisms that explain how exercise should be able to be used to treat and manage obesity were given. Following this, I argued that comparing results of RCTs does not give us a good insight into why there is a difference in effectiveness of these two types of intervention. If, however, we examine more closely details of the mechanisms that lead to differences in observed effectiveness, we can provide a much more detailed explanation. Key to this, I argued, was examining social and psychological mechanisms, alongside physical and biological mechanisms. I, then, argued that understanding these mechanisms is very useful; for instance, in identifying adherence to interventions as a key causal factor, we may be able to improve these interventions in the future, without relying on supervision to ensure adherence. I also discussed some factors that could influence adherence. Given the breadth of the obesity epidemic, and the difficulty of ensuring adherence with supervision for such a large number of people, this helps to highlight the importance of looking at mechanism details beyond what can be gleaned from comparing results of intervention RCTs. In the next chapter, I will extend the lessons from this case study, and the FIFA 11+ case study, to make a general argument about the importance of investigating mechanism details in sports science.

Chapter 8

Understanding mechanisms in sports science: extending the lessons of case studies I and II

8.1 Introduction

In this chapter, I argue that the lessons drawn from the previous two chapters extend naturally to the general case in sports science. That is, I argue that we have sufficient grounds to say that the importance of being able to provide details of mechanisms in sports science applies in the general case, and not just in those discussed. This clears up potential concerns about cherry-picking cases, and also deals with worries about how representative the case studies from the previous chapters are. This chapter, as well as extending the lessons to general cases, deals with the concern that, though sciences in general are concerned with being able to provide explanations, as EBP is primarily concerned with being able to inform practice, it need not concern itself with explanations. As Anjum and Mumford raise (2018, 89), the importance of explanations is diminished in evidence-based fields. In these fields, there is often a trend towards seeking to establish causation without understanding what underlies that causal relationship, placing an emphasis on evidence from association rather than mechanistic studies (Anjum and Mumford, 2018, 89). I argue that EBP should be concerned with being able to provide details of mechanisms by demonstrating that providing details of mechanisms can help improve our theories relating to causal relations and, then, arguing that this can improve the way we interpret research, design and test interventions, and prescribe interventions. Finally, I argue that these benefits should matter a good deal to EBP. I, then, reinforce these points with a sports science example. The work of this chapter is informed by arguments made by Gillies in Chapter 11 of his 2019 book *Causality, Probability, and Medicine*, and Anjum and Mumford's 2018 book *Causation in Science and the Methods of Scientific Discovery.*

8.2 Gillies on seeking explanations with mechanisms

Gillies argues that we may use evidence of mechanisms to develop deeper explanations in medicine, in line with the aims of science (2019, chapter 11). This view is important, in the context of my argument, as it helps to give a framework that explains how providing some details of a mechanism can fit into a larger explanation of a causal relationship. It also provides a more general illustration than the one given in the previous two chapters for how providing mechanism details can improve our understanding of causal relationships. Once I have explained his view, I will argue, using my case studies, that this applies in the case of sports science as well as medicine. Whilst Gillies analyses mechanisms as causal networks, this is controversial, as it cannot account for the organisation of entities and activities in a mechanism. This is important as understanding the organisation of entities and activities in a mechanism is part of what helps use mechanisms to provide explanations. However, the argument Gillies makes is not supported by the particular way he analyses mechanisms. This means that we can accept his conclusions without also accepting his view of mechanism.

Though I have argued we only need to know a mechanism exists in order to establish a causal claim, Gillies argues that the more we know about a mechanism M, the deeper our understanding of the relationship between a cause A and effect B is. He gives the example of the historical case of sheep dying of anthrax. In this instance, a basic causal relationship was proposed, and, then, through examining details of the mechanism that underlies that relationship, a deeper understanding of that relationship was developed. In this instance the original causal relationship A causes B was proposed:

- A: sheep graze in anthrax fields, and
- B: some sheep die of anthrax symptoms.

When the causal relationship was first proposed, so says Gillies, the fields were called anthrax fields because the sheep died of the same symptoms that would be caused by anthrax if it were present. Gillies explains that, following this discovery, research into the mechanism between A and B showed that 'anthrax fields' had anthrax spores in them. It was discovered that sheep ingested these spores, and were poisoned by them. Understanding the mechanism in this way, according to Gillies, allowed A and B to be restated as:

- A: 'sheep grazing in fields where there are many anthrax spores', and
- B: 'some of them die with symptoms of anthrax' (2019. 190)

Gillies remarks that this refinement as a result of improving our understanding of relevant mechanisms counts as a deeper understanding of the causal relationship. He uses the example in a way that is intended to be illustrative of the way providing of details of mechanisms can deepen our theories surrounding causal relationships. Gillies finishes his argument by suggesting that if one theory corrects, modifies, or refines another theory, and explains the theory, it is a deeper theory than the original. As understanding a mechanism allows us to do these things, when we understand a mechanism, we have a deeper theory. Gillies claims that this accords with, but improves, Popper's view in *The Aim of Science* (1972), that the deeper theory is the one that corrects and explains previous theories. Popper's view was intended to account for the advancement of sciences like physics, whereas Gillies wanted to extend it in such a way that it can account for other sciences, particularly medicine. Gillies' view extends Popper's by being able to account for the fact that deeper theories are those that can refine or modify original theories, as well as correct and explain. This is important so that it can account for examples like the anthrax case. It should be clear that something which furthers the aims of science, such as by providing deeper explanations, is a good thing. As such, evidence that contributes to this should be taken seriously, rather than being dismissed. Providing mechanism details, then, and evidence that allows us to do this, should be taken seriously. However, this is not yet enough of a reason for EBP to take the importance of providing details of mechanisms seriously. First, an argument must be made that this same argument can apply in sports science. Secondly, as EBP is concerned primarily with practice, and is concerned less with expanding knowledge bases, it must be argued that understanding these mechanisms has benefits that lead to the type of practical improvements EBP is concerned with.

8.3 Seeking explanations in sports science

In the previous two chapters, I provided case studies in which employing an understanding of relevant mechanisms allowed us to explain observed outcomes in a way that could not be done if the analysis of results was kept to the analysis of observed correlations, with no attention paid to details of the mechanisms that give rise to these correlations. These examples are not as simple as the anthrax example Gillies gives. This is because, as well as explaining a causal relationship between A and B, the case studies I employ explain why there is a causal relationship in some instances, and not in other similar instances. For instance, in the case of the FIFA 11+ example, there is a relationship between adopting the injury prevention programme and reduced injury rates in some populations, and not in others. If we couch this similarly to Gillies' anthrax example, we may perhaps say that:

- A_1: A number of populations are prescribed the FIFA 11+ IPP.
- B_1: A reduction in injury rate is only seen in some of these populations.

We may even be able to compare results from association studies to provide a slightly deeper theory:

- A_1: A number of populations are prescribed the FIFA 11+ IPP.
- B_1: A reduction in injury rate is seen in adolescent women, but not men of intermediate skill and above, and veteran men.

As I commented in chapter 6, before we examine details of relevant mechanisms, this is about as much detail as we are able to give in our explanation of this causal relationship. We know that the 11+ is effective in some populations, and not others, but we are unable to give a good explanation for why this is the case. However, once we look for causal factors relevant to the way in which the 11+ can intervene on injury rate, like strength, adherence, and proprioceptive ability, we are better able to explain why the 11+ is ineffective in some populations. For instance, using this information, we are able to restate the causal relationship between A_1 and B_1 as follows by adding in details X_1 between A_1 and B_1 that explain the observed outcome:

- A_1: A number of populations are prescribed the FIFA 11+ IPP to reduce injury rates.
- X_1: The 11+ intervenes on injury rates in populations where it is sufficient to improve strength, proprioceptive ability, and it is adhered to.
- B_1: A reduction in injury rate is seen in adolescent women, but not intermediate men and veteran men.

We can see here that, even when stated simply, incorporating an understanding of relevant mechanisms into our examination of the causal link between A and B allows for a deeper theory by refining and explaining the previous theory. In virtue of the discussion of these mechanisms in chapter 6, it would be possible to explain this causal relationship in even greater detail than was given here. For instance, we could fill in details explaining why adherence, and the ability to train strength, are important and why they may not occur. This would contribute further to motivating how important understanding these mechanisms is for deepening our ability to interpret study findings and understand intervention effects. However, as this has already been done, it will not be repeated. This type of explanation is a good deal better than the one offered in Soligard et al. in their discussion of why the 11+ was effective in some groups. Recall, Soligard et al. claim that:

> Our prevention programme is multifaceted and addresses many

> factors that could be related to the risk of injury (jogging and active stretching for general warm up, strength, balance, awareness of vulnerable hip and knee positions, technique of planting, cutting, landing, and running), and it is not possible to determine exactly which exercises or factors might have been responsible for the observed effects. (2008, 8)

This claim highlights nicely the limitations to the discussion of an intervention if the mechanisms by which the intervention may have its effect are omitted from that discussion.

The same is true in the case of supervised versus unsupervised exercise interventions for obesity:

- A_2: Exercise interventions are prescribed for people with obesity.
- B_2: Many people who have been prescribed these interventions are still unable to manage their obesity.

This case is particularly interesting as we may refine our theory, at first, by comparing results from association studies. This may lead people to suggest that evidence from mechanistic studies is of little importance, as we can simply understand causal relationships by comparing results from association studies. For instance, we may refine the theory by adding details X_2 between A_2 and B_2 as follows:

- A_2: Exercise interventions are prescribed for people with obesity.
- X_2 Supervised exercise interventions can produce the energy balance changes necessary to manage obesity, and unsupervised interventions cannot.
- B_2: People who are prescribed supervised exercise interventions manage their obesity better than those prescribed unsupervised exercise interventions.

This of course deepens our theory by refining it. It allows us to go some way towards explaining *why* some interventions are effective and some are not. From this, it seems like supervision is the explanation. What it does not do, however, is explain why supervision is important to the outcome of these interventions and why the presence of supervision improves intervention outcomes. If we can explain that, we once again have a better understanding of the causal relationship. Once we examine supervision, we can see that it may influence both adherence and effort. Thus, we can restate the causal relationship, including details X_2 and Y_2 as:

- A_2: Exercise interventions are prescribed for people with obesity.
- X_2 People regularly adhere to supervised exercise interventions and do not regularly adhere to unsupervised interventions.
- Y_2 Prescribed exercise interventions can only create an energy deficit if adhered to.
- B_2: People who are prescribed supervised exercise interventions manage their obesity better than those prescribed unsupervised exercise interventions.

With these two cases, what I have illustrated is that, by deepening our understanding of mechanisms relevant to causal relationships, we can provide a better explanation for that causal relationship in sports science, just as we can in medicine.

One may potentially read this and buy the claim that providing details of mechanisms is useful because it can help us improve our theories surrounding causal relationships. They may, however, not buy that this motivates the evaluation of evidence from mechanistic studies in order to do this. Can we not, as I have done in some instances, simply compare the results of lots of similar RCTs in order to uncover mechanism details or, use RCTs as to identify mechanisms? To this, I reply: yes, in theory.

However, we must remember that practicality and feasibility are necessary concerns in sports science. Considering this, and the arguments made in chapter 2, particularly those relating to sample size, it is clear that it is more or less unfeasible to try and uncover mechanism details simply by comparing different RCTs. Not only have I given reasons to believe that RCTs in sports science often do not provide strong evidence in the first place, it is also highly unlikely that it will be possible, in many instances, to conduct sufficient RCTs to uncover mechanism details as easily as can be done using mechanistic studies. There are also some cases, for instance cases in biomechanics, where we simply cannot provide details of mechanisms without conducting some amount of mechanistic research. Consider the difficulty of trying to understand *how* muscle protein synthesis can be spiked, as is necessary for the 11+ IPP to work, without doing some bench research.

An EBM or EBP proponent may, of course, counter this by suggesting that evidence-based research is not concerned with explanations at all. It is for other parts of science to concern itself with that, and for evidence-based researchers to concern itself with finding and applying interventions. I will deal with this objection in the following sections.

8.4 Anjum and Mumford on the importance of mechanistic explanations

Gillies is not the only person to discuss the importance of understanding mechanisms as it relates to theories in science. In their 2018 book, Anjum and Mumford also discuss the importance of understanding causal theories and mechanisms in science. Their work has a metaphysical bent. They make the claim that different methods of looking for causation make different claims about what causation is. Whatever method one adopts to look for causes, they argue, is used as it should capture 'some aspect of the true nature of causation' (2018, 246). For instance, if one searches for evidence of causation using methods that look to see if the presence of A makes a difference to B, this aligns one, at least in part, with a view of causation that includes difference making. As mentioned in the introduction, this thesis is focusing on epistemological issues, leaving relevant metaphysical issues for future work. As such, I will neither adopt nor argue against this claim in this thesis.

Just like Gillies, Anjum and Mumford argue that it is a goal of science that we must construct theories, this being the case despite trends found in epidemiology and evidence-based policy (2018, 89). Anjum and Mumford push forward this reasoning in a way that further benefits my argument. Not only do they argue that providing details of mechanisms can help us to understand causal claims, they also discuss some benefits afforded scientific ventures when this is done, beyond improving understanding for understanding's sake. This helps to motivate the importance of providing details of mechanisms in evidence-based spheres. The first claim of importance to my work that Anjum and Mumford make is that: 'any causal theory that is lacking in a mechanistic aspect will be incomplete in some sense because it will have no account of *how* the effect is produced by the cause'[1] (2018, 109). This compliments the idea that 'mechanistic and qualitative evidence should be best to answer how and why questions even though it tends to be lowly ranked in a standard evidence pyramid' (2018, 235). What reasoning supports this? Data of the type one may get from an RCT, they claim, can provide evidence that causal relations exist, but do not provide good explanations for how those relations arise. For instance, this type of evidence can tell us that men live longer when married, but it does not tell us what it is that is about marriage that improves life expectancy (2018, 108). We need to look at evidence that provides explanations for the underlying mechanisms to see that life expectancy is affected by 'stress, diet, lifestyle, loneliness, depression, which can be counteracted by some of the benefits marriage

[1] Italics my own, for emphasis.

brings' (2018, 108). So, without some kind of theory to explain what underlies a causal relationship, even in cases where we can effectively intervene on that relationship, we cannot properly interpret the findings of studies about those interventions as we cannot explain why they have that effect. These points made by Anjum and Mumford relate very closely to my arguments about the FIFA 11+ and exercise interventions for obesity where I argue that without examining details of mechanisms, we cannot explain the causal relationships. Their arguments help to push forward the more general case.

In addition to the claim about mechanistic evidence being important for answering how and why questions, Anjum and Mumford claim that: 'deep causal understanding comes when we have a rich theory: one that tells us not just what causes what but also how or why... a richer theory enables us to reason counterfactually about a variety of interventions and changes.' (2018, 249). To the second point, they invoke the Semmelweis case (discussed in subsection 4.5.1), as a means of illustration. They claim that as Semmelweis did not understand the mechanism explaining germ transfer, Semmelweis only knew that washing his hands reduced mortality, and not how (2018, 249). Anjum and Mumford go on to say that Semmelweis could have changed his routine in a number of ineffective ways before determining that handwashing was a key variable. To this effect, having a good idea about what may work is preferable to trial and error. For instance, if Semmelweis made different changes to his routine to see what followed, he may have put lives at risk. Again, I echo this sentiment in both of my case studies. In both case studies, I discuss how an understanding of mechanism details can help us to adapt, improve, and design interventions. Using an understanding of the mechanisms relevant to an intervention or exposure, we can reason, for instance in the FIFA 11+ case, that there is a good chance that if we increase the intensity of resistance exercises for the hamstrings for adult male populations, this may reduce injury rates by better improving hamstring strength. This argument by Anjum and Mumford, and its relevance to my sports science cases, again helps to make arguments in favour of providing mechanism details more general.

8.5 Providing mechanism details helps to interpret results, and why EBP should care

So far, in this chapter, I have argued that we may improve the theories related to causal relationships by seeking to provide details of the mechanisms that support them. I, then, argued that this can be used to answer how or why questions in sports science, and can help to interpret study results by helping us to see why the causal relationship observed arises. If EBP is practice-focused, however, and follows the evidence-based trends mentioned by Anjum and Mumford, why would being able to explain the results of trials matter to EBP? In this section, I give a number of reasons.

Practitioners should care about the interpretation of results using mechanisms, for instance, as it relates to extrapolation. That is, EBP practitioners should care that we can use our understanding of mechanisms to help determine if an intervention, which is effective in a trial population, is effective in as similar population outside a trial, and is also effective in less similar populations outside a trial. As argued by EBM+ (see subsection 5.3.2), providing details of intervention-relevant mechanisms can aid this by allowing us to compare how an intervention has its effect in a population, and how this mechanism operates in other populations.

A number of authors have commented on how one may do this, see for example: Parkkinen and Williamson, 2020; Steel, 2007. Steel (2007), is one of the first people to discuss mechanisms-based extrapolation in depth. For illustrative purposes, I will briefly explain his account. Steel argues that we may perform mechanisms based extrapolation by Comparative Process Tracing. The following is summarised from Steel, 2007, section 5.3.2. Comparative Process Tracing involves determining the mechanism in a model or test group by which something has its effect, and then comparing the mechanism and its stages in the target where the mechanisms are most likely to be different. We trace the process by which the mechanism unfolds along its causal pathway and examine the mechanism at the last point where target and model population's mechanism of action are most likely to differ. The basis for extrapolation is given based on the similarity of the two mechanisms at those crucial stages. Of course, the reliability of this method relies on how well we have identified relevant mechanisms and differences between them, but the more similar the mechanisms are, 'the stronger the basis for extrapolation' (Steel, 2007, 89).

For example, suppose there is a mechanism:

$$A \to X \to Y \to B \to Z$$

where X, Y, and Z are the points where a target and a model organism are likely to have differences in how the mechanism acts and A and B are places where they're very similar. In this instance, changes in X and Y will result in changes in Z. Steel takes this to mean that if we follow comparative process tracing, that we need only examine the mechanism at point Z. This is because any differences caused by X and Y will also lead to differences in the mechanism's working at Z. Z can be seen as a kind of "bottleneck" where the preceding differences in the mechanisms are visible (Steel, 2007, 90). If we then compare the mechanisms at Z, we will see if we have a viable basis for extrapolation if the mechanisms are suitably similar.

Despite there being a number of differing accounts of mechanism based extrapolation, I will not be arguing for a preferred one, or suggesting solutions to the problems faced by any account. What these accounts share, however, is what is important here. This is, the idea that by having some level of understanding of mechanisms relevant to an intervention and relevant populations, we can use this as part of the evidence that an intervention will work somewhere other than where it has been trialled. There are many reasons we may wish to use mechanism based extrapolation instead of, or as well as, other methods for examining interventions. This could be for ethical and practical reasons, such where intervention tests are conducted on animals before humans (Steel, 2007). Mechanisms based extrapolation is used to explain why, using mechanistic evidence, we are justified in making an extrapolation we are considering (Steel, 2007). Mechanisms based extrapolation is also important where funding does not exist to conduct a full gamut of high quality RCTs for every relevant population, or determining whether an intervention effective in one place will be effective in another place, such as with social policy (Cartwright and Hardie, 2012). Through the lens of the RWT, for example, where sufficient evidence of correlation has been collected, comparing mechanisms may be a sufficient justification for a causal claim in a different population. Or, in some cases, it may be evidence that, as mechanisms differ, further studies need to be conducted in order to be justified in making a causal claim in a new population.

For instance, providing details of the mechanism by which the 11+ works helps us to see not just that it is ineffective for adult male populations, it also helps to tell us why this is the case. This is important as it tells us that we should not employ it in those populations, and what it is about the intervention or population that makes this the case. Similarly, understanding these mechanisms gives us good reason to think that the 11+ may be effective in non-Scandanavian adolescent girls, given the relevant mechanisms differ far less in these instances, perhaps without even the need for further RCTs.

Similarly, the wide use of caffeine as an ergogenic aid can provide us with a useful example. Its effectiveness has been tested in lab conditions in relatively few sports, and at relatively few athletic competency levels compared to its widespread use (Burke, 2008). For instance, relatively few studies have been conducted on elite athletes, the majority having been conducted on low-level athletes. Despite the mechanisms by which caffeine has its ergogenic effect not being fully known, enough is known about the similarity between mechanisms involved in performance in endurance sports to be able to extrapolate its effect between sports, and between ability levels (L. Burke, 2008). For example, we know that caffeine can increase the strength of muscular contractions by increasing the amount of calcium available to muscles, and we also know that increasing the amount of calcium available to muscles can improve performance in endurance sports (L. Burke, 2008; Tarnopolsky, 2008). What this means is that, despite not having tested the use of caffeine in some sports, we know that it can have its ergogenic effect, and therefore improve performance in those sports because the mechanism of action is consistent across those sports. Through the lens of the RWT, we know a correlation exists between caffeine consumption and enhanced performance in some sports. Further, by looking at details of relevant mechanisms we know that, for endurance sports, one way in which this happens is by increasing muscle calcium availability. We also know that performance in other endurance sports can be enhanced by improving muscle calcium availability. As such, when we know that muscle calcium availability is a limiting factor in performance in a sport, we can, and do, justifiably infer that supplementing caffeine can enhance performance in those sports.

There is also a mirror for this type of practice in medicine. As Wilde and Parkkinen (2019) discuss, when IARC establish the carcinogenicity of certain agents, it is not always possible to conduct studies in humans. In these instances, the similarity to mechanisms in animals (where tests can be conducted) can be assessed to determine if the mechanism by which something is a carcinogen can be sufficient to infer a causal relationship in humans. This was the case, they argue, with benzo(a)pyrene. In this instance, there were association studies on benzo(a)pyrene, but these were insufficient to establish a causal claim. However, when relevant mechanisms were identified, they were sufficient to establish the causal claim.

From these examples, we can clearly see that providing details of these mechanisms is useful as it gives us reasons to justify inferences about the effects of different interventions across sports, and between ability levels. This is particularly useful in cases where we do not want to, or cannot, conduct high-quality association studies in all areas where we may want to employ an intervention or exposure.

There is a second prong to the discussion of mechanisms and populations. Providing details of intervention relevant mechanisms can also help us to stratify populations, which is useful to see if further research may need to be conducted to establish efficacy in other groups. For instance, if we provide intervention relevant mechanism details, this can help to uncover populations that may respond differently to an intervention, and whether this has an effect on effectiveness. The 11+ can be used as an example. Adult men of intermediate and above skill, and veteran men, are sufficiently dissimilar from adolescent girls, and sufficiently similar to each other, that we may consider them to be one relatively homogeneous population that react similarly to the 11+. However, by examining intervention relevant mechanism details we can ascertain that the 11+ is ineffective in both groups, but for different reasons. By looking at details of relevant mechanisms, this helps us to split two otherwise seemingly similar sets of people into different populations.

The rehydration case discussed in the introduction, and again in subsubsection 2.3.2.1, also furnishes us with a useful example of the uses of providing mechanism details, here. Recall, based on evidence from a trial in which it was noted that runners with higher internal body temperatures at the end of the race had also lost more body-weight during a race through sweat loss; it was suggested that the sweat loss *caused* the high internal temperature. This led to the widespread advice that athletes should drink as much as is tolerable during exercise to reduce the risk of heat illness, and to improve performance. We now know that this advice is flawed and that maximal fluid ingestion can be dangerous. Further, we now understand the mechanism linking sweat loss and high internal temperatures, and know that they have a common cause: metabolic rate. The faster a runner is, the higher their metabolic rate and, therefore, the greater their sweat loss and the higher their internal temperature. This example is useful for my argument as it provides us with a case where, if we had understood, or sought to understand, the mechanism underlying the proposed link between internal temperature and sweat loss, we would have known that rather than that relationship being causal, there is a common cause. We would, then, not have promoted a dangerous intervention. This case also illustrates how knowing details of intervention relevant mechanisms can help us to explain that observed correlations are not causal, and how they instead arise.

Providing details of intervention relevant mechanisms also helps us to contextualise the results of association studies. Take, for instance, the case of β-alanine, discussed in section 5.5. It provides a clear example in which an understanding of the mechanism by which β-alanine works was needed to explain the results observed by RCTs. Providing details of the mechanism by which β-alanine has its effect was vital to explaining why it improves exercise performance only for specific durations of exercise. Providing details of mechanisms in this sense can help practitioners decide if supplementing it is suitable, given the event and performance outcomes they are concerned with. This example can, of course, be seen in a more broad sense. Not only does understanding mechanisms help us by providing some evidence of where an intervention may work, it can also help us to understand if an intervention will be useful given the outcomes we are concerned with.

The case of exercise interventions for obesity also provides motivation for the importance of providing details of intervention relevant mechanisms. Unsupervised exercise interventions are widely, and ineffectively, prescribed. For instance, as is detailed in chapter 7, the NHS prescribes exercise without concern for supervision. Where details of the relevant mechanisms are understood, perhaps they would not be prescribed, and suitable alternatives could be found. One could suggest that prescribing largely ineffective interventions such as this is, on a public health level, a cost-effective solution. However, as will be seen in the next section, providing mechanism details can be used to develop other cost-effective interventions.

So, providing details of mechanisms is useful in an EBP context, as it can help to interpret study findings by telling us where an intervention may or may not be useful. Further, comparing details of mechanisms may help to support evidence from RCTs in other populations. This is of particular concern to EBP practitioners because it can stop them employing ineffective interventions, perhaps wasting athlete time, money, and effort. It can also help us to understand and contextualise findings, helping us to see if intervention outcomes are relevant to our desired ones. It can also tell them that perhaps an intervention could be adapted in such a way that it may be effective in a population in which it currently is not, which leads to the next section.

8.6 Providing mechanism details aids intervention development, and why EBP should care

In the previous two chapters, I argued that if we deepen our understanding of the mechanisms relevant to an intervention, for instance by identifying mechanism details and key causal factors, this can aid further research and intervention development, as well as providing explanations for causal relationships. These examples are intended as illustrative of what may be possible in sports science more widely, not simply those two cases. For instance, in the case of the FIFA 11+ IPP, we have an intervention that we know is highly effective in some populations, but that is ineffective for the most at risk group in football. Given the risk of injury in football, it is important to develop an effective IPP for this population.

When we examine the mechanisms relevant to the 11+, this helps us to understand why this intervention is ineffective in some populations. For instance, in the case of the 11+, it tells us that it was not sufficiently high in volume and intensity to train strength and proprioceptive ability in some populations in which it was ineffective. As was discussed in great detail in section 6.4, when examined, this can help us see ways in which we could develop a new and improved intervention to target these groups. As was suggested, this improved understanding provides us with some evidence that if we were to increase the volume and intensity of resistance exercises in the 11+, it may be effective for men of intermediate skill levels and above at reducing injury rate. We have evidence for this because providing mechanism details tells us that hamstring strength is a key causal factor in hamstring injury, and our understanding of strength training tells us that the inability of the 11+ to intervene on this in some populations could be changed if intensity and volume of hamstring training were increased in the IPP.

Of course, in line with the RWT, utilising our understanding of the relevant mechanisms in this way is unlikely to provide sufficient evidence to *establish* a causal relationship between an improved 11+ and reduced injury rates in intermediate men, even if the mechanisms supporting this claim may be established. Association studies are still likely to be needed. However, as was suggested by Anujum and Mumford, our understanding of these mechanisms can be used to streamline this process. For instance, by developing interventions we have a good idea will work for specific reasons before we test them, we are better served than testing a multitude of new interventions with different changes until something works. This second option is the type of thing that may happen when your explanation for the effectiveness of the 11+ leaves the mechanisms underlying the causal relationship as black-boxes. A black-box does not allow you to suggest which parts of the 11+ were actually beneficial for reducing injury rates, as was illustrated in a section 8.3.

The case of exercise interventions for obesity also gives us an illuminating example of how intervention development can be aided in the sports sciences with an understanding of relevant mechanisms. As I suggested in chapter 7, supervision is a key causal factor in exercise interventions for obesity. If we, then, seek to understand how supervision plays a role in the effectiveness of these interventions, we can see that it can increase both adherence to exercise, and the effort used in exercise. These can, in turn, improve the effectiveness of exercise interventions for obesity. From this, we can tell that prescribing interventions that improve adherence and effort can possibly be effective methods of improving exercise interventions for obesity. I also suggested that, given the scale of the obesity epidemic, it is almost impossible to supervise the exercise of every member of the public who may be prescribed an exercise intervention for obesity. As such, improving adherence and effort via supervision is unfeasible and another method of improving those factors must be found. This allowed me to look at alternative methods of improving adherence used in other obesity interventions, such as simulating supervision by employing regular weekly group meetings. Using relevant mechanism details allows us to suggest these types of improvements to interventions for further testing with good justification. If we did not examine the way in which supervision had an effect on intervention outcomes, and instead were only able to point to supervision, we would have a harder time suggesting and justifying these new types of intervention for trial.

To generalise the points from this section: providing details of the mechanisms relevant to interventions helps us suggest ways we can improve and adapt interventions for future testing, and gives us good justification for these improvements and adaptions to these interventions. If we do not do this, we may be left in a position where we have to keep trying different interventions until something works, or run a whole gamut of iterative association studies, unable to know what part of the intervention is effective, without comparing different iterations. So, understanding relevant mechanisms can potentially cut down research time and costs. It should be clear why this, and evidence which helps us do this, should be important to EBP. If we do not provide details of mechanisms and cannot answer important how and why questions, we miss out on an important avenue of evidence that could help us to adapt, improve, and develop the best possible practices, and evidence that can help us streamline this process.

8.7 A more in depth example from sport

Here, I provide two examples: one from medicine, and one from sport, These examples help to illustrate and provide more force to the arguments from this chapter. They exemplify the importance of being able to provide details of mechanisms in an evidence-based context, and show how important it is in helping us to practice effectively.

In medicine, the case of the investigation of streptomycin as a treatment for tuberculosis is used to illustrate the importance of being able to provide details of mechanisms when assessing how to employ a treatment. Gillies (Gillies, 2017a) gives an overview of the case. In the instance of streptomycin, RCTs were carried out and recovery rates from tuberculosis were much higher in groups who received streptomycin than those who were given bed rest as an active control (Gillies, 2017a, 59). However, 5 years after the trial, the difference in number of patients who died in the control and treatment groups was no longer statistically significant, raising concerns about the viability of streptomycin as a treatment (Gillies, 2017a, 59). The reason this measurement was taken 5 years post trial is because the researchers conducting the trial took note of the mechanism by which streptomycin has its intended effect (Gillies, 2017a, 59). It was known that streptomycin disposed of the strain of bacteria responsible for tuberculosis slowly in comparison to other antibiotics. Accordingly, it was supposed that during streptomycin treatment, streptomycin resistant strains of the bacteria responsible for tuberculosis may develop in patients. As such, where this occurred, not only would a relapse be likely, but a new treatment with streptomycin would be ineffective. Because this was noted, it was decided to try combination treatments of streptomycin and another drug, para-amino-salicylic acid (PAS) which would inhibit the growth of the bacteria responsible for tuberculosis, limiting the development of antibiotic resistant strains. It was later found that a streptomycin and PAS combination treatment not only limited the development of antibiotic resistant strains, but also allowed for repeated courses of streptomycin treatment when needed (Gillies, 2017a, 59-60). Gillies uses this case to argue in favour of the need to have evidence of mechanisms when establishing causal claims, a case I already made in Part II of this thesis. This case is useful in this part of the thesis because it illustrates how, when we can provide details of mechanisms, it helps us interpret the results of trials, and helps us know how to employ a treatment in medicine. Interesting parallels for this case can be found in sports science, as I will argue.

The importance of providing details of the mechanisms related to an outcome of interest, as it relates to interpreting intervention effectiveness and applicability, can be highlighted by a recent review (Kuikman et al., 2021) discussing the treatment of Relative Energy Deficiency in Sport (RED-S). This review considered how important understanding the mechanisms by which RED-S arises are when determining how to treat RED-S. The review combined evidence from RCTs on RED-S treatments and a mechanistic understanding of RED-S causes and treatments to propose treatments that, given these mechanisms, are more viable than those that may be suggested without taking these mechanisms into account. RED-S is caused when an athlete's energy intake is chronically lower than their exercise energy expenditure. RED-S is defined as 'impaired physiological functioning caused by relative energy deficiency and includes, but is not limited to, impairments of metabolic rate, menstrual function, bone health, immunity, protein synthesis, and cardiovascular health' (Mountjoy et al., 2014, 491). The review noted that commonly prescribed treatments for RED-S include lowering an athlete's energy output by getting them to exercise less, whilst increasing their energy intake by asking them to eat more, with the aim of achieving a healthy energy balance and restoring healthy function (Kuikman et al., 2021, 268-269). The review notes, however, by pointing out key causal factors in the mechanism by which athletes get RED-S, that it will often be unwise to treat RED-S by encouraging athletes to exercise less and eat more. For instance, athletes who get RED-S as a result of having a very high-energy output will likely be highly committed to their training and unwilling to cut the amount of exercise they do in order to treat RED-S (Kuikman et al., 2021, 268). In addition to this, RED-S treatment is associated with weight gain, as a result of the shifted energy balance, and athletes may worry that this can impact performance (Kuikman et al., 2021, 268). Further, some athletes may find it too difficult to eat the amount needed to sustain their training (Kuikman et al., 2021, 272).

Noting these key causal factors in the mechanism by which RED-S may occur in athletes, the review discussed alternative treatment strategies which are more likely to be effective, and which account for this mechanism. For instance, it was noted that changing *when* athletes ate could improve symptoms of RED-S by reducing the duration over which an athlete's lack of energy intake caused muscle breakdown (Kuikman et al., 2021, 271). Increasing the amount of an athlete's diet that is made up of carbohydrate, and reducing fibre intake, in relation to other nutrients was also noted as improving RED-S symptoms (Kuikman et al., 2021, 271-272). Interventions like this take note of mechanisms which may make high-energy intake difficult for athletes, whilst still improving

RED-S symptoms. Symptoms of RED-S can be further reduced, without an athlete drastically reducing exercise volume in order to decrease energy output, which they may be unwilling to do, by including exercises that increase bone strength, such as resistance exercises and weight lifting (Kuikman et al., 2021, 272). In addition to this, athletes can treat RED-S symptoms by engaging in stress management techniques (Kuikman et al., 2021, 273).

What the case of RED-S treatments shows, similarly to the case of streptomycin, is that, although we have evidence from some trials that some treatments *should* be effective treatments for RED-S, this does not give us a full picture of how those treatments will work in non-clinical settings. Once details of the relevant mechanisms are considered, this allows us to better interpret the results of these trials and propose more effective treatments. This, then, provides an important example of how providing details of mechanisms in sports science helps us to see where different interventions may be effective, and why. It also helps us to see how we may change or adapt the interventions we have in order to improve their efficacy.

8.8 Conclusion

In this chapter, I explained Gillies' view on using mechanisms to improve our theories, in line with the aims of science. I, then, argued that the view, as he argues for it in medicine, applies also to sports science by illustrating it with my case studies from the two previous chapters. I have also illustrated Anjum and Mumford's arguments that having deeper theories that utilise understanding of mechanisms is important, as it helps us to answer important how and why questions in science. I, then, argue that this is important as it helps both in the interpretation and contextualisation of results from studies, and, also, in the development of new interventions. Both of these, I argued, should be more important to EBP than current trends suggest that they are. As such, I argue, evidence in the sports sciences which can help to improve our theories should be taken seriously. So, evidence from mechanistic studies is useful in the sports sciences.

Chapter 9

Conclusion

9.1 Summary

This thesis has motivated the importance of assessing evidence from mechanistic studies in EBP and sports science. I supported this with three broad claims. In Part I, I argued that the evidence produced by RCTs and RCT-style N of 1 trials in sports science will often be low quality, often unavoidably. In response to this, in Part II of this thesis, I argued for the **Better Evidence Thesis**. This is the claim that we better fulfil the goal of the relying on best possible evidence when we perform and assess evidence from association studies and mechanistic studies together than if we assess only evidence from association studies. In Part III of the thesis, I argue that we ought to assess mechanistic studies beyond their ability to help provide evidence of mechanism for establishing causation. This is because assessing mechanistic studies aids us in both our ability to interpret the results of studies, and also in the development and adaption of interventions. These two claims motivate the main argument by giving strong reasons why the assessment of evidence from mechanistic studies aligns with the goals of EBP and sports science.

Part I of the thesis argued that the goal of relying on the best possible evidence, a guiding principle of EBP, is not well fulfilled if the general EBP guideline of taking RCTs as the best primary evidence gathering source are followed. This was because the nature of the type of research conducted in sports science means that RCTs very often produce low-quality evidence as they fail to rule out explanations other than that the intervention or exposure being tested was the cause of observed outcomes. Where this is the case, the evidence produced is of a low quality. I argued that this was the case using the **Excluded Explanations Argument**. This argument states that:

- An RCT is meant to provide high-quality evidence on the grounds that it rules out alternate explanations for trial outcomes.
- If things other than the intervention or exposure being tested can explain the outcome of a trial, we cannot rule in the intervention or exposure being tested as the only explanation for observed outcomes.
- The less well RCTs rule out alternate explanations for differences in observed outcomes, the lower the quality of evidence those RCTs produce is. This is, in part, because we have less confidence in our ability to rule *in* the intervention or exposure being tested as a cause of those differences.

This, taken with my arguments that RCTs in the sports sciences will often unavoidably be unable to fulfil the requirements needed for an RCT to rule out alternate explanations for observed outcomes, explains why RCTs in the sports sciences will often provide low-quality evidence.

Following this, I defended the RWT in medicine, and argued that it also applies in sports science. I did this by arguing that to establish causal claims, sports science needs mechanisms to show that an observed correlation is causal, and correlations to provide evidence for net effects of mechanisms. This provides epistemic rationale for the RWT. I also used the historical case of what it took to establish the effects of creatine supplementation on performance outcomes as a practical example of the application of the RWT in sports science. Part II of the thesis concluded by taking the practices of IARC, and the work of EBM+ (which is informed by the RWT), as motivation for the claim that we may better fulfil the goal of relying on the best possible evidence in sports science if evidence from mechanistic studies is assessed alongside evidence from association studies like RCTs, motivating the **Better Evidence Thesis**. This is grounded in the idea that assessing both types of study helps us to establish both a correlation and mechanism claim, and that it shores up the evidential foundations for causal claims above those made on the basis of evidence from RCTs only.

Part III of this thesis continued to motivate the importance of assessing evidence from mechanistic studies in sports science and EBP by arguing that, not only should we seek to establish that mechanisms exist, but, also, we should seek to try and provide relevant details of those mechanisms. The basis for this claim was that where we can provide the details of a mechanism we can improve study interpretation, and improve and streamline intervention design. I argued that study interpretation, for instance, is improved where we provide details of mechanisms by helping us to see where the results of some set of studies may be applied in the real world. I employed two case studies in order to make these arguments, the FIFA 11+ IPP, and the case of exercise interventions for obesity. In both of these cases, I argued that without providing details of relevant mechanisms, we cannot explain important intervention relevant outcomes. Presenting the case studies also motivated the feasibility of this type of work. I, also, argued, using these cases, that if we provide details of these mechanisms, we can improve future intervention design and development. Using these two case studies, and work by Gillies, and Anjum and Mumford, in chapter 8, I argued that we can extend the lessons from the two case studies naturally to make a general case. I provided an argument about the importance of providing details of mechanisms, as it relates to, not just improving our theories, but also the benefits providing these mechanism details afford the sciences.

Just as EBM+ can be seen as a step forwards in the methodology of evidence assessment paradigms in medicine, this thesis can be seen as providing an argument in favour of further improving evidence assessment paradigms in sports science, so we can find out what to do, and know we are justified in doing it.

9.2 Work for the future

As we are at the end of this thesis, it will be important to talk about work for the future.

One of the first things to consider is areas where the research done here may also have applications in other fields. One key thing this thesis did was look at the work of the EBM+ group, which, by focusing their attention on medicine, argued for a method of evidence evaluation in a science that is largely biological. There are, of course, notable exceptions. Russo, for instance, treats epidemiology as a social science in her book Causality and Causal Modelling in the Social Sciences (2008). By approaching sports science, a field where physical and biological, and psychological and social mechanisms intertwine and produce complex multi-causal relationships, this thesis helps to provide the beginning of similar arguments in similar sciences where these types of multi-causal relationships exist, particularly those areas where an aim is to inform practice using an evidence-based framework. For instance, sciences concerned with climate change, and intervening on it, will involve looking at mechanisms that concern sociological and biological mechanisms, as well as physical mechanisms that affect how the world works, such as those from physical geography and the atmospheric sciences. Thus, some arguments in this thesis may be found to have application outside sports science in other sciences where there is a broad overlap between these types of mechanisms.

Discussions of metaphysical implications that the arguments of this thesis have, and metaphysical implications that work this thesis may rely on, have been sidestepped in order to make room for the epistemological and practical content of this thesis. This practical and epistemological focus has been maintained throughout this thesis, mainly because the goals of EBP itself are its practical ends. However, understanding what metaphysical commitments are made by the arguments in this thesis will be an important project for the future as it may have entailments for how we understand and interpret research findings outside the practical sphere, on a more theoretical level. One thing that will be important to look at is whether, as Anjum and Mumford suggest, different research methods make different metaphysical assumptions about causality. This may have implications, for instance, where different parts of sports science use different methods of investigation. Does this mean causality is, in some way, different between those different parts of sports science? For instance, physical and social mechanisms are researched in different ways in the sports sciences (McFee, 2009). This may say something about how those mechanisms exist, and whether it means that they exist in different ways.

Related to this previous point, work needs to be done to develop tools specific for sports science that allow the integration of evidence of both physical and social mechanisms and correlations together, and to understand and represent how strong the evidence we have for certain claims, and causal factors underlying those claims, is. A call is being made in sports science for the use of complex systems analysis in the understanding of phenomena (Bittencourt et al., 2016). This is motivated by its use in fields like epidemiology, economics, and biology. One place that may be interesting to look is at the use of causal networks to represent causal relationships. These may be useful for helping to represent the multi-causal and indeterministic causal relationships in sports science, as Gillies argues they are in the case of medicine (Gillies, 2019). Using causal networks to represent causal relationships may be useful because it provides a range of intervention entry points for researchers (Greenland et al., 1999; Joffe et al., 2012). What this means is that by examining causal networks that represent relationships, we may be able to determine more easily important parts of the causal relationship that may be intervened on, what affects them, and what changing them will affect. The use of *conjectural* causal networks in epidemiology may also have uses in sports science. In epidemiology, causal networks can be conjectured, where the quality of evidence for a link between causal factors can be represented visually on the network (Joffe et al., 2012). Because of this, we can update the conjectured network as research is conducted. Conjectural causal networks allow us (Joffe et al., 2012, 9):

- ‘to make assumptions and hypotheses explicit for discussion;
- to place hypotheses in the public domain prior to testing - a conjecture that is open to refutation;
- to plan data collection;
- to structure the statistical analysis of the hypothesised pathways;
- to identify evidence gaps and therefore generate a research agenda’

All of which may be useful in sports science.

Another thing not discussed in this thesis, that will be important to look at in future work, and for EBP, is how expert opinion and understanding can be integrated into sport. I have laid out a way by which we may better fulfil the goal of relying on best possible evidence, in order to establish causal claims, motivating the **Better Evidence Thesis**, but what I have not done is talk about how we may move from those causal claims to really understanding how and when to practice based on those claims. Of course, EBP sees expert opinion as a low-quality source of evidence, and I do not contest this, but there will still be a need for experts to exist to interpret and apply practices once the evidence is produced. How this may be done, what limitations exist, and whether there may be better ways of doing this, is left for future work.

Chapter 10

Glossary of key arguments

10.1 The disambiguated 'Russo-Williamson Thesis'

The disambiguated RWT can be stated as follows:

> In order to establish a causal claim in medicine one normally needs to establish two things: first, that the putative cause and effect are appropriately correlated; second, that there is some mechanism which explains instances of the putative effect in terms of the putative cause and which can account for this correlation (Williamson, 2019, 33)

This thesis is supported by the epistemic rationale given in section 4.2. It is also supported by historical cases such as the Semmelweis case discussed in subsection 4.5.1. The epistemic rationale can be summarised simply.

Evidence that a mechanism exists is insufficient to establish that an intervention has an effect because it does not tell us whether that intervention has a net effect on an outcome, and if it does have a net effect, it does not tell us the extent of that effect. Conversely, the observation of a correlation between an intervention and a measured outcome is insufficient to establish that the intervention causes that effect because the observed correlation may have other explanations, such as chance or confounding. However, evidence of mechanism counters the flaws faced by evidence of correlation, and evidence of correlation counters the flaws of evidence of mechanism. Where we have identified a mechanism, if we find a correlation, we can see what net effect, if any, the intervention has. Where we observe a correlation, finding that a suitable mechanism exists tells us that there is a way in which the correlation we observe can be caused by the intervention under investigation. Accordingly, to establish that a causal claim is genuine, we must establish both that a correlation exists between intervention and outcome, and we must establish that a suitable mechanism exists to explain that correlation.

The disambiguation of the RWT makes it clear that what is needed to establish causal claims is that we establish that a mechanism exists, we do not need to establish the details of that mechanism. Clearly, however, finding the details of a mechanism can be used to support the claim that a mechanism exists. What the disambiguation the RWT does is to explain how we may, in some instances, establish causal claims without knowing the details of a mechanism, such as in cases where effectiveness is established on the basis of association studies alone. Williamson's inference gives one set of sufficient conditions for establishing that a mechanism exists on the basis of evidence from association studies is given in section 4.6.

10.2 The EBM+ position

The EBM+ position states that, in medicine, when assessing causal claims, we ought to assess evidence from association studies like RCTs, and mechanistic studies, together. This is supported by the RWT, but is not to be conflated with it. EBM+ provides a methodology that guides us in establishing causal claims in medicine, given the implications of the RWT.

The rationale for the EBM+ position can be explained as follows:

> In general, association studies provide relatively good evidence for correlations, but relatively poor evidence in favour of the existence of mechanisms. Conversely, mechanistic studies provide relatively good evidence that a mechanism exists by providing evidence for details of those mechanisms, but relatively poor evidence of correlation. What this means is that, as we need to establish both a mechanism and a correlation when assessing a causal claim, we ought to assess evidence from mechanistic studies and association studies in order to establish a correlation and a mechanism.

This does not mean that we cannot establish a causal claim on the basis of evidence from either mechanistic or association studies alone. What it means is that as it is difficult to establish both a correlation and a mechanism on the basis of one type of study, we ought to assess both. There are of course some instances where we may establish causal claims on the basis of only one type of study, however. For instance, mechanistic studies may suffice to establish causal claims where effect sizes are suitably large, and RCTs may establish causal claims where they meet the sufficient conditions given in section 4.6.

10.3 The ‘Excluded Explanations Argument’

The excluded explanations argument states that RCTs in the sports sciences regularly provide, often unavoidably, low-quality evidence for causal claims.

RCTs provide stronger evidence the closer they are to ideal. Close-to-ideal RCTs have large sample sizes, adequate placebo controls, and adequate double blinding. The smaller sample sizes are, the less adequate placebo controls are, and the less good blinding is, the worse the evidence an RCT produces is. This is because, the worse these criteria are met, the more likely it is that something other than the intervention or exposure under investigation explains the observed outcomes of the trial. This means that the evidence they produce in favour of a causal claim is of low quality because, in principle, future evidence could change the confidence we have in a claim supported by that evidence. This is the case because, in principle, we could provide evidence for the causal claim that rules out all other alternate explanations for observed outcomes which would change our confidence in that causal claim.

So, RCTs in sports science often provide low-quality evidence because they are often far from meeting these criteria, as is argued in section 2.3. The excluded explanations argument also applies to N of 1 trials because, by being hard to blind and adequately placebo control, they also likely fail to exclude alternate explanations for observed outcomes with a high degree of rational confidence. As is also argued in chapter 3 and chapter 2, these difficulties in adequately blinding, placebo controlling, and sampling, will be unavoidable in much of sports science due to the nature of the science, meaning the quality of evidence that can be produced by RCTs in sports science is often unavoidably low.

10.4 The ‘Better Evidence Thesis’

EBP, motivated and informed by EBM, seeks to inform decisions and base practice on the best possible evidence. As is explained in chapter 1, this manifests itself in the privileging of evidence from RCTs, and the dismissal of evidence from other sources. The **Better Evidence Thesis** is motivated by the spirit of the desire to inform decisions and base practice on the best possible evidence, in line with the goals of EBP. The **Better Evidence Thesis** however, is a rejection of the idea that RCTs should be privileged when this is the goal. The **Better Evidence Thesis** states that, in sports science, if we assess evidence from mechanistic studies and association studies like RCTs, we will be relying on better evidence than if we rely primarily on evidence from RCTs alone. This thesis is supported by three main claims. Firstly, the standard of evidence from RCTs and N of 1 trials in sports science is particularly low as the trials are often unavoidably far from ideal; as such, they are not a good guide for establishing intervention effectiveness, and accordingly are not good guides for practice or decision-making. Secondly, the thesis is supported by the fact that the RWT applies in sports science, meaning that in order to establish causal claims we need to establish both the existence of a mechanism and a correlation. Third, the practice of assessing evidence from both mechanistic studies and association studies like RCTs is used fruitfully in medicine, such as in the example of IARC, and has strong epistemic grounding in the case of EBM+. Taken together, what these points mean is that RCTs in sports science will often, unavoidably, be insufficient to establish both a correlation and a mechanism, and we may have better evidential support for causal claims when we assess evidence from association studies and mechanistic studies together, particularly when those association studies provide low-quality evidence. Accordingly, we reach the **Better Evidence Thesis**: when assessing causal claims in sports science, we should assess evidence from mechanistic and association studies together instead of privileging evidence from RCTs alone in order to better fulfil the EBP goal of relying on the best possible evidence.

10.5 The 'Harm Profile Thesis'

Put the most simply, the **Harm Profile Thesis** states that, in general, the harms associated with sports science interventions are often less critical harms than those associated with medical interventions. This argument is explicated in section 3.4. It is supported by illustrative examples of the harm profiles associated with sports and medical interventions, and Stegenga's argument (2018, 144, and throughout) that many medical interventions have particularly troubling harm profiles. A caveat is given to the thesis, however. This is that the perceived impact of harms associated with sports science interventions may be variable. If a sports intervention has no physical or mental harms associated with it, but is a poor performance enhancer, particularly when other more beneficial interventions may be used, more elite athletes may consider that to be a harm.

The **Harm Profile Thesis** is used to support two main claims. Firstly, it supports the claim that, when conducting N of 1 trials, the problem of lack of generalizability of findings is not as pressing in sports science as it is in medicine. This is because, whilst we will still not be able to generalize claims about intervention effects from N of 1 trials, we do not need to worry that when conducting N of 1 trials we are not collecting generalizable evidence about potential harms, because they are likely minimal.

The second claim the **Harm profile Thesis** supports can be found in subsection 5.6.4. The claim is that although evidence supporting the effectiveness of interventions in sports science may be insufficient to establish effectiveness in some cases, we may find that there are instances where it is still suitable to engage in a practice, particularly where it is better to do something, than do nothing. The **Harm Profile Thesis** supports this because, unlike many cases in medicine, even where an intervention does not have its intended beneficial effect, we need not worry that it will have some critical harm profile associated with its adoption.

Bibliography

Ahtianinen, J., Pakarinen, A., Alen, M., Kraemer, W., & Hakkinen, K. (2003). Muscle hypertrophy, hormonal adaptations and strength development during strength training in strength-trained and untrained men. *European Journal of Applied Physiology*, *89*, 555–563.

Alderwick, H., & Dixon, J. (2019). The NHS long term plan.

Amonette, W., English, K., & Ottenbacher, K. (2010). Nullius in verba: A call for the incorporation of evidence-based practice into the discipline of exercise science. *Sports Medicine*, *40*(6), 449–457.

Amonette, W., English, K., & Kraemer, W. (2016). *Evidence-based practice in exercise science: The six-step approach*. Human Kinetics.

Anjum, R. L., & Mumford, S. (2018). *Causation in science and the methods of scientific discovery*. Oxford University Press, USA.

Araujo, A., Julious, S., & Senn, S. (2016). Understanding variation in sets of n-of-1 trials. *PLoS One*, *11*(12). https://doi.org/10.1371/journal.pone.0167167

Arnason, A., Anderson, T., Holme, L., Engebresten, L., & R, B. (2008). Prevention of hamstring strains in elite soccer: An intervention study. *Scandanavian Journal of Medicine and Science in Sports*, *18*(1), 40–48.

Ashcroft, R. E. (2004). Current epistemological problems in evidence based medicine. *Journal of Medical Ethics*, *30*(2), 131–135.

Atkinson, G., & Nevill, A. (2001). Selected issues in the design and analysis of sport performance research. *Journal of Sports Sciences*, *19*(10), 811–827.

Auker-Howlett, D., & Wilde, M. (2019). Reinforced reasoning in medicine. *Journal of Evaluation in Clinical Practice*, *26*(2), 458–464.

Bahr, R., & Krosshaug, T. (2005). Understanding injury mechanisms: A key component of preventing injuries in sport. *British Journal of Sports Medicine*, *39*(6), 324–329.

Baker, S., Jung, A., & Petrella, J. (2011). Presence of observers increases one repetition maximum in college-age males and females. *International Journal of Exercise Science*, *4*(3), 199–203.

Ballor, D., & Keesey, R. (1991). A meta-analysis of the factors affecting changes in body mass, fat mass and fat-free mass in males and females. *International Journal of Obesity*, *15*(11), 717–726.

Balsom, P. D., Söderlund, K., & Ekblom, B. (1994). Creatine in humans with special reference to creatine supplementation. *Sports Medicine*, *18*(4), 268–280.

Bandegan, A., Courtney-Martin, G., Rafii, M., Pencharz, P. B., & Lemon, P. W. (2017). Indicator amino acid–derived estimate of dietary protein requirement for male bodybuilders on a nontraining day is several-fold greater than the current recommended dietary allowance. *The Journal of Nutrition*, *147*(5), 850–857.

Barbosa, G. M., Trajano, G. S., Dantas, G. A., Silva, B. R., & Vieira, W. H. B. (2020). Chronic effects of static and dynamic stretching on hamstrings eccentric strength and functional performance: A randomized controlled trial. *The Journal of Strength & Conditioning Research*, *34*(7), 2031–2039.

BASES. (n.d.). *About sport and exercise science* [date accessed = 26/6/19]. https://www.bases.org.uk/spage-about_us-about_sport___execise_science.html

Beach, D. (2021). Evidential pluralism and evidence of mechanisms in the social sciences. *Synthese*, 1–21.

Beauchamp, M. R., Ruissen, G. R., Dunlop, W. L., Estabrooks, P. A., Harden, S. M., Wolf, S. A., Liu, Y., Schmader, T., Puterman, E., Sheel, A. W., et al. (2018). Group-based physical activity for older adults (goal) randomized controlled trial: Exercise adherence outcomes. *Health Psychology*, *37*(5), 451–461.

Beedie, C., Benedetti, F., Barbiani, D., Camerone, E., Lindheimer, J., & Roelands. (2020). Incorporating methods and findings from neuroscience to better understand placebo and nocebo effects in sport. *European Journal of Sport Science*, *20*(3), 313–325.

Beedie, C., Whyte, G., Lane, A. M., Cohen, E., Raglin, J., Hurst, P., Coleman, D., & Foad, A. (2018). Caution, this treatment is a placebo. it might work, but it might not: Why emerging mechanistic evidence for placebo effects does not legitimise complementary and alternative medicines in sport. *British Journal of Sports Medicine*, *52*(13), 817–818.

Befort, C., Donnelly, J., Sullivan, D., Ellerbeck, E., & Perri, M. (2010). Group versus individual phone-based obesity treatment for rural women. *Eating Behaviors*, *11*(1), 11–17.

Behera, M., Kumar, A., Soares, H., Sokol, L., & Djulbegovic, B. (2007). Evidence-based medicine for rare diseases: Implications for data interpretation and clinical trial design. *Cancer Control*, *14*(2), 160–166.

Bernhardsson, S., Klintberg, I. H., & Wendt, G. K. (2011). Evaluation of an exercise concept focusing on eccentric strength training of the rotator cuff for patients with subacromial impingement syndrome. *Clinical Rehabilitation*, *25*(1), 69–78.

Billings, J., Hashem, F., & Macvarish, J. (2010). Am i bovvered? a participative action research study to develop, implement and evaluate physical activity interventions with girls. phases two and three report. *Centre for Health Services Studies, University of Kent.*

Birkett, N., Al-Zoughool, M., Bird, M., Baan, R., Zielinski, J., & Krewski, D. (2019). Overview of biological mechanisms of human carcinogens. *Journal of Toxicology and Environmental Health, Part B*, *22*(7-8), 288–359.

Bittencourt, F. N., Meeuwisse, W. H., Mendonça, L. D., Nettel-Aguirre, A., Ocarino, J. M., & Fonseca, S. T. (2016). Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition-narrative review and new concept. *British Journal of Sports Medicine*, *50*(21), 1309–1314.

Bizzini, M., & Dvorak, J. (2015). FIFA 11+: An effective programme to prevent football injuries in various player groups worldwide—a narrative review. *British Journal of Sports Medicine*, *49*(9), 577–579.

Bizzini, M., Junge, A., & Dvorak, J. (2013). Implementation of the FIFA 11+ football warm up program: How to approach and convince the football associations to invest in prevention. *British Journal of Sports Medicine*, *47*(12), 803–806.

Bleakley, C., & MacAuley, D. (2002). The quality of research in sports journals. *British Journal of Sports Medicine*, *36*(2), 124–125.

Bompa, T., & Buzzichelli, C. (2015). *Periodization training for sports* (3rd ed.). Human Kinetics.

Braun, B., & Miller, B. (2008). Introduction to sports nutrition: Energy metabolism and exercise. In I. W. Judy Driskell (Ed.), *Sports nutrition: Energy metabolism and exercise* (pp. 1–24). CRC Press.

Bricca, A., Juhl, C. B., Bizzini, M., Andersen, T. E., & Thorborg, K. (2018). There are more football injury prevention reviews than randomised controlled trials. time for more RCT action! *British Journal of Sports Medicine*, *52*(22), 1477–1478.

Broadbent, A. (2011). Inferring causation in epidemiology: Mechanisms, black boxes, and contrasts. In P. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 45–68). Oxford University Press.

Buchheit, M., Eirale, C., Simpson, B. M., & Lacome, M. (2019). Injury rate and prevention in elite football: Let us first search within our own hearts. *British Journal of Sports Medicine*, *53*(21), 1327–1328.

Burgess, E., Hassmén, P., & Pumpa, K. L. (2017). Determinants of adherence to lifestyle intervention in adults with obesity: A systematic review. *Clinical Obesity*, *7*(3), 123–135.

Burke, L. (2008). Caffeine and sports performance. *Applied Physiology, Nutrition, and Metabolism*, *33*(6), 1319–1334.

Burke, L. M., & Hawley, J. A. (2018). Swifter, higher, stronger: What's on the menu? *Science*, *362*(6416), 781–787.

Campaner, R. (2011). Understanding mechanisms in the health sciences. *Theoretical Medicine and Bioethics*, *32*(1), 5–17.

Cancer Research UK. (2016). *Being obese or overweight could cause 7.6 million cases of disease by 2035* [Accessed 3 July 2019]. https://www.cancerresearchuk.org/about-us/cancer-news/press-release/2016-06-20-being-obese-or-overweight-could-cause-76-million-cases-of-disease-by-2035?dm_i=21A8,49XJN,MJLS87,FRTM1,1

Carpenter, T. (2012). *Uneasy bedfellows: Amateurism and coaching traditions in twentieth century British sport* (Doctoral dissertation). Manchester Metropolitan University.

Cartwright, N., & Munro, E. (2010). The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice*, *16*(2), 260–266.

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, *2*(1), 11–20.

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies*, *147*, 59–70.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Cerulli, G., Benoit, D., Caraffa, A., & Pontrggia, F. (2001). Proprioceptive training and prevention of anterior cruciate ligament injuries in soccer. *Journal of Orthopaedic & Sports Physical Therapy*, *31*(11), 655–660.

Chalmers, D. (2002). Injury prevention in sport: Not yet part of the game? *BMJ*, *8*(supplement IV), 22–25.

Chang, C., Liu, W., Zhao, X., Li, S., & Yu, C. (2008). Effect of supervised exercise intervention on metabolic risk factors and physical fitness in chinese obese children in early puberty. *Obesity Reviews*, *9*(s1), 135–141.

Chanutin, A. et al. (1926). The fate of creatine when administered to man. *Journal of Biological Chemistry*, *67*(1), 29–41.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2013). The evidence that evidence-based medicine omits. *Preventive Medicine*, *57*(6), 745–747.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi*, *33*(2), 339–360.

Clarke, B., & Russo, F. (2017). Mechanisms and biomedicine. In P. Illari & S. Glennan (Eds.), *The routledge handbook of mechanisms and mechanical philosophy* (pp. 319–331). Routledge.

Cooke, R., & Jones, A. (2017). Recruiting adult participants to physical activity intervention studies using sport: A systematic review. *BMJ Open Sport & Exercise Medicine*, *3*(1), 1–9.

Cooper, S., & Nevill, A. (2005). *Do statistical methods replace reasoning in exercise science research?* (M. McNamee, Ed.). Routledge.

Craighead, L., & Blum, M. (1989). Supervised exercise in behavioral treatment for moderate obesity. *Behavior Therapy*, *20*(1), 49–59.

Crim, M. C., Calloway, D., & Margen, S. (1975). Creatine metabolism in men: Urinary creatine and creatinine excretions with creatine feeding. *The Journal of Nutrition*, *105*(4), 428–438.

Croisier, J.-L. (2004). Factors associated with recurrent hamstring injuries. *Sports Medicine*, *34*(10), 681–695.

Croisier, J.-L., Ganteaume, S., Binet, J., Genty, M., & Ferret, J.-M. (2008). Strength imbalances and prevention of hamstring injury in professional soccer players: A prospective study. *The American Journal of Sports Medicine*, *36*(8), 1469–1475.

Culbertson, J. Y., Kreider, R. B., Greenwood, M., & Cooke, M. (2010). Effects of beta-alanine on muscle carnosine and exercise performance: A review of the current literature. *Nutrients*, *2*(1), 75–98.

Dalle, R., Calugi, S., Centis, E., El Ghoch, M., & Marchesini, G. (2010). Cognitive-behavioral strategies to increase the adherence to exercise in the management of obesity. *Journal of Obesity*, *2011*, 1–11.

Daussin, F. N., Zoll, J., Dufour, S. P., Ponsot, E., Lonsdorfer-Wolf, E., Doutreleau, S., Mettauer, B., Piquard, F., Geny, B., & Richard, R. (2008). Effect of interval versus continuous training on cardiorespiratory and mitochondrial functions: Relationship to aerobic performance improvements in sedentary subjects. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *295*(1), R264–R272.

Day, D. (2011). Craft coaching and the 'discerning eye' of the coach. *International Journal of Sports Science & Coaching*, *6*(1), 179–195.

Department of Health and Social Care. (2011a). Equality analysis: A call to action on obesity in england.

Department of Health and Social Care. (2011b). Healthy lives, healthy people: A call to action on obesity in england.

Detmer, E., Fryback, G., & Gassner, K. (1978). Heuristics and biases in medical decision-making. *Academic Medicine*, *53*(8), 682–683.

Deurenberg, P., Andreoli, A., Borg, P., Kukkonen-Harjula, K., De Lorenzo, A., Van Marken Lichtenbelt, W. D., Testolin, G., Vigano, R., & Vollaard, N. (2001). The validity of predicted body fat percentage from body mass index and from impedance in samples of five european populations. *European Journal of Clinical Nutrition*, *55*(11), 973–979.

Deurenberg, P., & Yap, M. (1999). The assessment of obesity: Methods for measuring body fat and global prevalence of obesity. *Best*

Practice & Research Clinical Endocrinology & Metabolism, *13*(1), 1–11.

Doherty, M., & Smith, P. (2004). Effects of caffeine ingestion on exercise testing: A meta-analysis. *International Journal of Sport Nutrition and Exercise Metabolism*, *14*(6), 626–646.

Donnelly, J., Blair, S., Jakicic, J., Manore, M., Rankin, J., & Smith. (2009). American college of sports medicine position stand. appropriate physical activity intervention strategies for weight loss and prevention of weight regain for adults. *Medicine & Science in Sports & Exercise*, *41*(2), 459–471.

Donnelly, J., Jacobsen, D., Hill, J., Potteiger, J., Sullivan, D., Johnson, S., Heelan, K., Hise, M., Fennessey, P., & Sonko, B. (2003). Effects of a 16-month randomized controlled exercise trial on body weight and composition in young, overweight men and women: The midwest exercise trial. *Archives of Internal Medicine*, *163*(11), 1343–1350.

Donnelly, J., Smith, Dunn, L., Mayo, M., Jacobsen, D., Stewart, E., Gibson, C., & Sullivan, D. (2007). Comparison of a phone vs clinic approach to achieve 10% weight loss. *International Journal of Obesity*, *31*(8), 1270–1276.

dos Santos, W. D. N., Gentil, P., de Araújo Ribeiro, A. L., Vieira, C. A., & Martins, W. R. (2018). Effects of variable resistance training on maximal strength: A meta-analysis. *The Journal of Strength & Conditioning Research*, *32*(11), e52–e55.

Dowe, P. (2000). *Physical causation*. Cambridge University Press.

Ebell, M., Siwek, J., Weiss, B., Woolf, S., Susman, J., Ewigman, B., & Bowman, M. (2004). Strength of recommendation taxonomy (SORT): A patient-centered approach to grading evidence in the medical literature. *The Journal of the American Board of Family Practice*, *17*(1), 59–67.

Ekblom, B. (1996). Effects of creatine supplementation on performance. *American Journal of Sports Medicine*, *24*(6), S38–S39.

Emilien, G., van Meurs, W., & Maloteaux, J.-M. (2000). The dose-response relationship in phase I clinical trials and beyond: Use, meaning, and assessment. *Pharmacology & Therapeutics*, *88*(1), 33–58.

Ernst, E., & Pittler, M. (2006). Efficacy or effectiveness? *Journal of Internal Medicine*, *260*(5), 488–490.

Evans, C. R., & Dion, K. L. (2012). Group cohesion and performance: A meta-analysis. *Small Group Research*, *43*(6), 690–701.

Evans, D. (2003). *Placebo: The belief effect*. HarperCollins London.

Farrance, C., Tsofliou, F., & Clark, C. (2016). Adherence to community based group exercise interventions for older people: A mixed-methods systematic review. *Preventive Medicine*, *87*, 155–166.

FIFA. (2007). Fifa big count 2006: 270 million people active in football. *FIFA Communications Division, Information Services*, *31*, 1–12.

Finch, C. (2006). A new framework for research leading to sports injury prevention. *Journal of Science and Medicine in Sport*, *9*(1), 3–9.

Finch, C. (2011). Implementation and dissemination research: The time has come! *British Journal of Sports Medicine*, *45*(10), 763–764.

Flatt, J.-P. (2007). Differences in basal energy expenditure and obesity. *Obesity*, *15*(11), 2546–2548.

Fogelholm, M., & Kukkonen-Harjula, K. (2000). Does physical activity prevent weight gain – a systematic review. *Obesity Reviews*, *1*(2), 95–111.

Frederic, M. (2017). *Fundamentals of anatomy & physiology, global edition*. Pearson Education Limited.

Fu, F., & Stone, D. (2001). *Sports injuries : Mechanisms, prevention, treatment* (2nd ed). Lippincott Williams & Wilkins.

Fung, Y. (2015). The application of biomechanics to the understanding of injury and healing. In A. Nahum & J. Melvin (Eds.), *Accidental injury: Biomechanics and prevention* (pp. 1–11). Springer-Verlag New York.

Gabler, N., Duan, N., Vohra, S., & Kravitz, R. (2011). N-of-1 trials in the medical literature: A systematic review. *Medical Care*, *49*(8), 761–768.

Gallagher, D., Visser, M., Sepúlveda, D., Pierson, R. N., Harris, T., & Heymsfield, S. B. (1996). How Useful Is Body Mass Index for Comparison of Body Fatness across Age, Sex, and Ethnic Groups? *American Journal of Epidemiology*, *143*(3), 228–239.

Gallin, J. I., Alling, D. W., Malech, H. L., Wesley, R., Koziol, D., Marciano, B., Eisenstein, E. M., Turner, M. L., DeCarlo, E. S., Starling, J. M., et al. (2003). Itraconazole to prevent fungal infections in chronic granulomatous disease. *New England Journal of Medicine*, *348*(24), 2416–2422.

Galpin, A. J., Malyszek, K. K., Davis, K. A., Record, S. M., Brown, L. E., Coburn, J. W., Harmon, R. A., Steele, J. M., & Manolovitz, A. D. (2015). Acute effects of elastic bands on kinetic characteristics during the deadlift at moderate and heavy loads. *The Journal of Strength & Conditioning Research*, *29*(12), 3271–3278.

Garn, S. M., Leonard, W. R., & Hawthorne, V. M. (1988). Three limitations of the body mass index. *American Journal of Clinical Nutrition*, *48*(3), 691–692.

Garrow, J., & Summerbell, C. (1995). Meta-analysis: Effect of exercise, with or without dieting, on the body composition of overweight subjects. *European Journal of Clinical Nutrition*, *49*(1), 1–10.

Gatterer, H., Ruedl, G., Faulhaber, M., Regele, M., & Burtscher, M. (2012). Effects of the performance level and the FIFA "11" injury prevention program on the injury rate in italian male amateur soccer players. *Journal of Sports Medicine and Physical Fitness*, *52*(1), 80–84.

Gibson, O. R., James, C. A., Mee, J. A., Willmott, A. G., Turner, G., Hayes, M., & Maxwell, N. S. (2020). Heat alleviation strategies for athletic performance: A review and practitioner guidelines. *Temperature*, *7*(1), 3–36.

Gilbert, W., & Trudel, P. (2004). Analysis of coaching science research published from 1970–2001. *Research quarterly for exercise and sport*, *75*(4), 388–399.

Gillies, D. (2019). *Causality, probability, and medicine.* Routledge.

Gillies, D. (2005). Hempelian and Kuhnian approaches in the philosophy of medicine: the Semmelweis case. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(1), 159–181.

Gillies, D. (2017a). Evidence of mechanism in the evaluation of streptomycin and thalidomide. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *66*, 55–62.

Gillies, D. (2017b). Mechanisms in Medicine. *Axiomathes*, *27*(6), 621–634.

Giraldez-Costas, V., Gonzalez-Garcia, J., Lara, B., Del Coso, J., Wilk, M., & Salinero, J. J. (2020). Caffeine increases muscle performance during a bench press training session. *Journal of Human Kinetics*, *74*(1), 185–193.

Glaister, M., Howatson, G., Abraham, C. S., Lockey, R. A., Goodwin, J. E., Foley, P., & McInnes, G. (2008). Caffeine supplementation and multiple sprint running performance. *Medicine & Science in Sports & Exercise*, *40*(10), 1835–1840.

Good, I. (1959). A theory of causality. *The British Journal for the Philosophy of Science*, *9*(36), 307–310.

GRADE Working Group. (2004). Grading quality of evidence and strength of recommendations. *BMJ*, *328*(7454), 1490.

Greener, J., Douglas, F., & van Teijlingen, E. (2010). More of the same? Conflicting perspectives of obesity causation and intervention amongst overweight people, health professionals and policy makers. *Social Science and Medicine*, *70*(7), 1042–1049.

Greenland, S., Pearl, J., & Robins, J. (1999). Causal Diagrams for Epidemiologic Research. *Epidemiology*, *10*(1), 37–48.

Grimes, D. R. (2012). Proposed mechanisms for homeopathy are physically impossible. *Focus on Alternative and Complementary Therapies*, *17*(3), 149–155.

Grünbaum, A. (1981). The placebo concept. *Behavioural Research and Therapy*, *19*(2), 157–167.

Grünbaum, A. (1986). The placebo concept in medicine and psychiatry. *Psychological medicine*, *16*(1), 19–38.

Guyatt, G., Cairns, J., Churchill, D., Cook, D., Haynes, B., Hirsh, J., Irvine, J., Levine, M., Levine, M., et al. (1992). Evidence-based medicine: A new approach to teaching the practice of medicine. *Jama*, *268*(17), 2420–2425.

Guyatt, G., Meade, M. O., Jaeschke, R. Z., Cook, D. J., & Haynes, R. B. (2000). Practitioners of evidence based care: Not all clinicians need to appraise evidence from scratch but all need some skills. *BMJ*, *320*, 954–955.

Guyatt, G., Oxman, A., Sultan, S., Glasziou, P., Akl, E., Alonso-Coello, P., Atkins, D., Kunz, R., Brozek, J., Montori, V., et al. (2011). Grade guidelines: 9. rating up the quality of evidence. *Journal of Clinical Epidemiology*, *64*(12), 1311–1316.

Hagobian, T. A., & Evero, N. (2013). Exercise and weight loss: What is the evidence of sex differences? *Current Obesity Reports*, *2*(1), 86–92.

Hammes, D., Funten, K., Kaiser, S., Frisen, E., Bizzini, M., & Meyer, T. (2015). Injury prevention in male veteran football players - a randmised controlled tiral using "FIFA 11+". *Journal of Sports Sciences*, *33*(9), 873–881.

Hanson, D., Allegrante, J., Sleet, D., & Finch, C. (2014). Research alone is not sufficient to prevent sports injury. *British Journal of Sports Medicine*, *48*(8), 682–684.

Hanson, D., Finch, C., Allegrante, J., & Sleet, D. (2012). Closing the gap between injury prevention research and community safety promotion practice: Revisiting the public health model. *Public Health Reports*, *127*(2), 147–155.

Harrell, K. (2019). *Anatomy.* Wolters Kluwer.

Harvard School of Public Health. (n.d.). *Obesity prevention source: Why use BMI?* https://www.hsph.harvard.edu/obesity-prevention-source/obesity-definition/obesity-definition-full-story/ Accessed 23 July 19

Heiderscheit, B. C., Sherry, M. A., Silder, A., Chumanov, E. S., & Thelen, D. G. (2010). Hamstring strain injuries: Recommendations for diagnosis, rehabilitation, and injury prevention. *Journal of Orthopaedic & Sports Physical Therapy*, *40*(2), 67–81.

Heidt, R., Sweeterman, L., Carlonas, R., Traub, J., & Tekulve, F. (2000). Avoidance of soccer injuries with preseason conditioning. *The American Journal of Sports Medicine*, *28*(5), 659–662.

Herzog, W. (2017). Fairness in olympic sports: How can we control the increasing complexity of doping use in high performance sports? *Journal of Sport and Health Science*, *6*(1), 47.

Hill, A. B. (2015). The environment and disease: Association or causation? - reprint. *Journal of the Royal Society of Medicine*, *108*(1), 32–37.

Hodgson, A. B., Randell, R. K., & Jeukendrup, A. E. (2013). The metabolic and performance effects of caffeine compared to coffee during endurance exercise. *PloS One*, *8*(4), e59561.

Holman, B. (2015). Why most sugar pills are not placebos. *Philosophy of Science*, *82*(5), 1330–1343.

Holowchak, M. (2002). Ergogenic aids and the limits of human performance in sport: Ethical issues, aesthetic considerations. *Journal of the Philosophy of Sport*, *29*(1), 75–86.

Hootman, J. (2007). *Evidence-based sports medicine* (D. MacAuley & T. Best, Eds.). Wiley Online Library.

Howick, J. (2011a). Exposing the vanities—and a qualified defense of mechanistic reasoning in health care decision making. *Philosophy of Science*, *78*(5), 926–940.

Howick, J. (2011b). *The philosophy of evidence-based medicine.* John Wiley & Sons.

Howick, J. (2009). Questioning the methodologic superiority of 'placebo' over 'active' controlled trials. *The American Journal of Bioethics*, *9*(9), 34–48.

Howick, J., Chalmers, I., Glasziou, P., Greenhalgh, T., Heneghan, C., Liberati, A., Moschetti, I., Phillips, B., & Thornton, H. (2011). The 2011 oxford CEBM evidence levels of evidence (introductory document). *Oxford Center for Evidence Based Medicine*.

Hrysomallis, C. (2007). Relationship between balance ability, training and sports injury risk. *Sports Medicine*, *37*(6), 547–556.

Huang, J.-H., Su, Q.-M., Yang, J., Lv, Y.-H., He, Y.-C., Chen, J.-C., Xu, L., Wang, K., & Zheng, Q.-S. (2015). Sample sizes in dosage investigational clinical trials: A systematic evaluation. *Drug Design, Development and Therapy*, *9*, 305–312.

Hübscher, M., Zech, A., Pfeifer, K., Hänsel, F., Vogt, L., & Banzer, W. (2010). Neuromuscular training for sports injury prevention: A systematic review. *Medicine & Science in Sports & Exercise*, *42*(3), 413–421.

IARC. (1999). *Some chemicals that cause tumours of the kidney or urinary bladder in rodents and some other substances.* IARC Monographs on the Evaluation of Carcinogenic Risks to Humans Volume 73.

IARC. (2019). *Preamble to the IARC monographs on the identification of carcinogenic hazards to humans.*

Illari, P. (2011). Mechanistic evidence: Disambiguating the Russo-Williamson Thesis. *International Studies in the Philosophy of Science*, *25*(2), 139–57.

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119–135.

Ivarsson, A., & Andersen, M. B. (2016). What counts as "evidence" in evidence-based practice? searching for some fire behind all the smoke. *Journal of Sport Psychology in Action*, *7*(1), 11–22.

Jakicic, J., & Otto, A. (2006). Treatment and Prevention of Obesity: What is the Role of Exercise? *Nutrition Reviews*, *64*(supplement 1), S57–S61.

Joffe, M., Gambhir, M., Chadeau-Hyam, M., & Vineis, P. (2012). Causal diagrams in systems epidemiology. *Emerging Themes in Epidemiology*, *9*(1), 1–18.

Jukola, S. (2019). Casuistic reasoning, standards of evidence, and expertise on elite athletes' nutrition. *Philosophies*, *4*(2), 1–11.

Kahneman, D., & Tversky, A. (2000). *Choices, values, and frames.* Cambridge University Press.

Karamanou, M., Panayiotakopoulos, G., Tsoucalas, G., Kousoulis, A. A., & Androutsos, G. (2012). From miasmas to germs: A historical approach to theories of infectious disease transmission. *Le Infezioni in Medicina*, *20*(1), 58–62.

Keller, C. S., Noyes, F. R., & Buncher, C. R. (1987). The medical aspects of soccer injury epidemiology. *The American Journal of Sports Medicine*, *15*(3), 230–237.

Khan, M., Evaniew, N., Gichuru, M., Habib, A., Ayeni, O., Bedi, A., Walsh, M., Devereaux, P., & Bhandari, M. (2017). The fragility of statistically significant findings from randomized trials in sports surgery: A systematic survey. *The American Journal of Sports Medicine*, *45*(9), 2164–2170.

Khaodhiar, L., McCowen, K. C., & Blackburn, G. L. (1999). Obesity and its comorbid conditions. *Clinical Cornerstone*, *2*(3), 17–31.

Kinugasa, T. (2013). The application of single-case research designs to study elite athletes' conditioning: An update. *Journal of Applied Sport Psychology*, *25*(1), 157–166.

Kinugasa, T., Cerin, E., & Hooper, S. (2004). Single-subject research designs and data analyses for assessing elite athletes' conditioning. *Sports Medicine*, *34*(15), 1035–1050.

Klügl, M., Shrier, I., McBain, K., Shultz, R., Meeuwisse, W., Garza, D., & Matheson, G. (2010). The prevention of sport injury: An analysis of 12,000 published manuscripts. *Clinical Journal of Sport Medicine*, *20*(6), 407–412.

Knudson, D., Elliott, B., & Hamill, J. (2014). Proposing application of results in sport and exercise research reports. *Sports Biomechanics*, *13*(3), 195–203.

Knudson, D. (2005). Evidence-based practice in kinesiology: The theory to practice gap revisited. *The Physical Educator*, *62*(4), 212–221.

Knudson, D. (2009). Significant and meaningful effects in sports biomechanics research. *Sports Biomechanics*, *8*(1), 96–104.

Knudson, D., Elliott, B., & Ackland, T. (2012). Citation of evidence for research and application in kinesiology. *Kinesiology Review*, *1*(2), 129–136.

Kraemer, W., Duncan, N., & Volek, J. (1998). Resistance training and elite athletes: Adaptations and program considerations. *Journal of Orthopaedic and Sports Physical Training*, *28*(2), 110–119.

Kreider, R. B. (2003). Effects of creatine supplementation on performance and training adaptations. *Molecular and Cellular Biochemistry*, *244*(1-2), 89–94.

Kuikman, M. A., Mountjoy, M., Stellingwerff, T., & Burr, J. F. (2021). A review of nonpharmacological strategies in the treatment of relative energy deficiency in sport. *International Journal of Sport Nutrition and Exercise Metabolism*, *31*(3), 268–275.

Kwasnicka, D., Dombrowski, S. U., White, M., & Sniehotta, F. F. (2017). N-of-1 study of weight loss maintenance assessing predictors of physical activity, adherence to weight loss plan and weight change. *Psychology & Health*, *32*(6), 686–708.

Kwasnicka, D., & Naughton, F. (2020). N-of-1 methods: A practical guide to exploring trajectories of behaviour change and designing precision behaviour change interventions. *Psychology of Sport and Exercise*, *47*, 1–9. https://doi.org//10.1016/j.psychsport.2019.101570

Lamarche, L., Gammage, K., & Gabriel, D. (2011). The effects of experimenter gender on state social physique anxiety and strength in a testing environment. *The Journal of Strength & Conditioning Research*, *25*(2), 533–538.

Lauby-Secretan, B., Loomis, D., Baan, R., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Grosse, Y., & Straif, K. (2016). Use of mechanistic data in the IARC evaluations of the carcinogenicity of polychlorinated biphenyls and related compounds. *Environmental Science and Pollution Research*, *23*(3), 2220–2229.

Lauersen, J. B., Bertelsen, D. M., & Andersen, L. B. (2014). The effectiveness of exercise interventions to prevent sports injuries: A systematic review and meta-analysis of randomised controlled trials. *British Journal of Sports Medicine*, *48*(11), 871–877.

Laufs, U., Filipiak, K. J., Gouni-Berthold, I., Catapano, A. L., Mandraffino, G., & Benlian, P. (2017). Practical aspects in the management of statin-associated muscle symptoms (SAMS). *Atherosclerosis Supplements*, *26*, 45–55.

Laursen, P. B., Shing, C. M., Peake, J. M., Coombes, J. S., & Jenkins, D. G. (2002). Interval training program optimization in highly trained endurance cyclists. *Medicine & Science in Sports & Exercise*, *34*(11), 1801–1807.

Leuridan, B., & Weber, E. (2011). The IARC and mechanistic evidence. In J. Williamson, P. Illari, & F. Russo (Eds.), *Causality in the sciences* (pp. 91–109). Oxford University Press Oxford.

Lilford, R., Thornton, J., & Braunholtz, D. (1995). Clinical trials and rare diseases: A way out of a conundrum. *BMJ*, *311*(7020), 1621–1625.

Littlejohn, C. (2012). *Justification and the truth-connection.* Cambridge University Press.

Lloyd, J., Creanor, S., Logan, S., Green, C., Dean, S., Hillsdon, M., Abraham, C., Tomlinson, R., Pearson, V., Taylor, R., et al. (2018). Effectiveness of the Healthy Lifestyles Programme (HeLP) to prevent obesity in UK primary-school children: a cluster randomised controlled trial. *The Lancet. Child & Adolescent Health*, *2*(1), 35–45.

MacAuley, D. (2000). Sport and exercise medicine: Building the foundations of a new discipline. *Journal of Science and Medicine in Sport*, *3*(3), 254–259.

MacAuley, D., & Best, T. (2007). *Evidence-based sports medicine*. Wiley Online Library.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Maddocks, M., Kerry, R., Turner, A., & Howick, J. (2016). Problematic placebos in physical therapy trials. *Journal of Evaluation in Clinical Practice*, *22*(4), 598–602.

Martin, G. L., Thompson, K., & Regehr, K. (2004). Studies using single-subject designs in sport psychology: 30 years of research. *The Behavior Analyst*, *27*(2), 263–280.

McArdle, W., Katch, F., & Katch, V. (2008). *Sports and exercise nutrition* (Third). Lippincott Williams & Wilkins.

McDonald, S., Quinn, F., Vieira, R., O'Brien, N., White, M., Johnston, D. W., & Sniehotta, F. F. (2017). The state of the art and future opportunities for using longitudinal n-of-1 methods in health behaviour research: A systematic literature overview. *Health Psychology Review*, *11*(4), 307–323.

McDonald, S., Vieira, R., Godfrey, A., O'Brien, N., White, M., & Sniehotta, F. F. (2017). Changes in physical activity during the retirement transition: A series of novel n-of-1 natural experiments. *International Journal of Behavioral Nutrition and Physical Activity*, *14*(167).

McFee, G. (2009). *Ethics, knowledge and truth in sports research: An epistemology of sport*. Routledge.

McIntosh, A. S. (2005). Risk compensation, motivation, injuries, and biomechanics in competitive sport. *British Journal of Sports Medicine*, *39*(1), 2–3.

McKeon, P., Medina, J., & Hertel, J. (2006). Hierarchy of research design in evidence-based sports medicine. *Athletic Therapy Today*, *1*, 42–45.

McNamee, M. (2004). *Philosophy and the sciences of exercise, health and sport: Critical perspectives on research methods*. Routledge.

Medina, J., McKeon, P., & Hertel, J. (2006). Rating the levels of evidence in sports-medicine research. *International Journal of Athletic Therapy and Training*, *11*(5), 38–41.

Medina-McKeon, J., & McKeon, P. (2009). Assessment of the quality of clinically relevant research. *International Journal of Athletic Therapy and Training*, *14*(3), 4–9.

Miller, F., & Brody, H. (2002). What makes placebo-controlled trials unethical? *The American Journal of Bioethics*, *2*(2), 3–9.

Mjølsnes, R., Arnason, A., østhagen, T., Raastad, T., & Bahr, R. (2004). A 10-week randomized trial comparing eccentric vs. concentric hamstring strength training in well-trained soccer players. *Scandinavian Journal of Medicine & Science in Sports*, *14*(5), 311–317.

Moerman, D. E. (1983). General medical effectiveness and human biology: Placebo effects in the treatment of ulcer disease. *Medical Anthropology Quarterly*, *14*(4), 3–16.

Mountjoy, M., Sundgot-Borgen, J., Burke, L., Carter, S., Constantini, N., Lebrun, C., Meyer, N., Sherman, R., Steffen, K., Budgett, R., et al. (2014). The IOC consensus statement: Beyond the female athlete triad—relative energy deficiency in sport (RED-S). *British Journal of Sports Medicine*, *48*(7), 491–497.

Mulimani, P. S. (2017). Evidence-based practice and the evidence pyramid: A 21st century orthodontic odyssey. *American Journal of Orthodontics and Dentofacial Orthopedics*, *151*(1), 1–8.

National Health Service. (2019). *Treatment: Obesity*. https://www.nhs.uk/conditions/obesity/treatment/

National Strength & Conditioning Association. (2015). *Essentials of strength training and conditioning 4th edition* (4th ed.). Human Kinetics.

Nevill, A., Atkinson, G., & Hughes, M. (2008). Twenty-five years of sport performance research in the journal of sports sciences. *Journal of Sports Sciences*, *26*(4), 413–426.

Newsholme, E. A., & Beis, I. (1996). *Creatine and creatine phosphate: Scientific and clinical perspectives*. Academic Press.

Ney, P., Collins, C., & Spensor, C. (1986). Double blind: Double talk or are there ways to do better research. *Medical Hypotheses*, *21*(2), 119–126.

NHS. (2019). https://publichealthmatters.blog.gov.uk/2019/01/08/the-nhs-long-term-plan-10-key-public-health-points/

NICE. (2006). The guidelines manual. *London: National Institute for Health and Clinical Excellence*.

Nieuwlaat, R., Wilczynski, N., Navarro, T., Hobson, N., Jeffery, R., Keepanasseril, A., Agoritsas, T., Mistry, N., Iorio, A., Jack, S., et al. (2014). Interventions for enhancing medication adherence. *Cochrane Database of Systematic Reviews*, (11), 1–4.

Noakes, T. (2004). Can we trust rehydration research? In M. McNamee (Ed.), *Philosophy and the sciences of exercise, health and sport: Critical perspectives on research methods* (pp. 137–159). Routledge.

Noakes, T. (1991). *Lore of running*. Leisure Presss.

OCEBM Levels of Evidence Working Group. (2011). *The Oxford 2011 levels of evidence*. http://www.cebm.net/index.aspx?o=5653

OpenStax. (n.d.). *Anatomy and physiology 24.1: Overview of metabolic reactions* [date accessed = 26/6/19]. https://openstax.org/books/anatomy-and-physiology/pages/24-1-overview-of-metabolic-reactions

Ordovas, J. M., Ferguson, L. R., Tai, E. S., & Mathers, J. C. (2018). Personalised nutrition and health. *BMJ*, *361*, 1–7.

Orth, W., Madan, A., Taddeucci, R., Coday, M., & Tichansky, D. (2008). Support group meeting attendance is associated with better weight loss. *Obesity Surgery*, *18*(4), 391–394.

Osterberg, L., & Blaschke, T. (2005). Adherence to medication. *New England Journal of Medicine*, *353*(5), 487–497.

Osuka, Y., Jung, S., Kim, T., Okubo, Y., Kim, E., & Tanaka, K. (2017). Does attending an exercise class with a spouse improve long-term exercise adherence among people aged 65 years and older: A 6-month prospective follow-up study. *BMC Geriatrics*, *17*(170), 1–9.

Oxford Shoulder and Elbow Clinic. (2004). *Shoulder impingement* [date accessed: 4/10/21]. https://www.ouh.nhs.uk/shoulderandelbow/information/documents/A4ShoulderImpingeAppendix5.pdf

Paillard, T., Neo, F., Riviere, T., Marion, V., Montoya, R., & Dupui, P. (2006). Postural performance and strategy in the unipedal stance of soccer players at different levels of competition. *Journal of Athletic Training*, *41*(2), 172–176.

Paillard, T., & Noé, F. (2006). Effect of expertise and visual contribution on postural control in soccer. *Scandinavian Journal of Medicine & Science in Sports*, *16*(5), 345–348.

Papineau, D. (1994). The virtues of randomization. *The British journal for the philosophy of science*, *45*(2), 437–450.

Parkkinen, V.-P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M. P., Norell, C., Russo, F., Shaw, B., & Williamson, J. (2018). *Evaluating evidence of mechanisms in medicine: Principles and procedures*. Springer.

Parkkinen, V.-P., & Williamson, J. (2020). Extrapolating from model organisms in pharmacology. In A. LaCaze & B. Osimani (Eds.), *Uncertainty in pharmacology: Epistemology, methods, and decisions* (pp. 59–78). Springer International Publishing.

Pasman, W., Van Baak, M., Jeukendrup, A., & De Haan, A. (1995). The effect of different dosages of caffeine on endurance performance time. *International Journal of Sports Medicine*, *16*(4), 225–230.

Patsopoulos, N. A. (2011). A pragmatic view on pragmatic trials. *Dialogues in Clinical Neuroscience*, *13*(2), 217–224.

Pedro-Botet, J., Climent, E., & Benaiges, D. (2019). Muscle and statins: From toxicity to the nocebo effect. *Expert Opinion on Drug Safety*, *18*(7), 573–579.

Perri, M., Nezu, A., McKelvey, W., Shermer, R., Renjilian, D., & Viegener, B. (2001). Relapse prevention training and problem-solving therapy in the long-term management of obesity. *Journal of Consulting and Clinical Psychology*, *69*(4), 722–726.

Petersen, J., & Hölmich, P. (2005). Evidence based prevention of hamstring injuries in sport. *British Journal of Sports Medicine*, *39*(6), 319–323.

Petersen, J., Thorborg, K., Nielsen, M. B., Budtz-Jørgensen, E., & Hölmich, P. (2011). Preventive effect of eccentric training on acute hamstring injuries in men's soccer: A cluster-randomized controlled trial. *The American Journal of Sports Medicine*, *39*(11), 2296–2303.

Petridou, A., Siopi, A., & Mougios, V. (2019). Exercise in the management of obesity. *Metabolism*, *92*, 163–169.

Philippe, P., & Mansi, O. (1998). Nonlinearity in the epidemiology of complex health and disease processes. *Theoretical Medicine and Bioethics*, *19*(6), 591–607.

Phillips, L. (2000). Sports injury incidence. *British Journal of Sports Medicine*, *34*(2), 133–136.

Pike, J. (2018). Therapeutic use exemptions and the doctrine of double effect. *Journal of the Philosophy of Sport*, *45*(1), 68–82.

Pontzer, H., Yamada, Y., Sagayama, H., Ainslie, P. N., Andersen, L. F., Anderson, L. J., Arab, L., Baddou, I., Bedu-Addo, K., Blaak, E. E., et al. (2021). Daily energy expenditure through the human life course. *Science*, *373*(6556), 808–812.

Pope, H. G., Wood, R. I., Rogol, A., Nyberg, F., Bowers, L., & Bhasin, S. (2014). Adverse health consequences of performance-enhancing drugs: An endocrine society scientific statement. *Endocrine Reviews*, *35*(3), 341–375.

Popper, K. R. (1972). The aim of science. *Objective knowledge.* Oxford University Press Oxford.

Prentice, W. (2014). *Athletic training: Principles of evidence-based clinical practice.* McGraw-Hill Education.

Public Health England. (n.d.). *Health matters: Obesity and the food environment* [Accessed 3 July 2019]. https://www.gov.uk/government/

publications/health-matters-obesity-and-the-food-environment/health-matters-obesity-and-the-food-environment--2

Pyne, D., Hopkins, W., & Martin, D. (2010). Inadequate sample sizes in studies of athletic performance at the 2012 ACSM annual meeting. *Sportscience*, *14*, 1–11.

Quarrie, K. L., Gianotti, S. M., Hopkins, W. G., & Hume, P. A. (2007). Effect of nationwide injury prevention programme on serious spinal injuries in new zealand rugby union: Ecological study. *BMJ*, *334*(7604), 1150–1153.

Reinchenbach, H. (1978). The principle of causality and the possibility of its empirical confirmation. *Hans Reichenbach selected writings 1909–1953* (pp. 345–371). Springer.

Renjilian, D., Perri, M., Nezu, A., McKelvey, W., Shermer, R., & Anton, S. (2001). Individual versus group therapy for obesity: Effects of matching participants to their treatment preferences. *Journal of Consulting and Clinical Psychology*, *69*(4), 717–721.

Ricciardi, F., Nazareth, I., & Petersen, I. (2019). General practitioners' adherence to prescribing guidelines for statins in the united kingdoms. *bioRxiv*, 625236. https://doi.org/https://doi.org/10.1101/625236

Richardson, L. (2020). Peloton as a facilitator of hope: Pathways to initiate and sustain behaviors that enhance well-being [Master of Applied Positive Psychology (MAPP)]. *Capstone Projects*.

Risberg, M., Mork, M., Jenssen, H., & Holm, I. (2001). Design and implementation of a nueromuscular training programme following anterior cruciate ligament reconstruction. *Journal of Orthopaedic & Sports Physical Therapy*, *31*(11), 620–631.

Ross, R., Pedwell, H., & Rissanen, J. (1995). Effects of energy restriction and exercise on skeletal muscle and adipose tissue in women as measured by magnetic resonance imaging. *The American Journal of Clinical Nutrition*, *61*(6), 1179–1185.

Rothman, K. J. (2008). BMI-related errors in the measurement of obesity. *International Journal of Obesity*, *32*(S3), S56–S59.

Ruberg, S. J. (1995). Dose response studies I. some design considerations. *Journal of Biopharmaceutical Statistics*, *5*(1), 1–14.

Russo, F. (2012). Public health policy, evidence, and causation: lessons from the studies on obesity. *Medicine, Health Care and Philosophy*, *15*, 141–151.

Russo, F. (2015). Causation and correlation in medical science: Theoretical problems. *Handbook of the philosophy of medicine* (pp. 839–850). Springer.

Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, *21*(2), 157–170.

Russo, F., & Williamson, J. (2011). Epistemic Causality and Evidence-Based Medicine. *History and Philosophy of the Life Sciences*, *33*(4), 563–582.

Russo, F. (2008). *Causality and causal modelling in the social sciences: Measuring variations*. Springer.

Sacher, P. M., Kolotourou, M., Chadwick, P. M., Cole, T. J., Lawson, M. S., Lucas, A., & Singhal, A. (2010). Randomized controlled trial of the MEND program: A family-based community intervention for childhood obesity. *Obesity*, *18*(S1), S62–S68.

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ*, *312*, 71–72.

Sainani, K. L., Lohse, K. R., Jones, P. R., & Vickers, A. (2019). Magnitude-based inference is not bayesian and is not a valid method of inference. *Scandinavian Journal of Medicine & Science in Sports*, *29*(9), 1428–1436.

Salmon, W. C. (1998). *Causality and explanation*. Oxford University Press.

Salomons, G. S., & Wyss, M. (2007). *Creatine and creatine kinase in health and disease: Subcellular biochemistry volume 46*. Springer Science & Business Media.

Sands, W., Cardinale, M., McNeal, J., Murray, S., Sole, C., Reed, J., Apostolopoulos, N., & Stone, M. (2019). Recommendations for measurement and management of an elite athlete. *Sports*, *7*(105), 1–17.

Saunders, B., Elliott-Sale, K., Artioli, G. G., Swinton, P. A., Dolan, E., Roschel, H., Sale, C., & Gualano, B. (2017). β-alanine supplementation to improve exercise capacity and performance: A systematic review and meta-analysis. *British Journal of Sports Medicine*, *51*(8), 658–669.

Sbrocco, T., Carter, M., Lewis, E., Vaughn, N., Kalupa, K., King, S., Suchday, S., Osborn, R., & Cintrón, J. (2005). Church-based obesity treatment for african-american women improves adherence. *Ethnicity & Disease*, *15*(2), 246–255.

Scanlan, A., & MacKay, M. (2001). Review of soccer injury prevention strategies. *Sports and recreation injury prevention strategies: Systematic review and best practices*. BC Injury Research; Prevention Unit, Children's Hospital of Eastern Ontario.

Schoenfeld, B. (2010). The mechanisms of muscle hypertrophy and their application to resistance training. *Journal of Strength and Conditioning Research*, *24*(10), 2857–2872.

Schoenfeld, B. (2020). https://twitter.com/BradSchoenfeld/status/1221876277277274115

Schoenfeld, B., Ogborn, D., & Krieger, J. (2017). Dose-response relationship between weekly resistance training volume and increases in muscle mass: A systematic review and meta-analysis. *Journal of Sports Sciences*, *35*(11), 1073–1082.

Schünemann, H., Brożek, J., Guyatt, G., & Oxman, A. (2013). *GRADE handbook for grading quality of evidence and strength of recommendations. updated october 2013*. The GRADE Working Group.

Schweizer, G., & Furley, P. (2016). Reproducible research in sport and exercise psychology: The role of sample sizes. *Psychology of Sport and Exercise*, *23*, 114–122.

Scott, L., Scott, D., Bedic, S., & Dowd, J. (1999). The effect of associative and dissociative strategies on rowing ergometer performance. *The Sports Psychologist*, *13*(1), 57–68.

Select Committee on Science and Technology. (2012a). *Sport and exercise science and medicine: Building on the olympic legacy to improve the nation's health.* Published by the Authority of the House of Lords. HL Paper 33, 1st Report of Session 2012-13.

Select Committee on Science and Technology. (2012b). *Sport and exercise science and medicine: Building on the olympic legacy to improve the nation's health.* Published by the Authority of the House of Lords. HL Paper 33, 1st Report of Session 2012-13.

Shapiro, A., & Morris, L. (1978). The placebo effect in medical and psychological therapies. In S. Garfield & B. A (Eds.), *Handbook of psychotherapy and behavioural change: An empirical analysis* (pp. 369–410). John Wiley & Sons.

Shapiro, A., Shapiro, E., & Harrington, A. (1999). Is it much ado about nothing. In A. Harrington (Ed.), *The placebo effect: An interdisciplinary exploration* (pp. 12–36). Harvard University Press.

Sheridan, A., Marchant, D., Williams, E., Jones, H., Hewitt, P., & Sparks, A. (2019). Presence of spotters improves bench press performance: A deception study. *The Journal of Strength & Conditioning Research*, *33*(7), 1755–1761.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Sinclair, C., & Geiger, J. (2000). Caffeine use in sports. a pharmacological review. *The Journal of Sports Medicine and Physical Fitness*, *40*(1), 71–79.

Smalley, K. J., Knerr, A. N., Kendrick, Z. V., Colliver, J. A., & Owen, O. E. (1990). Reassessment of body mass indices. *The American journal of clinical nutrition*, *52*(3), 405–408.

Smith, G. C., & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ*, *327*(7429), 1459–1461.

Sniehotta, F. F., Presseau, J., Hobbs, N., & Araújo-Soares, V. (2012). Testing self-regulation interventions to increase walking using factorial randomized n-of-1 trials. *Health Psychology*, *31*(6), 733–737.

Soares, I., & Carneiro, A. V. (2002). Intention-to-treat analysis in clinical trials: Principles and practical importance. *Revista portuguesa de cardiologia: orgao oficial da Sociedade Portuguesa de Cardiologia= Portuguese journal of cardiology: an official journal of the Portuguese Society of Cardiology*, *21*(10), 1191–1198.

Society for Transparency, Openness, and Replication in Kinesiology. (2020). https://storkinesiology.org/#

Soligard, T., Myklebust, G., Steffen, K., Holme, I., Silvers, H., Bizzini, M., Junge, A., Dvorak, J., Bahr, R., & Andersen, T. E. (2008). Comprehensive warm-up programme to prevent injuries in young female footballers: Cluster randomised controlled trial. *BMJ:Online First*, *337*, 1–9.

Solomon, M. (2015). *Making medical knowledge*. Oxford University Press, USA.

Soomro, N., Sanders, R., Hackett, D., Hubka, T., Ebrahimi, S., Freeston, J., & Cobley, S. (2015). The efficacy of injury prevention programs in adolescent team sports: A meta-analysis. *The American Journal of Sports Medicine*, *44*(9), 2415–2424.

Soria-Gila, M. A., Chirosa, I. J., Bautista, I. J., Baena, S., & Chirosa, L. J. (2015). Effects of variable resistance training on maximal strength: A meta-analysis. *The Journal of Strength & Conditioning Research*, *29*(11), 3260–3270.

Sothern, M. (2001). Exercise as a modality in the treatment of childhood obesity. *Pediatric Clinics of North America*, *48*(4), 995–1015.

Sothern, M., Loftin, M., Suskind, R., Wilson, J., Singley, C., Udall, J., & Blecker, U. (1997). The impact of level of obesity on physiologic function during rest, submaximal and maximal weight-bearing exercise in obese youth. *Obesity Research*, *5*(1), 35S.

Sport England. (2017). Review of the evidence on the outcomes of sport and physical activity: A rapid evidence review.

Steel, D. (2007). *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press, Oxford.

Stegenga, J. (2014). Down with the hierarchies. *Topoi*, *33*(2), 313–322.

Stegenga, J. (2018). *Medical nihilism*. Oxford University Press.

Steves, R., & Hootman, J. (2004). Evidence-based medicine: What is it and how does it apply to athletic training? *Journal of athletic training*, *39*(1), 83–87.

Stovitz, S. D., Verhagen, E., & Shrier, I. (2017). Misinterpretations of the 'p value': A brief primer for academic sports medicine. *British Journal of Sports Medicine*, *51*(16), 1176–1177.

Straus, S., Glasziou, P., Richardson, S., & Haynes, B. (2019). *Evidence-based medicine e-book: How to practice and teach EBM*. Elsevier Health Sciences.

Summerbell, C., Waters, E., Edmunds, L., Kelly, S., Brown, T., & Campbell, K. (2003). Interventions for preventing obesity in children [CD001971]. *Cochrane Database of Systematic Reviews*, *3*.

Suppes, P. (1973). A probabilistic theory of causality. *British Journal for the Philosophy of Science*, *24*(4), 409–410.

Szabo, A., & Griffiths, M. D. (2007). Exercise addiction in British sport science students. *International Journal of Mental Health and Addiction*, *5*(1), 25–28.

Tarnopolsky, A. (2008). Effect of caffeine on the neuromuscular system - potential as an ergogenic aid. *Applied Physiology, Nutrition, and Metabolism*, *33*(6), 1284–1289.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials:

The 15-item risk of bias in n-of-1 trials (RoBiNT) scale. *Neuropsychological Rehabilitation*, *23*(5), 619–638.

Teixeira, M. Z. (2011). Scientific evidence of the homeopathic epistemological model. *International Journal of High Dilution Research-ISSN 1982-6206*, *10*(34), 46–64.

Temple, R., & Ellenberg, S. (2000a). Placebo-controlled trials and active controlled trials in the evaluation of new treatments, part 1: Ethical and scientific issues. *Annals of Internal Medicine*, *133*, 455–463.

Temple, R., & Ellenberg, S. (2000b). Placebo-controlled trials and active controlled trials in the evaluation of new treatments, part 2: Practical issues and specific cases. *Annals of Internal Medicine*, *133*, 464–470.

The Association of UK Dieticians. (2018). *Policy statement uk government's childhood obesity strategy*. https://www.bda.uk.com/improvinghealth/healthprofessionals/policy_statements/policy_statement_-_uk_governments_childhood_obesity_strategy

the NNT. (2013). *Statins given for 5 years for heart disease prevention (with known heart disease)* [date accessed: 12/03/2020]. https://www.thennt.com/nnt/statins-for-heart-disease-prevention-with-known-heart-disease/

Thiel, C., Foster, C., Banzer, W., & De Koning, J. (2012). Pacing in olympic track races: Competitive tactics versus best performance strategy. *Journal of Sports Sciences*, *30*(11), 1107–1115.

Thompson, P. D., Panza, G., Zaleski, A., & Taylor, B. (2016). Statin-associated side effects. *Journal of the American College of Cardiology*, *67*(20), 2395–2410.

Thorborg, K., Krommes, K. K., Esteve, E., Clausen, M. B., Bartels, E. M., & Rathleff, M. S. (2017). Effect of specific exercise-based football injury prevention programmes on the overall injury rate in football: A systematic review and meta-analysis of the FIFA 11 and 11+ programmes. *British Journal of Sports Medicine*, *51*(7), 562–571.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Umer, M., Qadir, I., & Azam, M. (2012). Subacromial impingement syndrome. *Orthopedic Reviews*, *4*(2), 79–82.

Urbach, P. (1993). The value of randomization and control in clinical trials. *Statistics in Medicine*, *12*(15-16), 1421–1431.

US Food and Drug Administration. (2016). *Step 3: Clinical research* [date accessed: 14/11/19]. https://www.fda.gov/patients/drug-development-process/step-3-clinical-research

van Beijsterveldt, A. M. C., van de Port, I. G. L., Krist, M. R., Schmikli, S. L., Stubbe, J. H., Frederiks, J. E., & Backx, F. J. G. (2012). Effectiveness of an injury prevention programme for adult male amateur soccer players: A cluster-randomised controlled trial. *British Journal of Sports Medicine*, *46*(16), 1114–1118.

Wasan, A. D. (2014). Efficacy vs Effectiveness and Explanatory vs Pragmatic: Where Is the Balance Point in Pain Medicine Research? *Pain Medicine*, *15*(4), 539–540.

Waters, E., de Silva-Sanigosrki, A., & Hall, B. (2011). Interventions for preventing obesity in children [CD001871]. *Cochrane Database of Systematic Reviews*, *12*.

Welsh, A. H., & Knight, E. J. (2015). Magnitude-based inference: A statistical review. *Medicine and Science in Sports and Exercise*, *47*(4), 874–884.

Wilde, M., & Parkkinen, V.-P. (2019). Extrapolation and the Russo–Williamson thesis. *Synthese*, *196*(8), 3251–3262.

Williams, M. H., Kreider, R. B., & Branch, J. D. (1999). *Creatine: The power supplement.* Human kinetics.

Williamson, J. (2013). How can causal explanations explain? *Erkenntnis*, *78*, 257–275.

Williamson, J. (2019). Establishing causal claims in medicine. *International Studies in the Philosophy of Science*, *32*(1), 33–61.

Williamson, J. (2021). Establishing the teratogenicity of zika and evaluating causal criteria. *Synthese*, *198*(1326), 2505–2518.

Williamson, J. (2015). Deliberation, judgement and the nature of evidence. *Economics and Philosophy*, *31*(1), 27–65.

Wing, R. (1999). Physical activity in the treatment of the adulthood overweight and obesity: Current evidence and research issues. *Medicine and Science in Sports and Exercise*, *31*(11 Suppl), S547–S552.

Wong, P., & Hong, Y. (2005). Soccer injury in the lower extremities. *British Journal of Sports Medicine*, *39*(8), 473–482.

Woodruff, K. (2016). *Sports nutrition.* Momentum Press.

Wootton, D. (2007). *Bad medicine: Doctors doing harm since Hippocrates.* Oxford University Press.

World Health Organisation. (2015). https://www.who.int/dietphysicalactivity/childhood/en/

World Health Organisation. (2021). https://www.who.int/westernpacific/health-topics/obesity

World Health Organization. (2003). Adherence to long-term therapies: Evidence for action.

Worrall, J. (2002). What evidence in evidence-based medicine. *Philosophy of Science*, *69*(3).

Worrall, J. (2007). Why there's no cause to randomize. *The British Journal for the Philosophy of Science*, *58*(3), 451–488.

Yang, J., Tibbetts, A., Covassin, T., Cheng, G., Nayar, S., & Heiden, E. (2012). Epidemiology of overuse and acute injuries among competitive collegiate athletes. *Journal of Athletic Training*, *47*(2), 198–204.

Yusuf, S., Hawken, S., Ounpuu, S., Bautista, L., Franzosi, M. G., Commerford, P., Lang, C. C., Rumboldt, Z., Onen, C. L., Lisheng, L., et al. (2005). Obesity and the risk of myocardial infarction in 27 000 participants from 52 countries: A case-control study. *The Lancet*, *366*(9497), 1640–1649.

Zatsiorsky, V., & Kraemer, W. (2006). *Science and practice of strength training, second edition* (2nd ed.). Human Kinetics.