KTH ROYAL INSTITUTE
OF TECHNOLOGY

Doctoral Thesis in Philosophy

Philosophical Aspects of Evidence and Methodology in Medicine

JESPER JERKERT

Stockholm, Sweden 2021

Philosophical Aspects of Evidence and Methodology in Medicine

JESPER JERKERT

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Monday the 17th of May, 2021, at 9:00 a.m. (online)

Doctoral Thesis in Philosophy
KTH Royal Institute of Technology
Stockholm, Sweden 2021

TRITA-ABE-DLT-214
ISSN 1650-8831
ISBN 978-91-7873-838-0

Printed by: Universitetsservice US-AB, Sweden 2021

Contents

Abstract

Jesper Jerkert, *Philosophical Aspects of Evidence and Methodology in Medicine*, Doctoral thesis in Philosophy. *Theses in Philosophy from the Royal Institute of Technology* 65. Stockholm, 2021, 197 pp. With a summary in Swedish. ISBN 978-91-7873-838-0. TRITA-ABE-DLT-214.

The thesis consists of an introduction and five papers. The introduction gives a brief historical survey of empirical investigations into the effectiveness of medicinal interventions, as well as surveys of the concept of evidence and of the history and philosophy of experiments. The main ideas of the EBM (evidence-based medicine) movement are also presented.

Paper I: Concerns have been raised that clinical trials do not offer reliable evidence for some types of treatment, in particular for highly individualised treatments, for example traditional homeopathy. With respect to individualised treatments, it is argued that such concerns are unfounded. There are two minimal conditions related to the nature of the treatments that must be fulfilled for evaluability in a clinical trial, namely (1) the proper distinction of treatment groups and (2) the elimination of confounding variables or variations. These conditions do not preclude the testing of individualised medicine.

Paper II: Traditionally, mechanistic reasoning has been assigned a negligible role in the EBM literature. When discussed, mechanistic reasoning has almost exclusively been positive—both in an epistemic sense of claiming that *there is* a mechanistic chain and in a health-related sense of there being claimed *benefits* for the patient. Negative mechanistic reasoning has been neglected. I distinguish three main types of negative mechanistic reasoning and subsume them under a new definition. One of the three distinguished types, which is negative only in the health-related sense, has a corresponding positive counterpart, whereas the other two, which are epistemically negative, do not have such counterparts, at least not that are particularly interesting as evidence. Accounting for negative mechanistic reasoning in EBM is therefore partly different from accounting for positive mechanistic reasoning.

Paper III: Evidence hierarchies are lists of investigative strategies ordered with regard to the claimed strength of evidence. They have been used for a couple of decades within EBM, particularly for the assessment of evidence for treatment recommendations, but they remain controversial. An under-investigated question is what the order in the hierarchy means. Four interpretations of the order are distinguished and discussed. The two most credible ones are, in rough terms, "typically stronger" and "ideally stronger". The GRADE framework seems to be based on the "typically stronger" reading. Even if the interpretation of an evidence hierarchy were established, hierarchies appear to be rather unhelpful for the task of evidence aggregation. However, specifying the intended order relation may help sort out disagreements.

Paper IV: There are three main arguments for randomisation that connect inseparably to theoretical concepts: (1) Randomisation is useful for performing null hypothesis testing. (2) Randomisation is needed for plausible causal inferences from treatment to effect. (3) Randomisation is acceptable and computationally convenient in a Bayesian setting. A critical scrutiny of these arguments shows that (1) is acceptable in the context of clinical trials. As for (2), it is argued that randomisation only provides weak reasons for drawing causal inferences in the context of real (as opposed to theoretically ideal but unrealistic) clinical trials. Argument (3) is weak because it is controversial among Bayesians, and because formally Bayesian analyses of trial results are rarely asked for.

Paper V: Practical arguments for randomisation are arguments with no necessary connections to theoretical frameworks like null hypothesis testing or causal inferences. Four common practical arguments in the context of clinical trials are distinguished and assessed: (1) Randomisation contributes to allocation concealment. (2) Randomisation contributes to the baseline balance of treatment groups. (3) Randomisation decreases self-selection bias. (4) Randomisation removes allocation bias. Argument (1) is rejected. Arguments (3) and (4) are approved. Argument (2) is rejected if it is formulated so as to be independent from (3) and (4), but it is true that randomisation contributes to balance through the mechanisms mentioned in (3) and (4). It is judged that (4) may be the strongest single argument.

Thesis composition

This thesis consists of an introduction and the following five research papers:

I. Jesper Jerkert: "Why alternative medicine can be scientifically evaluated", in Massimo Pigliucci and Maarten Boudry (eds), *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*, Chicago: University of Chicago Press, 2013, 305–320.

II. Jesper Jerkert: "Negative mechanistic reasoning in medical intervention assessment", *Theoretical Medicine and Bioethics* 36(6), 2015, 425–437.

III. Jesper Jerkert: "On the meaning of medical evidence hierarchies", *Philosophy of Medicine* 2(1), 2021, 1–21.

IV. Jesper Jerkert: "Assessing the theoretical arguments for randomisation in clinical trials", manuscript.

V. Jesper Jerkert: "Assessing the practical arguments for randomisation in clinical trials", manuscript.

In the introduction, the research papers will be referenced with their Roman numerals as per above.

There is also a comprehensive summary in Swedish.

Acknowledgments

Sven Ove Hansson (professor at the Division of Philosophy, KTH) was my main supervisor until December 31, 2020, when he retired from this capacity *de jure*, though not necessarily *de facto*. He has provided comments on everything included in this thesis, as well as overall support and encouragement. Participants—too numerous to be listed—in the PhD student seminar at the Division of Philosophy, KTH, have supplied comments on texts included in this thesis. The seminars have been chaired mainly by John Cantwell (professor at the Division of Philosophy, my assistant supervisor until December 31, 2020, main supervisor from January 1, 2021) and Tor Sandqvist (associate professor at the Division of Philosophy, assistant supervisor), both of whom gave useful feedback for several included texts.

Portions of what is included in Paper III were presented at the conference *Issues in Medical Epistemology* in Cologne in December 2017, which led to new insights on the part of the author.

Till Grüne-Yanoff (professor at the Division of Philosophy, KTH) acted as the advance reviewer of the thesis and pointed out several vaguenesses and inconsistencies, for which I thank him.

I am grateful for all comments and suggestions that I have received. Of course, no one but me is responsible for the arguments and conclusions contained in this thesis.

Introduction

1. The questions

This thesis is about evidence and methodology in medical research. Naturally, the thesis does not answer all conceivable questions within its subject area. Rather, the included papers, numbered I through V, try to answer a limited number of more precise questions. Those questions are, for each of the five papers:

I. What conditions must be fulfilled for a medical treatment to be eligible for a scientific evaluation of its effectiveness?

II. What roles can be played by mechanistic reasoning in a scientific evaluation of medical treatment effectiveness?

III. What could the order in a so-called evidence hierarchy mean, and what does it reasonably mean in the context of evaluating the evidence for recommending the best medical treatments?

IV. Which theoretical arguments are tenable out of those that are regularly offered in favour of randomisation in clinical trials?

V. Which practical arguments are tenable in favour of randomisation in clinical trials?

We will return to these questions and, naturally, to their answers (skip to Section 5 for a summary). To understand their meaning and significance, and to put them in context, some background material will be presented in this introduction. Naturally, for many readers it will not be necessary to digest the introduction in order to understand the included papers, as the latter are supposed to be self-contained. Conversely, the introduction is self-contained, too, as it includes research paper summaries (Section 5).

Anyone who is not interested in details may therefore read the introduction only. This introduction, then, has a dual purpose: first, if it is read without the paper summaries, it sets the stage for the research papers I–V; secondly, if it is read with the paper summaries, it is a surrogate for the whole thesis.

Inquiries into the methodology of a particular scientific field can always be claimed to belong to the specific discipline rather than to philosophy. Of course, medical scientists—at least some of them—should be interested in answers to questions such as those above; they are certainly not the concern of philosophers only. But it ought be uncontroversial to claim that they are also philosophical. Each inquiry in this thesis takes an outside look at a practice—in our case a scientific practice—and investigates the rules that govern (or that should govern) the practice. Such a outside view or assessment of a topic is typical for philosophy.

Each inquiry tries to increase clarity in its subject, and the result of each inquiry is argumentative: the conclusions are drawn on the basis of arguments that have been discussed and defended, rather than on the basis of empirical data collection. Furthermore, all appended papers involve roughly the same methodological elements, in roughly the same order. First, some methodological claim (made at least by some people) is identified.[1] It is then suggested that the claim can be doubted, and/or is in need of clarification. The claim is therefore analysed (and/or clarified) and I argue for some specific view or conclusion with respect to it. In this analysis, where the main work is being done, general ideas about how to perform scientific research or how to reason with evidence may be invoked. Since the enterprise is supposed to be informative for real medical research, connections are made to what is actually done and is feasible within such research. Possible counterarguments are discussed, too.

Since the investigations in this thesis are strongly connected to medicine, the work could be claimed to belong under the "philosophy of medicine" heading. The main alternative heading would be the philosophy of science. Particularly Papers IV and V are strongly connected to medical research, and some results in those papers cannot straightforwardly be transported to other scientific disciplines. This fact speaks in favour of a philosophy of medicine categorisation. On the other hand, there are other results, particularly in Papers II (on mechanistic reasoning) and III (on evidence hierarchies), that I believe could be useful and informative in other scientific disciplines, as

[1]The concrete claims can be formulated as follows. *Paper I*: Alternative medicine cannot be put to test in clinical trials because it is too individualised. *Paper II*: Mechanistic reasoning should have very low evidential value in medicine. *Paper III*: Evidence hierarchies are useful as methodological guides in medicine. *Papers IV–V*: Randomisation is useful in clinical trials for a number of specified reasons.

well as in the general philosophy of science. Also, the concept of evidence is prominent in the present work, which is generally true of the philosophy of science, but less so of the philosophy of medicine.[2]

The philosophy of medicine seems to be devoted mainly (if not exclusively) to questions about the nature of medicine and of the most important concepts within its realm. In a recent book entitled *Philosophy of Medicine* (Broadbent, 2019), we find the questions discussed therein to be the following:

- What is medicine? (How to define and demarcate the subject area.)
- What is the goal of medicine? (Is it curing? Is it inquiring? Is it something else?)
- What is health? What is disease?
- What should we think of medical schools of thought such as evidence-based medicine (EBM), "medical nihilism", and alternative medicine?

I think it is uncontroversial that these questions belong to the philosophy of medicine. In each of the questions, the medical connection provides essential meaning, and the questions cannot so easily (if at all) be transported to some other area "the philosophy of X". I do not try to answer any of the listed questions in the present thesis (though the answers that are provided to *my* questions can be used to partly answer Broadbent's questions about evidence-based medicine and alternative medicine). In summary, I prefer to view this thesis as a work within the philosophy of science, but a philosophy of medicine categorisation is not totally implausible.[3]

Of course, there are a lot of important ethical questions raised in medical practice and research. Such questions are usually sorted not under the philosophy of medicine heading but under the label of bioethics, an established field in its own. However, since ethical questions are not important in the present thesis, there is no overview of bioethical topics in this introduction.

[2] Obviously, the latter fact cannot be the sole determinant of where the work belongs. In academic philosophy, the concept of evidence is perhaps featured most prominently in the philosophy of law.

[3] Since this thesis is so clearly connected to a specific non-philosophical discipline (medicine), it would also seem possible to claim that it constitutes "applied philosophy (of science)". I am unwilling to agree, however. First, "applied philosophy" seems to be used predominantly for *ethics* being applied in various scenarios and disciplines (Archard, 2017: 18), but the present thesis is not about ethics. Secondly, the philosophy of science ought not be concerned solely with theories and principles that are never connected to, or compared with, how science is actually practiced. Making such connections and comparisons should, in my opinion, count as a core activity in the philosophy of science proper, and therefore does not justify the "applied" label.

As already mentioned, this thesis revolves a lot around evidence. Evidence is a general term, used in science as well as in philosophy (and elsewhere). Section 2 provides a rather elementary overview of the concept of evidence. The thesis also revolves, to a substantial degree, around clinical trials—particularly randomised clinical trials (RCTs). A clinical trial is an experiment. Section 3 briefly discusses selected topics in the history and philosophy of experiments, with special subsections on the concepts of blinding and randomisation. Since the 1990's, medical evaluation research has been debated particularly in relation to the evidence-based medicine (EBM) movement (or school of thought). But of course, medical evaluation research has a history long before the advent of EBM. Section 4 provides, first, a historical sketch of medical intervention research and, secondly, a brief presentation of EBM, including some common criticisms. Section 5 contains summaries of the research papers, mainly intended for those that do not plan to read, or cannot electronically access, the research papers.

To a professional philosopher, much of the introduction is likely to contain familiar stuff, and will hence not be needed in order to understand the appended papers. This should be particularly true of Sections 2 and 3, whereas Section 4 (which is less philosophical) may be more useful as a precursor to the papers. To a medical researcher, the usefulness of the introductory sections may be reversed as compared to the philosopher.

2. Evidence

2.1. General characteristics

The word "evidence" is derived via Middle English and French from classical Latin *evidentia*, from *evidens*, which means clear or obvious (to the senses or to the mind). The *-videns* part of the word comes from the verb *videre* (to see). Etymologically, then, evidence is visible, or clear for everyone to see.

In some contexts, such as law, journalism, and historical investigations, evidence tends to be something concrete: a physical object that can be exhibited and touched. Philosophers, on the other hand, allow for evidence to be considerably more abstract. Not all philosophers agree on what evidence is at its most basic level, however (Kelly, 2016). Empiricists have thought of evidence as that which is perceived by the senses ("sense data"). Some Bayesians have thought of a person's evidence as those beliefs of which the person is convinced that they are true. And there are other suggestions as well.

In view of such suggestions, evidence can easily be a fundamental concept in most epistemological questions. For example, consider the challenge from philosophical scepticism: do we have knowledge about the external world, in the sense that we are able to conclusively reject the possibility that we are being consistently misled by our senses, possibly by some Cartesian demon? This can be recast as a question about the reliability of our evidence, if evidence is that which is perceived by the senses.

The most obvious role played by evidence—in philosophy and in science—is to be *something which justifies belief*.[4] This function of evidence is ubiquitous in science. Scientists often say that they "follow the evidence"; i.e., they claim to adjust what they believe according to the evidence they are aware of. But beliefs can be erroneous, of course, in science as well as in other settings. Typically, then, when a belief in science is erroneous, this is due to misleading evidence. Thus, what we *take as* evidence is not necessarily something that justifies a belief, but rather something that we *think* justifies a belief.

There are several historical examples where scientists have considered their evidence at hand as strongly supporting a particular hypothesis, but where subsequent research has produced a quite different conclusion. In 1883, German physicist Heinrich Hertz (1857–1894) exerted an external electrostatic force on cathode rays in an evacuated tube. He did not observe the slightest deflections on the part of the cathode rays. From this obtained piece of evidence, Hertz considered the hypothesis that cathode rays are not electrically charged to be strongly supported. In 1897, however, British physicist J. J. Thomson (1856–1940) performed similar experiments, but obtained a much higher degree of evacuation in the tube. Under these circumstances, Thomson found the cathode rays to be deflected as if negatively charged. He concluded that the hypothesis that cathode rays were electrically charged was now strongly supported, and hence that Hertz's earlier opposite hypothesis had been strongly disconfirmed. Thomson also provided a physical argument for why the degree of evacuation in the tube would matter (Achinstein, 2001: 13–19).

As already mentioned, evidence could be physically concrete (like in a criminal investigation) as well as very abstract (like the entirety of someone's justified beliefs, according to some philosophical accounts). In the natural sciences, a piece of evidence is often presented in a way that lies between these extremes: it may take the form of a report of some empirically established fact. For example, consider this:

[4] I take it that if something is evidence, then it normatively affects belief. I doubt that the reversed implication is generally correct (if something normatively affects belief, then it is evidence), but I will not discuss the matter here.

(⋆) Vestigial organs can be found in innumerable animal species.

Someone might claim that (⋆) is (a piece of) evidence in favour of the theory of evolution. We understand roughly what this means, but for (⋆) to function properly as evidence for the theory of evolution, it must be combined with facts about the theory of evolution and with principles of reasoning. More generally, for something to be a piece of evidence E in favour of some hypothesis H, E cannot be solely a report of some empirical fact; E also has to connect to H and it must be made clear how E is supporting H. In other words, there has to be an *argument* for H that includes E. (⋆) is not by itself a self-contained argument for the theory of evolution, but it could certainly be a crucial part of an argument for the reality of evolution. Nevertheless, we often say that a fact-report such as (⋆) constitutes "evidence" or is "a piece of evidence". There is a conceptual tension at play here. On the one hand, evidence could intuitively be just some observed fact, like (⋆). On the other hand, if we understand evidence as something that speaks for or against a hypothesis, just an observation-report will not do. An observation-report does not speak for or against anything, only arguments do. When a mere observation-report is referred to as "evidence", then, the missing pieces to make a full argument are taken as implicit background information.[5]

We have said that evidence justifies beliefs (if it is strong enough). And all beliefs must be justified and true to count as knowledge, according to the classical definition (Ichikawa and Steup, 2018). But this does not necessarily mean that justification requires evidence; whether this is so depends on exactly how evidence is conceptualised. Some statements seem to be true in virtue of their meaning. Example statements include "2 + 2 = 4" and "all bachelors are unmarried". If they are true in virtue of their meaning, then it seems that we only need to understand them properly to be justified in believing them. Hence, no *further* evidence is required for our belief to be justified. (Indeed, such statements are often called "self-evident".) Could simply the correct understanding of a statement count as evidence? I will not take a stand in this issue here. Anyone who wishes to speak only of evidence that includes (or is identical to) some observation-report could use the term "empirical evidence", for clarity. Statements like "2 + 2 = 4" and "all bachelors

[5]We here talk about evidence E as being (or not being) evidence with respect to some hypothesis H. In discussions about medical evidence hierarchies, many seem to take for granted that one can speak about different generic strengths of evidence with respect to different ways of obtaining evidence (for example, RCTs or observational studies). Strength of evidence with respect to ways of obtaining evidence is not the same as strength of evidence with respect to hypotheses. The matter is further discussed briefly in Section 4.3.

are unmarried", then, clearly do not require supporting empirical evidence to be believed.

Apart from self-evident statements like "2+2 = 4", there are also empirical statements that are considered certain to such a degree that no further evidence in their favour is needed. For example, "the earth is roughly spherical" and "life forms on earth have developed in accordance with the principles of biological evolution" are such statements. Of course, there is a principled difference to the effect that we cannot even imagine "2 + 2 = 4" to be false, whereas it is conceivable that the empirical statements are false. There is a connection here to the dominant conceptualisation of a "hypothesis" in science, namely a factual statement which is open to testing and hence open to strengthening as well as to weakening from emerging evidence (Belsey, 1995). Statements about factual matters that are already considered known do not count as hypotheses on this account; only statements whose veracity is judged to be at least somewhat uncertain can function as scientific hypotheses and warrant search for further evidence.

As for the upper limit of evidential strength, we can call evidence "conclusive" if it demonstrates H to be true. Though this may seem entirely clear, a further distinction can nonetheless be introduced. We call evidence conclusive if we are very certain that it makes the hypothesis true, but where there may still be a tiny amount of doubt. However, a statement that logically entails the truth of another statement leaves no room for doubt. But such "evidence" is rarely of interest in a scientific setting. Therefore, it could be argued that logical entailment ought to be excluded from the extension of the concept of evidence. Indeed, Achinstein (2001: 169) suggests that E must not logically entail H to count as evidence: "The fact (e) that I am wearing a blue suit today is not evidence that (h) I am wearing a suit. It is too good to be evidence." This exclusion of logical entailments is reasonable in view of the fact that confirmation theory is normally construed as a theory about non-deductive reasoning (Crupi, 2020; Glymour, 1980: 63).

2.2. Two common principles of evidential reasoning

In order to reason about evidence, we need to understand reasoning principles that ought to be obeyed. As will be discussed briefly in the next subsection, evidence is often conceived of in terms of probabilities. Naturally, there will then be principles of probabilistic reasoning to be complied with. But first, independently of whether evidence is to be understood in probabilistic terms or not, we shall mention two other principles, or lines of thought, that have often been invoked in evidential reasoning.

The first is the principle of total evidence. This principle is regularly postulated in Bayesian accounts of evidence (Bayesian epistemology), but you need not be a Bayesian to find it reasonable. Its idea is that even if some particular piece of evidence E speaks in favour of (or against) some hypothesis H, it is not solely E that determines the fate of H. (Or, formulated in accordance with the main role of evidence as per above: it is not solely E that determines whether we are justified in believing H or not.) For even if E speaks in favour of H, there may well be some other evidence E' that speaks against the correctness of H (or vice versa). To assess H, we need to take E as well as E' into account; and generally, what to reasonably believe about some hypothesis H depends on the totality of evidence for or against H. Compiling and assessing all evidence relevant for H could be a formidable undertaking. It could be particularly overwhelming under certain conceptualisations of evidence. If, for example, you subscribe to the view that your evidence amounts to the totality of your justified beliefs, and if, furthermore, your beliefs are connected in the sense that certain beliefs follow from the combination of others (which seems like a reasonable position), then an assessment of what to think about H would effectively require *all* your beliefs to be taken into account. Whether your task is to weigh in all your beliefs or just those that more directly support or oppose H, then, you might be particularly interested in finding ways of reliably judging that some piece of evidence has negligible influence on the assessment of H and thus can be discarded from further consideration. We shall not discuss how to make such assessments here, however. The first formulation of the principle of total evidence is sometimes ascribed to Carnap (1947). But he, in turn, says that Keynes mentioned it in his 1921 *A Treatise on Probability*, ascribing it there to Bernoulli (Carnap, 1947: 138 *n*). Thus, it seems to be a principle of considerable age. The principle is invoked at one point in Paper IV in the present thesis.

A second principle of evidential reasoning, which seems to build upon the principle of total evidence, is the following. The strength of some E with respect to some H is not a function exclusively of the content of E and the content of H, but is dependent also on the existence (and knowledge) of other competing hypotheses.[6] One striking example is provided by Kelly (2016). Up to the early 1800's, according to the dominant belief, humans,

[6] Already from the principle of total evidence, it follows that it is not sufficient to take only the piece of evidence E into account in assessing what to believe about H. The second principle says that it is likewise insufficient to consider only hypothesis H, although it is the strength of E with respect to H that we wish to assess. I consider hypotheses not to be evidence. Rather, a hypothesis connects different pieces of (potential) evidence. Therefore, I consider this second principle to be different from (though related to) the principle of total evidence.

animals, and plants had been created more or less in their current forms by some creator (God) in a distant past. Let us label this belief H_1. Various evidence (or, at least, what was then perceived as evidence) was invoked for this hypothesis, e.g., "intelligent" anatomical features of animals, the fine balance between predators and prey, between parasites and hosts, etc. But importantly, until the writings of Charles Darwin there was no credible alternative to a supernatural act of creation. As soon as the theory of evolution (which we may call H_2) had been formulated, there was an alternative way of interpreting the evidence previously used in favour of H_1. Arguably, then, this example shows that the available evidence for H_1 was perceived as strong before the advent of H_2, but once H_2 had been formulated, the very same evidence was judged to be considerably weaker for H_1.

This second general principle—for which I am not aware of any standard name in the literature—is a pragmatic principle of evidential reasoning.[7] By this I mean that it stems from, and receives credibility from, the fact that in real life we rarely have perfect knowledge about the situation that we assess. Because of this, we cannot do better than base our assessments on what we know. If it turns out, later on, that we have been ignorant by not being able to formulate a credible alternative hypothesis, then as soon as we find out about such hypotheses we should be open to adjusting the strength of evidence that we assign to the evidence in relation to the original hypothesis. Of course, it would be possible to hold that in some idealised sense (or: in an idealised world, where we have perfect knowledge), the strength of the piece of evidence E with respect to the hypothesis H depends exclusively on E and H, and not at all on whether we are aware of credible alternative hypotheses. But this is hardly helpful in practice. In science, as well as in other evidence-assessing activities where there is uncertainty (e.g., in jurisprudence), the assessed strength of E with respect to H is, arguably, being subject to the pragmatic principle above. Because of this principle, an important task for anyone who wishes to assess the strength of some piece of evidence E in favour of H is to try to chart the territory of alternative hypotheses.[8] The

[7] Although I am not aware of a standard name of the principle, there is an affinity with a so-called inference to the best explanation (IBE). According to a standard work on this topic, an inference to the best explanation must be understood as an inference relative to *potential* explanations, meaning that "[w]e have to produce a pool of potential explanations, from which we infer the best one" (Lipton, 2004: 58). Producing a pool of potential explanations in the context of IBE seems similar to finding alternative hypotheses in the context of evidence assessment.

[8] The principle is related another idea, namely that as the sheer collection (amount) of evidence increases on which some argument for (or against) some hypothesis rests, it is less likely that some further piece of evidence will have a substantial impact on the overall probability assessment of the hypothesis. The amassed evidence used to formulate an argument may be referred to as the "weight" or the "Keynesian weight" of the argument. The idea was formulated

principle is used (but not explicitly mentioned) in Paper II, for example when I argue that the existence of a certain mechanistic chain between intervention and outcome is not necessarily strong evidence in favour of the intervention producing the outcome, since there may be other mechanistic chains that suggest that the outcome is not produced by the intervention.

2.3. Probabilistic accounts of evidence

According to one common view of evidence, its main roles are those of confirming (or at least supporting) or of opposing (disconfirming) hypotheses. Generically, then, any theory of how this works can be called "confirmation theory". In such a theory, some measure of confirmation (or certainty) is needed, as are procedures for accommodating new evidence, for handling the uncertainty of the relevant background information, and more. Probability theory provides much of what is wanted. It is not surprising, then, that probability theory has been used to formulate and discuss what happens when evidence changes the assessment of some hypothesis.

Rudolf Carnap (1891–1970) published a number of articles in the 1940's, culminating in the 1950 book *The Logical Foundations of Probability*. This book seems to have been the first to comprehensively and convincingly show how to give a probabilistic account of confirmation theory. Confirmation theory then essentially got probability theory as its core, supplemented with explications of various notions and activities that can be performed in the probabilistic framework. A central feature of probabilistic accounts of confirmation theory is that the "degree of confirmation" (or whatever term is chosen) of the hypothesis H from the evidence E corresponds to the formation of a probability for H conditioned on the evidence E. In other words, the degree of confirmation corresponds to $\Pr(H|E)$, where "Pr" means probability, and the vertical line | has its standard meaning "given" (or "on the assumption of").

When is something (a piece of) evidence for a hypothesis? In probabilistic terms, a popular answer is that E is evidence in favour of hypothesis H if, and only if, E raises the probability of H. This is often referred to as the *positive relevance* criterion (or condition):

$$\Pr(H|E) > \Pr(H).$$

It is found in Carnap. But Carnap also discussed an alternative criterion, according to which E is evidence for H just in case the probability of H given

by John Maynard Keynes (1921: 71–78). In jurisprudence, and in some other fields, one might also speak of the "robustness" of the available evidence, which seems to refer to the same idea.

E is larger than some threshold:

$$\Pr(H|E) > k,$$

where k is the threshold. The threshold could be context-dependent, or could be a fixed number to be used in all settings (for example, $k = \frac{1}{2}$). This can be called the *threshold criterion* for evidence. An alternative name is the *high probability criterion* for evidence, since k is taken to be some (sufficiently) high probability.

Both criteria can also be stated in variants where some background information B is assumed. The positive relevance criterion then becomes

$$\Pr(H|E \wedge B) > \Pr(H|B);$$

and the threshold criterion becomes

$$\Pr(H|E \wedge B) > k.$$

According to the positive relevance criterion, the greater the increase in the probability of H from some evidence, the stronger the evidence.

We can now make formal sense of comparative evidence statements, such as "E is stronger evidence for H than is E'".[9] According to the positive relevance criterion, this corresponds to

$$\Pr(H|E) > \Pr(H|E')$$

(where B has been left out). The statement "E is stronger evidence for H than for H'" could be written

$$\Pr(H|E) - \Pr(H) > \Pr(H'|E) - \Pr(H')$$

if we take "stronger" to correspond to a larger increase in the difference between the probabilities for the hypothesis with and without taking E as given.[10] If instead we use the threshold criterion (again without B being written out), "E is stronger evidence for H than is E'" can be translated into

$$\Pr(H|E) > \Pr(H|E')$$

[9] The suggested formalisations given here were borrowed from Achinstein (2001: 47–48), who in turn claims to have taken them from Carnap.

[10] I mention the difference (i.e., the arithmetic difference) because it would seem possible to instead take "stronger" to correspond to a larger increase in the *quotient* between the probabilities for the hypothesis with and without taking E as given. This is not discussed by Achinstein, however.

(that is, identical to what was found under the positive relevance criterion), and "E is stronger evidence for H than for H'" can be translated into

$$\Pr(H|E) > \Pr(H'|E).$$

How reasonable is it to talk about scientific evidence in probabilistic terms? This question can be broken down into several smaller questions. Some of these are:

(1) What notion of probability are we talking about? There are a number of interpretations of probability around.

(2) Probabilities are required to conform to the axioms of probability calculus. What does that mean for statements about evidence in probabilistic terms, or for probabilistically modelled degrees of belief with respect to evidential statements?

(3) We have presented two different criteria for what it means to be evidence in favour of a hypothesis: the positive relevance criterion and the threshold criterion. Is one of these the right one?

I will comment on these questions briefly. There are yet other questions of considerable philosophical interest, which will not, however, be further discussed here. One such question is the "problem of old evidence" (e.g., Snyder, 1994).

Let us first note that probabilistic accounts of scientific evidence can be criticised in ways seemingly independent on how questions (1)–(3) are answered. Clark Glymour has argued, for example, that there is a considerable difference between probabilistic accounts of evidence and how scientists themselves have talked about their evidence throughout the history of science. He asserts that probabilistic arguments have seldom been given at important points in the history of science, and this makes him doubt the relevance of probabilistic accounts. He summarises: "[P]robabilistic analyses remain at too great a distance from the history of scientific practice to be really informative about that practice" (Glymour, 1980: 65). Although I believe that Glymour is factually correct in the sense that probabilistic reasoning has been used rather little in the history of science as a whole, I still suspect that probabilistic accounts can shed interesting light on the use of evidence in science, particularly in certain scientific disciplines. The issue would certainly deserve a comprehensive treatment, but it will have to be done elsewhere.

Glymour presents more critique that is worth mentioning. Even though a probabilistic account of confirmation may capture some interesting things

that really go on in scientific confirmation, there are also many other things that happen in science, things that come up in discussions about scientific methodology and progress, but these are largely absent from the probabilistic account:

> [T]here are a variety of methodological notions that an account of confirmation ought to explicate, and methodological truisms involving these notions that a confirmation theory ought to explain: for example, variety of evidence and why we desire it, ad hoc hypotheses and why we eschew them, what separates a hypothesis integral to a theory from one "tacked on" to the theory, simplicity and why it is so often admired, why "deoccamized"[11] theories are so often disdained, what determines when a piece of evidence is relevant to a hypothesis, what, if anything, makes the confirmation of one bit of theory by one bit of evidence stronger than the confirmation of another bit of theory (or possibly the same bit) by another (or possibly the same) bit of evidence. Although there are plausible Bayesian explications of some of these notions, there are not plausible Bayesian explications of others. (Glymour, 1980: 67–68)

In my opinion, this critique deserves attention, but it can also be avoided to some extent. On the one hand, it is true that the things mentioned by Glymour are regularly discussed as central methodological issues, and hence ought to be covered by a comprehensive theory of evidence. (I believe that what is stated in the quotation is still reasonably true today, although it was published in 1980.) On the other hand, a Bayesian could hold that the Bayesian theory is not claimed to be *that* comprehensive, so Bayesianism is true but has a narrower scope than what is expected by Glymour. I personally think that criticism emerging from the more specific questions (1)–(3) is of even greater interest than Glymour's general critique.

As for question (1), however, I will not review the standard probability interpretations here (see, e.g., Hájek, 2019) but will be content to note that some interpretations make the assignment of a probability to some E in favour of some H an *a priori* matter, whereas others make it dependent on

[11] "Deoccamisation" is a kind of opposite to Occam's razor. To deoccamise a theory, one replaces a theoretical term with of combination of new terms. For example, in Newtonian theory *force* may be replaced by *gorce* and *morce* (Curd and Cover, 1998: 651), where the combination of the latter two works exactly as the original *force*. The new theory would then entail exactly the same evidence (have the same observational consequences) as the old one, and they should, consequently, be assigned the same likelihoods. For a Bayesian to reject one of them (presumably the one with *gorce* and *morce*), then, that theory would have to be assigned a lower prior probability. But Glymour thinks that it is difficult to explain why a theory with more theoretical terms should be assigned a lower prior probability without introducing arbitrary restrictions on what evidence to be allowed for consideration (Glymour, 1980: 77).

empirical considerations. In a scientific context, it seems to me to be an incredible claim that the strength of evidence (in the form of probability) can generally, or always, be determined *a priori*. This then rules out Carnap's theory, according to which a statement about $\Pr(H|E)$ expresses a logical relationship, and the correct value of the probability can be computed *a priori* with reference to the rules that govern the language used (Achinstein, 2001: 49). However, it is plausible that some evidence that is useful in science can be assessed *a priori* whereas the assessment of other (and probably most) evidence requires empirical facts (*cf.* Achinstein, 2001: 101–102).

With respect to question (2), the term "probabilism" is used by Hájek (2008: 794) to denote the view that "an agent's beliefs come in degrees, which we may call *credences*; and that these credences are rationally required to conform to the probability calculus". Probabilism, then, is a necessary part of Bayesianism, and question (2) has been much discussed in order to make an overall assessment of Bayesianism. What reason do we have to think that people have degrees of belief that conform to probability theory? The standard Bayesian answer is: because we can measure those probabilities from wagers that people are willing to accept; and under the assumption that people are willing to accept wagers expected to give a gain but unwilling to accept a wager expected to give a loss, the betting odds acceptable to the person determine the degrees of belief, and they turn out to obey the axioms of probability.[12] But there are counterarguments to this answer. Glymour writes:

> [T]he subject may not believe that the bet will be paid off if he wins, or he may doubt that it is clear what constitutes winning, even though it is clear what constitutes losing. Things he values other than monetary gain (or whatever) may enter into his determination of the expected utility of purchasing the bet: for example, he may place either a positive or a negative value on the risk itself. And the very fact that he is offered a wager on P may somehow change his degree of belief in P. (Glymour, 1980: 70–71)

Furthermore, a person "may not have adopted the policy of acting so as to

[12] More carefully spelled out, the Bayesian argument that one is obliged to treat degrees of belief ("credences") as probabilities and to conform to the probability calculus is known as the *Dutch book* argument, which relies on the so-called Dutch book theorem. It says that if you do not conform to probability calculus, then there exists a set of bets, all of which you consider fair, that collectively guarantees that you lose money. Hájek (2008) has pointed out that there is a mirror-image theorem, the "Czech book" theorem, which says that if you violate probability theory, then there exists a set of bets, all of which you consider fair, that collectively guarantees that you *gain* money. Hájek shows that if you reformulate the Dutch book and Czech book theorems in terms of fair-or-favourable bets, then the argument for probabilism can be saved.

maximize our expected gain or our expected utility" (Glymour, 1980: 72). These criticisms derive their force from the fact that people may not behave as perfectly rational agents, which is presupposed in the Bayesian theory.

With respect to question (3), Peter Achinstein has presented a number of counterexamples to the two evidence criteria. The view that he wishes to promote is that evidence (as opposed to probability) is a threshold concept. And so is belief, Achinstein thinks. Nevertheless he rejects the threshold criterion as a sufficient *and* necessary condition for evidence. We shall review his argumentation as it appears in Achinstein (2001: 69–94).

Achinstein's first counterexample is as follows. Let B be that all of the 1000 tickets that made up a lottery were sold on Monday, of which John bought 100 and Bill bought 1. One ticket was planned to be drawn on Wednesday. Let E be that on Tuesday all lottery tickets were destroyed except those 101 that were bought by John or Bill. On Wednesday one of the remaining tickets was drawn. Consider hypothesis H: Bill won. Using just the background information B, we would assign the probability $\Pr(H|B) = \frac{1}{1000}$. Incorporating the evidence, we would say $\Pr(H|E \wedge B) = \frac{1}{101}$. In other words, the probability of H is nearly multiplied by ten when E is taken into account. According to the positive relevance criterion, then, E clearly constitutes evidence for H. But something is strange here (or so says Achinstein). Even though the probability that Bill won is raised considerably when E is taken into account, the probability that *John* won when E is taken into account is raised even more; in fact, this probability is $\frac{100}{101}$. In view of this overwhelming probability, one would be inclined to say that E is evidence that *John* won, not that Bill won. This counterexample shows, according to Achinstein, that an increase in probability is not sufficient for being evidence.

In Achinstein's second counterexample, B is that Steve is an olympic swimming team member who was in good shape on Wednesday morning. E is that Steve was training in the pool on Wednesday. H is that Steve drowned on Wednesday. Since Steve was in good shape, it was highly unlikely that he would drown on Wednesday. The probability that he would drown on Wednesday was nonetheless raised by the fact that he was training in water on that day. Therefore, according to the positive relevance criterion, E is evidence for H. Achinstein finds this to be absurd and concludes, again, that an increase in probability is not sufficient for being evidence.

Is an increase in probability necessary? Not so, says Achinstein. One of his counterexamples is claimed to show that even a *decrease* in probability may count as evidence, according to our intuitions. The example goes like

this. Let B be the following facts about two medicines M and M' against the symptoms S:

- M is 95 % effective in relieving S within two hours.
- M' is 90 % effective in relieving S within two hours and has fewer side effects than M.
- If M' is taken within 20 minutes of having taken M, then M' blocks the efficacy of M, but the efficacy of M' is unaffected.

Let E_1 be that David, having symptoms S, takes M on Monday at 10.00 A.M. Let E_2 be that David takes M' on Monday at 10.15 A.M. Now, consider H: David's symptoms S are relieved by noon on Monday. It seems reasonable to say that $\Pr(H|E_1 \wedge B) = 0.95$, whereas $\Pr(H|E_2 \wedge E_1 \wedge B) = 0.90$. Hence, when E_2 is taken into account (along with E_1), the probability that David's symptoms are relieved by noon has decreased compared to when only E_1 is considered (always in combination with B). According to the positive relevance criterion, then, E_2 is not evidence for H, given E_1 and B. Still, Achinstein believes that many of us would be inclined to say that E_2 *is* evidence for H (given E_1 and B), as the probability of H is no less than 90 % on the assumption of E_2.[13]

Finally, Achinstein presents a counterexample to the claim that high probability is sufficient for evidence. B is that Michael Jordan is a male basketball star. E is that Michael Jordan eats Wheaties. Consider H: Michael Jordan will not become pregnant. The probability of H is very high already given the background information B (where the most important piece of information, of course, is that Michael Jordan is male). If we add E, the probability of H is not changed, which means that the probability is still very high. Therefore, according to the criterion that something is evidence for a hypothesis if, and only if, the probability of the hypothesis given the evidence exceeds some threshold, E seems indeed to be evidence for H.

The message emerging from Achinstein's counterexamples is, to summarise, that an increase in probability is neither necessary nor sufficient for evidence, and that high probability is not sufficient for evidence. Achinstein does think that high probability is necessary, however. His argument relies on the (seemingly reasonable) assumption that if E is a good reason to believe H, then E cannot be a good reason to believe not-H (Achinstein, 2001: 116).

Sherrilyn Roush (2004) has defended positive relevance as a necessary (but not sufficient) condition for evidence. Consequently, she criticises Achinstein's example with David and the two medicines M and M'. (Roush also

[13] One will have to assume, although it is not explicitly stated in Achinstein, that the probability is low that S will disappear by noon on Wednesday if we do nothing.

discusses another Achinstein example claimed to show that positive relevance is not necessary for evidence, but since I have not accounted for it above, it is left out of the current discussion.) Roush accepts the probabilities $\Pr(H|E_1 \wedge B) = 0.95$ and $\Pr(H|E_2 \wedge E_1 \wedge B) = 0.90$, but she thinks that Achinstein's reasoning nevertheless fails on two accounts. First, when Achinstein says that we (or, at least, many people) intuitively accept that E_2 is evidence for H since the probability of H is 90 % on the assumption of E_2, this claim seems to rest on a high probability criterion for evidence: because 90 % is such a high probability, we ought to accept that E_2 is evidence for H. But as we have seen above (the Michael Jordan example), Achinstein is himself critical of the high probability criterion. And so is Roush, so her first counterargument is, essentially, that it does not follow from $\Pr(H|E_2 \wedge E_1 \wedge B) = 0.90$ that E_2 is evidence for H. The probability is high, but this does not make E_2 evidence for H.

Secondly, there is no doubt that $\Pr(H|E_2 \wedge B) = 0.90$, and this is presumed to be higher than $\Pr(H|B)$, although the latter probability is not assigned a numerical value in Achinstein. According to the positive relevance criterion, then, E_2 *is* evidence for H when there is no mention of E_1 (and hence no mention of David taking M). But Achinstein never asks us to compare these two probabilities, but rather the probabilities $\Pr(H|E_2 \wedge E_1 \wedge B) = 0.90$ and $\Pr(H|E_1 \wedge B) = 0.95$, where E_1 is given (assumed) in both. Roush hypothesises that the comparison between $\Pr(H|E_2 \wedge B)$ and $\Pr(H|B)$, which we are *not* asked to make, nevertheless influences some people's intuitions about the comparison between $\Pr(H|E_2 \wedge E_1 \wedge B)$ and $\Pr(H|E_1 \wedge B)$, which is the one we *are* asked to make. In itself, the fact that $\Pr(H|E_2 \wedge E_1 \wedge B) < \Pr(H|E_1 \wedge B)$ should have us conclude that E_2 is *not* evidence for H given $E_1 \wedge B$. But this is consistent with E_2 being evidence for H given *only* B. In other words, Roush flatly denies that E_2 is evidence for H given $E_1 \wedge B$.

I think Roush's counterarguments are well formulated, and I also think that the reply by Achinstein (2004) is weak (I will not go into details here). In sum, then, I am not convinced that Achinstein is right when he denies that positive relevance is a necessary condition for evidence. However, I am tentatively willing to subscribe to Achinstein's claims that neither positive relevance nor high probability is sufficient for evidence.

In the present thesis, the positive relevance and high probability criteria do not play crucial roles, though positive relevance is being mentioned in Paper IV. But the general conceptualisation of evidence—where the mentioned criteria could matter—is important in some additional places, particularly in Paper II.

3. Experiments

3.1. Basic characteristics

With the advent of the scientific revolution, the experiment soon emerged as the chief method of acquiring reliable data about the natural world.[14] It took some time for the term *experiment* to have its meaning fixed. The English word is derived from the Latin verb *experīrī*, "to try (out)". The modern sense of the word experiment, in a scientific setting, can be taken to be the following:

> An experiment is a series of observations of outcomes, where all factors believed to influence the outcomes are controlled and at least one such factor is manipulated by the experimenter.[15]

We shall first say a little about the meaning of "controlled" and "manipulated", respectively.

What we mean today by "control" in an experimental setting is rather a fusion of three reasonably separable but connected meanings of the word (Boring, 1954). The first meaning is check (or standard of comparison, verification). Something is controlled if there is a way of checking or verifying it. This is the original meaning of the word, which is derived from French *contre-role*, a duplicate register that can be used for checking or verifying the original register. The second meaning is restraint. Something is being controlled if there are restrictions on how it may vary. When we say that individual factors (variables) in an experiment are controlled, we mean that they do not vary (fluctuate, differ, change) in such a way that we lose track of what variations are responsible for what changes in the outcome variable. One possible restriction of a variable is, obviously, to keep it constant (unchanged); indeed, this is a common tactic in experimental designs. A factor can be controlled either through measures imposed by the experimenter, or without any such measures. It is thus not necessary, in order for a trial to count as an experiment, that every factor that could decrease our knowledge of how the outcome is being influenced is *actively* restrained by the experimenter; it is sufficient that the factors are judged, by the experimenter, not to differ in an unwanted way. Since variables in an experiment that ought to be controlled

[14] This subsection reuses material from Jerkert (2019: 71–82).

[15] With this formulation, computer simulations appear to be experiments. If one wishes to exclude simulations, one would have to add some condition with this effect. One candidate could be a requirement that the "factors" mentioned are materially similar to the targets that they represent. However, the question is complicated, and some have denied the significance of material similarity for distinguishing experiments from simulations (Winsberg, 2010).

in this sense of being restrained can be called confounders (or confounding variables), we might say that an important purpose of experimental control is to eliminate (or, at least, minimise the influence of) confounders. The third meaning is guidance (or management, direction). It is particularly the manipulated variable in an experiment that is being controlled in this third sense: it is being varied in a way known to the experimenter, thus putting the experimenter in a position of guiding or directing the experimental course of events.[16]

All three meanings of "control" can also be found outside of science. An inspector may control something in the first sense (check, verify). Someone may wish to control his or her behaviour in the second sense (restrain, restrict). And it is predominantly in the third sense (guide, manage) that a violin player controls his or her vibrato.

Manipulation, in turn, amounts either to actions taken by the experimenter, or, at least, to actions taken by someone on behalf of (or, according to instructions from) the experimenter. At least one factor believed to affect the outcome is to be manipulated in an experiment, and usually it is precisely the relation between the manipulated factor and the outcome that the researcher wishes to investigate by performing the experiment. (Sometimes the relationship between two or more manipulated variables and the outcome is of interest.) In order to extract as much and useful information about this relationship as possible, manipulations are normally imposed in a planned way, for example according to some predetermined order, or according to a random order.[17] There is no tension or contradiction between the concepts of experimental manipulation and experimental control: the experimental variable that is being manipulated is simultaneously being controlled (mainly in the third sense described above).

As a paradigmatic example of an experiment, consider Galileo's inquiries into the effect of gravity on free-falling objects. Galileo assumed that balls rolling down an inclined plane would accelerate in the same way as in free

[16] As a side-note, there is a difference between the similar-sounding "controlled experiment" and "control experiment". In a *controlled experiment* there is control, as described in the main text. A *control experiment* (or control trial, or simply "a control") is one in which a default or no-change condition is being applied (as opposed to the intervention of interest). Subjects that are only exposed to a control condition form a control group, the outcome statistics of which may be compared to that of the group that was exposed to the intervention of interest. Applying control conditions contributes to control, but control experiments are not in general necessary for control to be achieved in scientific experiments.

[17] When manipulation is imposed according to a random order, this is often done with the sole or partial purpose of preventing the experimenter from knowing which experimental condition is imposed in which experimental unit. This, then, will be "actions taken by someone on behalf of (or, according to instructions from) the experimenter", according to our description of manipulation, above. *Cf.* Section 3.3 on randomisation.

fall. (He preferred inclined planes since they would allow for more careful time measurements.) He released balls and measured the times elapsed when the ball had travelled many different, carefully recorded, distances along the plane. At different points along the plane, there were markers attached that would be touched by the ball (or strings stretched perpendicular to the ball track, which would be touched lightly by the top of the ball), creating an audible sound. By adjusting these markers, Galileo was able to create a steady rhythm from the sounds made by the ball as it accelerated. It is, arguably, possible to hear very small deviations from a steady rhythm for a musically trained ear.[18] Galileo found that the distance travelled between the markers was proportional to the time squared: $d \propto t^2$. This confirmed what he had hypothesised before he started making his measurements. The outcomes that Galileo measured were temporal durations (more precisely: equality of temporal durations). What he manipulated were the distances travelled by the balls (particularly: the distances between the markers). Factors that were believed to be able to affect the outcome in unwanted ways were controlled: Galileo used the same inclined plane, which was never moved, and he rolled the same balls over and over again.[19]

Sven Ove Hansson (2015) has made an interesting and useful distinction between two types of experiment. On the one hand, there are *epistemic* experiments. These are performed with the aim of gaining knowledge and understanding. Galileo's experiment with an inclined plane is a good example. It was performed to obtain knowledge about a general phenomenon (gravity), *not* to obtain knowledge about the effects of only precisely that which was actually done in the experiment, i.e., rolling balls down an inclined plane.

The other type is *directly action-guiding* experiments. Such an experiment is performed in order to find out whether a particular action yields a wanted result, more or less independently from any explanations that could be offered, or from any theories that could be tested in the experiment. In other words, the idea leading to a directly action-guiding experiment is simply this: "if you want to know if you can achieve Y by doing X, do X and see if Y occurs"

[18] Galileo came from a musical family. His father Vincenzo Galilei (1520–1591) was a composer and a music theorist, who also made experiments on the relation between pitch and string tension. The Galileo scholar Stillman Drake even argued that "[m]usic played not only a unique, but an essential role in leading Galileo to his new physics, a science of precise measurements, for music is an art demanding precise measurement and exact divisions" (1992: 15).

[19] There have been discussions in the Galileo literature as to whether this was the exact way in which the inclined plane experiments were performed; and also, were this the right way, as to whether it is realistic to obtain results that confirm so closely to the $d \propto t^2$ rule as did Galileo. Naylor reports unsuccessful attempts to replicate the experiment, concluding that "the experiments demonstrate the impracticality of searching for a rule linking time and distance by this means" (1980: 377). On the other hand, Riess, Heering, and Nawrath (2005) report success in replicating the experiment.

(Hansson, 2015: 85). Medical researchers who test a new treatment in a clinical trial are primarily performing a directly action-guiding experiment, because they want to see whether they can achieve a cure (health improvement) by administrating a specified treatment. I am making a reference to this distinction in Paper IV.

The distinction allows us to see that experimentation is a human activity that did not first emerge within science, as some people seem to think. It is true that epistemic experiments were unusual, perhaps even non-existent, before the advent of modern science in the 1600's. But directly action-guiding experiments, on the other hand, have been carried out for a much longer period of time, and have been performed independently in many parts of the world. A few examples may be offered to substantiate this claim (Hansson, 2015: 85–91):

– Incas have performed agricultural experiments in pre-colonial Latin America, systematically testing different crops under different microclimactic conditions. Several African people have performed agricultural experiments, too. The evidence comes both from archeological records and from the fact that many people still perform such experiments in ways seemingly independent from Western influence.

– Craftspeople have been experimenting with materials, building constructions, tools, music instruments, and more. To give just one example, there is a preserved instruction from ancient Egypt that the hardest bronze is made of 88 % copper and 12 % tin. This knowledge seems to have been obtained during a rather short period of time. It is difficult to see how it could have been obtained in any other way than through systematic experimentation.

We have reason to believe that modern science learned some experimental methodology from the earlier experimental traditions. This underlines the continuity between older and newer experimentation (Hansson, 2015: 98–106). However, two methodological features of modern scientific experiments seem not to have been imported into science from pre-scientific action-guiding experiments, but were instead developed over the last 150 years within science. I am thinking of blinding and randomisation, both of which are particularly salient in clinical trials. The developments of these two features are sketched in Sections 3.2 and 3.3.

If this wide notion of experiment is accepted, according to which experiments have been performed for many hundreds or even thousands of years (albeit only or mainly action-guiding ones before the 1600's), then it is

impossible to trace the very beginning of human experimentation. An important development that took place in the 1600's was, as is well-known, the integration of experiments with mathematical theorising. Indeed, in Galileo's experiment described above, he found the relationship $d \propto t^2$. Finding such a mathematical relationship between variables is, of course, not a necessary outcome from an experiment. But it is not unusual for epistemic experiments to have a mathematically expressible theoretical connection. The scientific revolution, as it is sometimes called, saw a rise and refinement of epistemic experiments.

In this connection, it is often emphasised that experiments are good at confirming (corroborating) or disconfirming hypotheses. Again, in Galileo's example case he had a hypothesis that distances and time durations were related according to the formula $d \propto t^2$, and this hypothesis was made more credible as the experimental results agreed with it. But again, testing a specific hypothesis is not a necessary condition for performing an experiment in general.

We shall be content with what has been said so far about experiments in general. Of course, there is a substantial literature on various aspects of the philosophy of experiments and experimentation. Classical topics include the relationship between experiments and theory (including the nature of the theory-ladenness of observation statements and, by extension, of experiments), the repeatability of experiments, the role of experiments in scientific realism, the charge that there is a fatal "experimenter's regress" with respect to instrument calibration, and much more (e.g., Hacking, 1983; Franklin, 1990).

3.2. A brief history of blinding

Experimental blinding (masking) amounts to intentional ignorance on some part. In the context of evaluating medical interventions, a participant is blinded (masked) if s/he does not know to which treatment group s/he belongs. Blinding conceived just as the concealment of information that could otherwise compromise some wanted condition is such a general method that it is bound to have been discovered and used independently many times. For example, blinding in this general sense is an essential feature of card games, which have been played for several hundreds of years. If we turn to the context of systematic knowledge gains, however, the history of blinding is more recent.

In a thorough exposé, Ted J. Kaptchuk (1998) distinguishes five phases in the history of blinding as a research methodology in medicine, psychology

or pharmacology. The earliest phase concerned the medical establishment's response to threatening unconventional healing methods such as mesmerism and homeopathy.[20] Mesmerism—named after its inventor Franz Anton Mesmer (1734–1815)—seems to have been the first unconventional healing system to be combatted with an armoury that included blinding. The earliest instances featured physical blindfolding with bandages in experiments performed under Benjamin Franklin's supervision. Benjamin Franklin was American ambassador to France at the time. He headed a commission made up of members of the Academy of Sciences, appointed by Louis XVI, the last king of France.[21] Female subjects, selected by a mesmerist because they were believed to be reliable, were asked to point to the body part directed towards the place from which the mesmeric energy was purportedly being emitted. When the women were able to see, they unfailingly pointed to body parts directed towards the "source". When blindfolded, they placed the sensations randomly with respect to the correct direction (Kaptchuk, 1998: 395). Blinding in the form of concealment or sham treatment was soon accepted as a standard feature. Furthermore, "[b]oth sides of the dispute adopted the strategy of blind assessment and argued that any evidence supporting the opponent could be attributed to imperfect or unfair experimental conditions or fraud" (Kaptchuk, 1998: 398).

Homeopathy was tested using blind conditions on several occasions during the 19th century. Some tests compared a homeopathic remedy with a placebo, others compared a homeopathic and an orthodox remedy. The most rigorous tests were even double blind, according to today's standards. A remarkably modern test with respect to methodology was performed in Nuremberg in 1835. J. J. Reuter, a local homeopath of some stature, had asserted that a homeopathic dilution of salt (NaCl) would have clear effects in healthy people, and after some heated debate it was agreed that a large trial was to be conducted. One hundred vials were split into two lots. Half were filled with distilled snow water, and the others were filled with a C30 dilution of salt in snow water, prepared according to Reuter's instructions. The vials were numbered 1–100 randomly with respect to their content. 54 vials were distributed, most of them at a large meeting, to citizens willing to participate. In a meeting three weeks later, the participants reported their experiences after ingesting the vial contents. Reports were obtained from 50 out of 54 participants. Of these, only eight reported anything unusual, of

[20] A few even earlier examples of blinding have been described (see, e.g., Kaptchuk's [1998] note 9), but these were isolated instances that cannot be considered part of a more general awareness of blinding issues in science.

[21] For an amusing account of Franklin's investigation into mesmerism, see Lopez (1993).

which five had received the homeopathic dilution and three had received snow water. Since Reuter had predicted that most participants who received the homeopathic dilution would have unusual experiences, it was concluded that he was wrong (Stolberg, 2006).

Later phases in the history of blind assessment concern, e.g., psychophysical experiments on the smallest differences in sensation that are discernible to humans, the existence of parapsychological phenomena such as telepathy, and the debate on whether hypnotism was or was not due to suggestion. Around the turn of the century 1900, blinding was gradually being established, in particular in German-speaking countries, as a routine precaution also in the testing of stimuli or substances not necessarily associated with strange phenomena such as telepathy or hypnotism. English-speaking countries soon followed suit, and blinding was incorporated into pharmacological testing. From this usage, it was quite natural a step to also introduce it in clinical research. Blinding as a methodological feature of controlled trials constitutes the last phase in Kaptchuk's review (1998: 421–432).

Kaptchuk distinguishes two historical motivations for using blinding in clinical trials. In research performed by German-speakers, the main rationale was the need to eliminate suggestion on the participants' part. To the Anglo-Americans, this motivation was originally not that important. Instead, the problem to be solved with blinding was that when one group of patients was to be used as a no-treatment control (which seems initially to have been the most common design), these patients were more difficult to recruit and more likely to drop out during the course of the study. The solution was to give something to the patients of this group too (a placebo remedy) but without informing them about it. In this way, the "recruitment and retention nightmare" (Kaptchuk, 1998: 423) could be avoided. Of course, the two motivations can be used in conjunction. If the control (comparison) group is given another active treatment rather than placebo or nothing, the "Anglo-American" motivation loses much of its force, however.

3.3. A brief history of randomisation

In the context of medical intervention assessments, randomisation amounts to the allocation of subjects to different treatments using some random mechanism. Put more generally, randomisation amounts to the selection of experimental conditions by dint of a random mechanism. The assignment of experimental condition to the experimental unit may be made in several steps. For example, one may first non-randomly assign experimental units to different groups and then, in a separate step, randomly decide which ex-

perimental condition will be assigned to which group, or to which members within the groups. In such a two-step set-up, the purpose of the first step is typically to stratify the subjects; i.e., to create smaller groups (*strata*) within each of which the members are similar with respect to some variables. The second step is the randomisation proper.[22] Whether or not randomisation is made in a single step (simple randomisation) or in a two-step procedure that includes stratification, an important feature is that each experimental unit is subject to the same probability of being assigned to a particular condition (treatment). The probability is thus identical across experimental units (e.g., trial participants), but is not necessarily identical across conditions (treatments), since one may wish to create treatment groups of unequal size.

Randomisation is a younger methodological feature than blinding, at least if we consider the time at which it had become a fairly widespread practice.[23] A trial published in 1948 on the use of streptomycin in pulmonary tuberculosis has been dubbed a "watershed" due to the careful use of randomisation (Doll, 1998). Not until then, after World War II, did randomisation gain its current status as a very important feature of well-performed clinical trials, and the 1948 streptomycin trial was instrumental in that development. In some writings one could get the impression that this trial was the first to use randomisation at all, but that is not true. For example, already in the 1920's and 1930's randomisation had an influential promoter in Ronald A. Fisher, whose textbooks (1925; 1935) went through several editions. Fisher, renowned as a statistician and as an evolutionary biologist, initially worked empirically with agricultural field trials, where different seeds, or identical seeds with different manures added, were grown in matched plots, and where Fisher insisted on the use of randomisation.

But Fisher was not first either, nor was agricultural science the first discipline in which randomisation was applied. An 1898 experiment conducted by Danish physician Johannes Fibiger (1867–1928) was at its centenary identified as the first controlled trial in a modern sense (Hróbjartsson *et al.*, 1998). The trial investigated the effect of serum treatment on diphtheria. Fibiger allocated patients to standard treatment or to standard treatment plus serum treatment according to day of admittance.

Which experiment is judged to be the first to have been randomised will obviously depend on what procedures we accept for yielding (sufficient) randomness. Today, one would hardly use the day of admittance as ran-

[22] More information on stratification is found in Paper V, Section 4.

[23] Of course, randomisation is, being by definition a feature of experiments, a special case of a more general use of random mechanisms such as lotteries for the allocation of burdens, resources, punishment, and more in contexts unrelated to scientific experiments. For a historical review of such lotteries, see Eckhoff (1989).

domisation variable. Perhaps one would expect the history of experimental randomisation to exhibit a steady progress towards better random number generation methods, but this is not so. The casting of lots has been used for thousands of years for creating justice or for divinatory purposes. In a medical setting, the Flemish physician Johannes Baptista van Helmont (1580–1644) suggested a trial to test humoural pathology, including the casting of lots for treatment allocation:

> Let us take from the itinerants' hospitals, from the camps or from elsewhere 200 or 500 poor people with fevers, pleurisy etc. and divide them in two: let us cast lots so that one half of them fall to me and the other half to you. I shall cure them without blood-letting or perceptible purging, you will do so according to your knowledge (…) and we shall see how many funerals each of us will have. (Helmont, 1648: 526–527; quoted from the translation in James Lind Library [no date])

Hence, Helmont is sometimes credited as the originator of experimental randomisation.[24] If competently executed, the casting of lots is a good procedure in terms of randomness. The same is true for a careful shuffling (e.g., of a deck of cards) and for coin flipping. However, as described by Chalmers *et al.* (2012), the most popular allocation schedule during the early decades of the 20th century to ensure fair comparisons in clinical trials was alternation (or rotation when there were more than two conditions). Alternation is vulnerable to bias if the underlying order of patients entering the study is predetermined or manipulated. In terms of the quality of randomness, then, alternation is not as good as the casting of lots or the drawing of cards from a properly shuffled deck.

Helmont's suggestion published in 1648 did not establish any practice of randomised trials. Ian Hacking (1988) has traced the *systematic* use of modern randomisation (i.e., using procedures that produce adequate randomness) to psychological research in the 1880's. As a prelude in 1883–84, Charles Sanders Peirce and Joseph Jastrow carried out psychophysical experiments on the discernibility of small weight differences. A post office balance was placed with one half visible and the other half hidden behind a screen from the subject's perspective. (The experimenter was on the other side of the screen.) A 1 kg mass was placed on the experimenter's pan, exerting a pressure on the subject's finger. Then an additional tiny weight could be added to and removed from the experimenter's pan at given points in time. Would the

[24] However, the trial proposed by van Helmont is unlikely to have taken place; see Donaldson (2016).

subject be able to discriminate the difference? To find out, two presentation orders were used; either the one just described (1 kg, then 1 kg plus a tiny extra, then 1 kg) or the reversed order (1 kg plus a tiny extra, then 1 kg, then 1 kg plus a tiny extra). These two orders of presentation were alternated using a random mechanism, namely drawing black or red cards from a shuffled pack. The subject was to determine whether the pressure was increased or decreased during the middle presentation. The Peirce–Jastrow study, though important, did not immediately change psychophysical experimentation.

Instead, it was within parapsychology (then called "psychical research") that randomisation first became more extensively discussed after its introduction. Almost concurrently with the Peirce–Jastrow experiment, Charles Richet (1850–1935), a French physiologist with a great interest in parapsychology, carried out card-guessing experiments in which each card was drawn at random. A "sender" concentrated on the card for some time, and a "receiver" (i.e., another person) then guessed the suit of the card. The main rationale for Richet's use of a randomiser seems to have been that it enabled inferences about the existence of hypothesised very weak telepathic effects in large trials, namely by allowing the expected number of successes from mere chance to be calculated. For example, in a series of 2927 guesses Richet recorded 789 successes, as compared to the chance level of 732 successes (Hacking, 1988: 438). Richet's experiments were given attention in parapsychological publications; hence his practice of random drawing surely became widely known in this community. Richet's method of using randomisation for allowing chance calculations against which to judge some empirical result seems to be a forerunner to the null hypothesis testing motivation that was later developed by Ronald Fisher and his followers.

A few years after Richet's experiments, randomisation was again debated in relation to telepathy when trials in the form of number-guessing were performed (though not by Richet). Critics pointed out that the selection of numerals to be transferred was crucial. For if people tend to think alike and share favourite integers among those available for the test, then we would expect performances above chance level although no telepathy had occurred. Randomised numerals could be used to remove this bias (Hacking, 1988: 442).

Later in history, and particularly in specific research settings, even more arguments in favour of randomisation have been offered. In Papers IV and V, I distinguish and assess seven common modern arguments for randomisation in clinical trials. Since randomisation can be claimed to achieve several goals in a single experimental setting, there is nothing strange with offering multiple, more or less independent, arguments for randomisation. (The same is also true for blinding.)

4. Medical intervention research

4.1. General history

There are several examples of empirical investigations into the effectiveness of medical interventions performed hundreds of years ago. Two examples may suffice here.

Before the 18th century, there were irregular and serious outbreaks of smallpox in large parts of the world. In the 1720's, inoculation (variolation) against smallpox was introduced in Europe, and its pros and cons were debated. At that time, the method of inoculation was already known and practised in, e.g., Greece, Armenia, and North Africa (Huth, 2006: 262). In Britain, medical doctors approached King George I and asked for prisoners to be made available for inoculation experiments. In August 1721, six prisoners were inoculated, with successful results. Further trials were performed on parish orphans (Boylston, 2010).

Scurvy plagued and killed seamen for centuries. Empirical evidence that lemon or orange juice could alleviate the symptoms or even cure the disease had accumulated for quite some time but had not been acknowledged by the medical establishment when James Lind (1716–1794) performed what has been credited as the first controlled prospective therapeutic trial. In 1747, he chose twelve similar scurvy cases while serving as a surgeon on *HMS Salisbury*. Two were given a quart of cider a day for two weeks, two were given 25 drops of sulphuric acid elixir three times a day for two weeks, two were given two spoonfuls of vinegar three times a day for two weeks, two were given half a pint of seawater a day for two weeks, two were given a purgative electuary three times a day for two weeks, and two were given two oranges and one lemon every day for six days (after which the quantity that could be spared had been consumed). Those who ate oranges and lemons experienced the greatest recovery from scurvy (Baron, 2009: 318).

There are not only individual historical examples of empirical investigations into the effectiveness of medical treatments, but also an interesting theoretical affinity between the empiricism favoured in today's EBM and certain trends in French medicine of the early 1800's. In the aftermath of the 1789 revolution—which had led to the abolishment of old hospitals and medical schools and the creation of new ones—Paris had emerged as the world centre of medical science, giving rise to a new type of medicine aptly called "hospital medicine", since it was, in the words of Erwin H. Ackerknecht (1967: 15), "only in the [new] hospital that the three pillars of the new medicine—physical examination, autopsy, and statistics—could be developed". These methods,

or at least their integration, were unknown to followers of ancient medicine.[25] Hospital medicine favoured an empiricist stance. Pierre Charles Alexandre Louis (1787–1872) was the most staunch promoter of a "numerical method", according to which statistics was made the basis of medicine (Ackerknecht, 1967: 9–10). Louis founded the Société d'Observation Médicale, and his thus practised *médecine d'observation* has been identified as a historical predecessor of EBM (Vandenbroucke, 1996).

Empirically oriented schools such as *médecine d'observation* and historical examples of empirical investigations into the effectiveness of medical interventions could create the impression that there is a straight line up to today's clinical trials of new medical interventions. However, on the contrary it has been argued that the question of effectiveness has been largely and curiously neglected in the history of medicine. Historian David Wootton, in his intriguing book *Bad Medicine*, has made this point forcefully, and has asked why the medical establishment has been so reluctant to turn discoveries in anatomy and in the natural sciences into practical use (and also why historians of medicine have been so uninterested in investigating this historical fact). As he notes, bloodletting was the main intervention in Western medicine well into the 1800's, and it seemed to make little difference to the practices of clinical medicine that the circulation of the blood was discovered in 1628, that oxygen was discovered in 1775, or even that the role of haemoglobin was established in 1862 (Wootton, 2007: 17).

At least since the days of Hippocrates (c. 460–377 BC), and influencing medicine for more than 2000 years, the dominant theory of the workings of the body held that there are four fluids, also known as *humours*, that are supposedly captivated in the human body: black bile, yellow bile (or choler), phlegm, and blood. Bad health was associated with an imbalance between these fluids, and it was the task of a medical doctor to restore the balance in an unhealthy person (Porter, 1998: 56–58). Given that the humoural theory of disease is wrong, it may not seem surprising that doctors for a very long time had little to offer in terms of effective treatments. If they were largely incorrect about how the body worked, how could they come up with effective treatments?

[25] One important reason why their integration was unknown to preceding generations is that hospitals in earlier times were not yet centres for medical discovery and for education of aspiring medical practitioners. These functions, which we take for granted today, were assigned to hospitals starting in the 18th century. Hospitals then became "modern". This explains Uddenberg's (2015: 300) claim that *Serafimerlasarettet* in Stockholm, founded in 1752, was Sweden's first hospital, in spite of the fact that the Danviken hospital, near Stockholm, is known to have been founded in 1558. Uddenberg's is, in effect, a claim that Serafimerlasarettet was the first *modern* hospital in the country.

Of course, it does not follow from the fact that you are ignorant of anatomy and physiology that you cannot test the effectiveness of medical treatments that are in use. The failure to test treatments empirically still demands an explanation. Perhaps a particularly staunch reliance on erroneous theories (such as humoural pathology) and perhaps also perceived connections between health and spiritual/religious beliefs, that were not so easy to challenge (Hansson, 2015), could explain the historical reluctance against testing treatments empirically? David Wootton has maintained that it is strange that medical treatments were not being put to test earlier:

> The real puzzle with regard to the history of medicine before germ theory, as with the history of astrology, is working out why medicine once passed for knowledge. The case of medicine is, at first sight, rather more intractable than that of astrology, for it is hard to disprove astrology (…). But medicine, it would seem, is quite different, for it is obvious how to set about testing the efficacy of a medical therapy. All that is needed is to take a group of patients with similar symptoms and treat some of them and not others. Moreover, (…) there is a convenient crude measure of success to hand: the ratio of those patients who are still above ground to those who are now below ground. If it is this easy to put medicine to the test, why did traditional medicine survive untested into the nineteenth century? (Wootton, 2007: 144)

The reluctance towards testing medical treatment claims empirically has faded by now. In the last 70 years or so, performing a so-called clinical trial has become widely accepted as the standard way of establishing whether a proposed medical intervention is effective or not. In a clinical trial, patients with a specific disease (or with a specified collection of symptoms) are divided into groups. The participants of one group get the treatment of interest, whereas the others are given other treatments (or no treatment). The outcomes are compared group-wise with statistical methods.

Only in certain delimited quarters is a sceptical attitude towards the internal methodology of clinical trials upheld.[26] Alternative medicine is the prime example. Arguments to the effect that alternative medicine, due to its highly individualised treatments, cannot be properly tested in clinical trials are scrutinised and found wanting in Paper I.

[26] By "internal methodology" I mean how a clinical trial is performed once it has started. On the other hand, there is a lot of discussion about the circumstances under which it is ethically permissible to recruit participants to clinical trials, given the uncertainty of the treatment benefits and harms. There is also considerable debate on the generalisability of clinical trial outcomes.

4.2. The origin of EBM

As was noted above, there are features of EBM that extend back several centuries in the history of medicine. Nonetheless, the *modern* movement of evidence-based medicine is much more recent and has a rather precise origin in time. Although the term "evidence-based medicine" had been used in print somewhat earlier, a *JAMA* article by a body called the Evidence-Based Medicine Working Group (1992) counts as the founding document. There were 31 signatories that comprised the Evidence-Based Medicine Working Group, with Gordon Guyatt designated as chair. Of these, 24 signatories were affiliated with departments at McMaster University, Hamilton, Ontario (Canada). The EBM movement thus also has a rather precise geographical origin. The 1992 article opens as follows:

> A new paradigm for medical practice is emerging. Evidence-based medicine de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient ground for clinical decision making and stresses the examination of evidence from clinical research. (Evidence-Based Medicine Working Group, 1992: 2420)[27]

The authors then introduce EBM by way of two scenarios. In both, a patient is admitted to a hospital after having experienced a grand mal seizure for the first time. The patient wants to know the risk of recurrent seizures. In the first scenario, "the way of the past", the physician, having taken some standard measures, is told that the risk is "high" by a senior physician. This information is conveyed to the patient. In the second scenario, "the way of the future" (i.e., the EBM way), the physician, having taken some standard measures, conducts a literature search and reads those papers that turn out to be relevant in the current situation. Thanks to this, the physician is able to give the patient much more precise figures about the risk of recurrent seizures.

Another often quoted statement on the nature (or definition) of EBM is the following, by some of its pioneers:

> Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means

[27] The authors are Kuhnians when they talk about a "new paradigm" (this is clear from their subsequent discussion). This is interesting from a philosophy of science perspective, but since the question of whether the emergence of EBM constitutes a Kuhnian paradigm shift has no bearing on the present thesis, it will not be further discussed here.

> integrating individual clinical expertise with the best available external clinical evidence from systematic research. (Sackett *et al.*, 1996: 71)

Both quotations above, from 1992 and 1996, stress the urgency of using clinical evidence. This is all well and good, but one also needs to explain *what* should count as clinical evidence, and how evidence from many studies is supposed to be aggregated (especially when evidence is drawn from different source types). Within EBM, this is accounted for in so-called evidence hierarchies, which are discussed in the next subsection.

To aggregate results from many studies on particular diseases and/or particular interventions so that one could give, e.g., an estimate of the probability of recurrent illness, such as in the grand mal seizure example above, has traditionally been the task of an epidemiologist. Indeed, EBM has been described as an offshoot from clinical epidemiology (Vandenbroucke, 1998: S14).

4.3. The EBM evidence hierarchies

An evidence hierarchy (or: a hierarchy of evidence, which is the term used in Paper II) is an ordered list of ways in which evidence can be obtained in some matter. The higher up in the list, the stronger the evidence obtained (for answering a specified research question), it is claimed. Evidence hierarchies were suggested before the 1992 seminal EBM article. One of the earliest, though not called an evidence hierarchy at the time of publication, is found in an article by the Canadian Task Force on the Periodic Health Examination[28] (1979: 1195), which used the following list to grade "the effectiveness of intervention" for a large number of conditions:

> – Evidence obtained from at least one properly randomized controlled trial.
> – Evidence obtained from well designed cohort or case–control analytic studies, preferably from more than one centre or research group.
> – Evidence obtained from comparisons between times or places with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin in the 1940s) could also be regarded as this type of evidence.
> – Opinions of respected authorities, based on clinical experience, descriptive studies or reports of expert committees.

[28] The Canadian connection between this article and the 1992 article is not spurious: several 1979 authors were affiliated with McMaster University, including the well-known EBM authority David L. Sackett.

Another well-known hierarchy is the one provided by Straus *et al.* (2005: 169):

- Systematic review with homogeneity of RCTs
- Individual RCT with narrow confidence interval
- All or none[29]
- Systematic review (with homogeneity) of cohort studies
- Individual cohort study (including low-quality RCT; e.g. <80 % follow-up)
- Systematic review (with homogeneity) of case–control study
- Individual case–control study
- Case series (and poor quality cohort and case–control studies)
- Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles".

A final example is the following, taken from the authoritative book by Gordon Guyatt and Drummond Rennie (2002: 7):

- N of 1 randomized controlled trial
- Systematic reviews of randomized trials
- Single randomized trial
- Systematic review of observational studies addressing patient-important outcomes
- Single observational study addressing patient-important outcomes
- Physiologic studies (studies of blood pressure, cardiac output, exercise capacity, bone density, and so forth)
- Unsystematic clinical observations.

Two further hierarchies are exemplified in Paper III, Section 2.[30] All hierarchies above are claimed, explicitly or implicitly, to be useful for assessments of whether an intervention helps or not. (For other research questions, the hierarchy might have to be changed.)

All research papers in this thesis connect more or less to evidence hierarchies. Paper III does so in the most obvious way, by asking what the order relation in such a hierarchy might mean. In Paper I, the claims of some alternative medicine proponents to the effect that their methods cannot be evaluated in clinical trials are scrutinised. Since high-quality clinical trials

[29] This is explained as follows: "Met when all patients died before the treatment became available, but some now survive on it; or when some patients died before the treatment became available, but now none die of it" (Straus *et al.*, 2005: 169).

[30] And for additional examples, see Liddle, Williamson and Irwig (1996); Preventive Services Task Force (1996: 862); Medicare Services Advisory Committee (2000). These three example hierarchies are handily summarised in Grossman and MacKenzie (2005: 522).

are found at the top of EBM's evidence hierarchies, what is at stake is also the applicability of EBM ways of thinking to alternative medicine. In Paper II, mechanistic reasoning is discussed, in particular negative variants thereof. As exemplified in the hierarchies quoted above, mechanistic reasoning either resides near the bottom of the hierarchies or is not even mentioned. (What I mean by "mechanistic reasoning" is explained in Section 4.4, below.) In Paper II, it is argued that mechanistic reasoning can sometimes provide strong evidence. If this is true, then the hierarchies have to be redesigned (if they are to be retained). Papers IV and V are about the value of randomisation in clinical trials, which is instrumental in any discussion on whether RCTs should occupy a higher tier than observational studies in evidence hierarchies. The discussion about the merits of randomisation is complicated by the fact that randomisation is claimed to achieve several, rather different, things. It is also made complex partly because one has to discuss both simple and stratified randomisation. According to McEntegart (2014), more elaborate allocation procedures than simple randomisation were used in over 80 % of the clinical trials that were listed in a database with over 1500 RCTs.

One theoretical concern in the way the concept of an evidence hierarchy is being used should be mentioned. In Section 2, above, we have seen that a piece of evidence E is often connected to some hypothesis H. Even if E could perhaps *exist* without being connected to some H (for example, when a person's evidence is taken to be those beliefs of which the person is convinced that they are true), at least the *strength* of E is dependent on which H it is being related to; thus, E may be strong with respect to H_1 but weak or completely irrelevant with respect to H_2. In the literature on evidence hierarchies, however, there is little talk about hypotheses at all. The central claim about evidence hierarchies is, as we have seen, that different ways of obtaining evidence (in Paper III, such ways are called "strategies", abbreviated S) can be assigned different (generic) strengths of evidence. But what about the fact that the strength of evidence depends on which H is being considered?

In the literature on evidence hierarchies, the range of claimed validity of a hierarchy is not commonly expressed with reference to hypotheses, but there are often specifications about the validity of some hierarchy with reference to a "question" or "purpose" (as already mentioned above). For example, the OCEBM Levels of Evidence Working Group (2011) suggests one hierarchy for the question "Does this intervention help?" but another hierarchy for the question "What are the common harms?". I therefore take it that there has to be a connection between hypotheses on the one hand and questions/purposes on the other. A question, to repeat, could be "Does

this intervention help [against this disease or collection of symptoms]?". But a hypothesis is more specific, for example: *paracetamol is efficient against fever*. (Perhaps the hypothesis should be clearer about what "efficient" means, what counts as "fever", etc., but these things do not matter for the point I am trying to convey here.) The statement *ofloxacin is efficient against pneumonia* is another hypothesis, but it has the same format as the first one; namely, that some intervention is efficient against some disease or symptom(s). So one way of making the needed connection is to say roughly the following: when it is claimed that some evidence hierarchy is valid for some question or purpose, this means that it is valid for all hypotheses that can be formulated according to a (specified) template that includes the question/purpose.

I will not try to spell out this connection in greater detail here (nor in Paper III), and in fact I suspect that it is a rather difficult task. A sceptic of evidence hierarchies could perhaps argue that the wanted connection cannot be established with the precision needed, which would then deal a theoretical blow to the very idea of using evidence hierarchies. But I will not discuss the matter further here.

4.4. Critique of EBM

Evidence-based medicine has been scrutinised and criticised from various viewpoints. I will give some examples, particularly ones that connect to my research papers. Naturally, my review of criticisms is not exhaustive. For instance, I will not review critique to the effect that EBM ways of aggregating and assessing evidence do not highlight or compensate gender imbalances in the underlying empirical research (Goldenberg, 2006: 2627).

Some medical theories and schools are sceptical of EBM because of its insistence on empirical, experimental investigations of treatment claims. In Paper I this topic is addressed with reference to alternative medicine, particularly to the claim that alternative medicine cannot be tested because it is applied and varied individually. Such a claim is found to be weak: it is generally untrue that just because a treatment can be adjusted individually, it cannot be tested in a clinical trial. But of course, one could formulate a more general critique against EBM to the effect that its great confidence (perhaps overconfidence) in RCTs may lead to an unfortunate suppression or neglect of the judgments that clinicians have to make in front of real patients. RCTs report average results, but clinical medicine is applied to individuals, and this gap must be filled with clinical judgment. For example, Mark R. Tonelli writes:

> Misunderstanding the nature of EBM or failing to adequately acknowledge its limitations, however, has potentially untoward consequences. Such consequences include the devaluation of the individual, a shift in the focus of medical practice from the individual to society at large, and the failure to appreciate and cultivate the complex nature of sound clinical judgment. (Tonelli, 1998: 1237)

Other frequently voiced criticisms of EBM concern the interpretation of evidence hierarchies. For example, Borgerson (2009: 218) writes that "certain research methods in medicine are thought to be categorically better than others", and she is critical of this. (However, she does not explain the exact meaning of "categorically better".) In Section 3.1 of Paper III, several additional examples of criticisms along those lines are given. But as I also argue in Paper III, this criticism seems unfair, since there are other possible interpretations of the order relation available, and since several EBM supporters have rejected the most strict interpretations. However, the question is also complicated by the fact that whereas EBM was initially presented as an activity (and/or an attitude) that the individual clinician could practice on an everyday basis, more recently, considerable emphasis has been put on how bodies of experts can assess in an EBM way the accumulated evidence in some matter of medical treatment. These contexts are so different from one another that an evidence hierarchy (and/or its particular interpretation) may be reasonable in one setting but unreasonable in the other.

Another line of criticism is that the hierarchies are not correctly designed (whether the interpretation of the order relation has been settled or not). An overall question here is whether the balance between empirical investigations and theoretical deliberations has been correctly struck in EBM. Kelly (2018), for example, calls for a "rationalist turn" in EBM. He argues that there is an

> enormous imbalance between the amount of effort that has gone into refining certain methods like the trial, and the associated statistics and epidemiology on the one hand and the forms of reasoning so often consigned to the bottom of the evidence hierarchy on the other. There are well-defined scientific protocols for methods and interpretation of results. The methods for understanding processes of inference and judgement beyond the method protocols are less well understood or articulated, and they should be developed. (Kelly, 2018: 1164)

Rationalism is not at all absent from EBM, according to Kelly, for he takes the very evidence hierarchies, and the ideas behind their designs, to be rationalist constructs: "The orthodoxy of EBM belongs firmly in the rationalist camp. It

is an overarching set of theoretical ideas and logical principles about methods" (2018: 1163).

A more specific question about EBM evidence hierarchies is whether RCTs really are evidentially superior to observational studies, as is consistently being claimed within EBM. Quite a lot of literature has been devoted to this question. Two examples follow. Concato *et al.* (2000) have argued that if RCTs were more reliable, one would expect observational studies to exaggerate treatment effects, but when RCTs and observational studies that assessed the same intervention have been compared, the average results have turned out to be very similar. In response to this and similar claims, Howick (2011: 39–62) has insisted that randomisation reduces more biases than observational studies. In addition, he denies that it is a general truth that RCTs and observational studies produce similar results. Again, however, I would like to notify the reader that in Paper III, I argue that the plausibility of saying that RCTs are evidentially superior to observational studies depends on which interpretation of the order relation is chosen. Perhaps RCTs provide stronger evidence than observational studies according to one interpretation, but not according to another. Specifying which interpretation of the order relation is being discussed does not, of course, make the RCTs vs. observational studies debate disappear, but could probably improve its quality.[31]

The low reliance on mechanistic reasoning within EBM has prompted a lot of debate. First I should explain roughly what is meant by mechanistic reasoning. Here is an example:

> The drug D contains the substance S_1 which, when disseminated in the body, gives rise to the reaction R_1, which in turn yields the reaction R_2, releasing substance S_2, which has a well-known inhibitory effect on the symptom Y. Consequently, intake of drug D will alleviate symptom Y.

Mechanistic reasoning typically consists of a chain of linked actions, eventually reaching the effect of interest. The chain is usually thought of as being causal. Normally, in each step of the chain there is a probability > 0 that the next step won't follow. If for no other reason, this is true since the human body is such a complex entity. Hence, there is always some uncertainty as to whether the chain is correct.

Within EBM there is a strong reliance on methods that foremost tell us *what* works, or *that* a treatment works. In order to show that something works, one does not need complete knowledge about the chain of action from

[31] The debate on the merits of RCTs vs. observational studies is also related to the question of how to justify randomisation; see Section 3.3 and Papers IV–V.

treatment to effect. But mechanistic reasoning is mostly about this chain. In this sense, mechanistic reasoning is somewhat detached from the main goal in EBM: assessing which treatments are beneficial and which ones are not. It has even been suggested that EBM started partly as "a reaction to what was perceived as a fairly widespread failure of mechanisms as evidence in clinical medical practice" (Andersen, 2012: 992).

When EBM adherents dismiss mechanistic reasoning, an important argument seems to be to show that mechanistic reasoning has gone astray on many occasions. Howick (2011: 154–157) presents a list of erroneous conclusions based on mechanistic reasoning from which the following examples are sampled:

- Antiarrhythmic drugs ought to reduce mortality due to sudden cardiac death according to mechanistic reasoning; trials have shown antiarrhythmic drugs to increase the mortality in question.
- Hormone replacement therapy ought to reduce menopausal symptoms according to mechanistic reasoning; comparative studies have shown the opposite.
- According to mechanistic reasoning, vitamin E reduces the risk of coronary heart disease and atherosclerosis; comparative studies have revealed no effect.

On a more principled level, mechanistic reasoning could be viewed with suspicion by EBM supporters because it often requires rather profound knowledge about theories and other (purported) generalisations from large amounts of empirical data. As mentioned above, some EBM literature asserts that normal practitioners of medicine (doctors, nurses, etc.) should be able to apply the EBM approach in their everyday clinical reality. This was stressed, for example, in the founding text of EBM (Evidence-Based Medicine Working Group, 1992). Mechanistic reasoning, being too specialised, does not fit this practically oriented perspective. Mechanistic reasoning also has an undeniable connection to expert opinion, which is met with equal scepticism in the EBM literature.

Still, it is not entirely clear how to move from an overall interest in what works and what is feasible in clinical practice to arguments against mechanistic reasoning. Perhaps mechanistic reasoning could be really helpful for finding out what works, but practitioners' ignorance impedes its use. A supporter of the standard EBM tenets could perhaps reply: I admit that mechanistic reasoning can (sometimes) be somewhat helpful, but the evidence

for or against some treatment emanating from mechanistic reasoning is always weaker than that provided by empirical studies. This seems to be the position taken by Howick, Glasziou and Aronson (2009: 189): "[A]lthough we believe that mechanistic evidence cannot be ignored, we acknowledge that mechanistic evidence should always play a subsidiary confirmatory role *vis-à-vis* direct evidence". I personally take this judgment to be problematic, but I will not discuss the matter further here.

5. Summaries of appended research papers

5.1. Paper I

Adherents of alternative medicine sometimes claim that their methods differ from the methods of established health care in that the former cannot be evaluated scientifically. In Paper I treatment requirements that have to be fulfilled for a scientific evaluation to be possible (in clinical trials or similar arrangements) are investigated. The requirements discussed concern the *treatment*, as opposed to requirements that (rightly) would pertain to the participants and their behaviour, to the correct handling of collected data, to the reliability of measurement equipment, *et cetera*.

Two requirements are presented and discussed in relation to a model situation in which two treatments *A* och *B* are to be compared in a clinical trial. The first requirement says that for each participant it must be possible to tell which treatment has been given. This *distinguishing criterion* (DC) is, somewhat simplified:

> For each patient involved in the trial, one must be able to tell, with the aid of a criterion formulated before the commencement of the trial, whether treatment *A* or *B* was given, using any available information recorded before or during the trial.

In addition, it is required that no participant is given both treatments *A* and *B*.

The second requirement is the *elimination of confounding variables* (ECV):

> There must be no variable present in the trial such that (i) there is a systematic discrepancy between the groups receiving treatments *A* and *B* in this variable, (ii) the health endpoint records in the groups receiving *A* and *B* have been substantially affected by the variable, and (iii) the variable is not part of the criterion in DC.

In addition, any treatment effect must be such that it does not regularly disappear when included in a trial.

Several misunderstandings concerning the relation between alternative medicine and science, or more generally concerning how to evaluate the effects of medical treatments, can be construed as violations of the two requirements. One example is the claim that it is impossible to scientifically evaluate treatment methods that are founded on "spiritual" or otherwise unscientific views and theories. Another example is the claim that individually adjusted, and hence non-standardised, treatments cannot be tested scientifically. In neither case can any support be gained from the requirements that I have formulated.

The presented requirements are in conjunction intended to be *sufficient* conditions for scientific testing. In other words, if both requirements are fulfilled for a given treatment method, then the effectiveness of the method can be evaluated in a scientifically acceptable treatment experiment (a clinical trial). However, the requirements are not *necessary*: there may well be treatment methods that can be scientifically evaluated without fulfilling the requirements. Admittedly, there may be treatment methods that do not fulfill both requirements but are nevertheless scientifically testable, and these cannot be diagnosed using sufficient conditions. On the other hand, there are two advantages with formulating sufficient conditions. First, sufficient conditions are arguably more useful in practice, for we may now tell the adherents of alternative medicine: look, if your treatment fulfills these two criteria, it is scientifically evaluable. With necessary conditions our message to alternative medicine adherents would be the less interesting: if your treatment method is scientifically evaluable, then it fulfills these two requirements.[32] Secondly, anyone who challenge my requirements as insufficient will most likely consider them as too lax. In other words, the critic would probably take a harder attitude towards alternative medicine than what follows from my requirements. My requirements are thus generous: it is enough to fulfill these modest conditions for a treatment method to be scientifically evaluable.

The overall picture formed by my two requirements is *inclusive* with regard to the scientific evaluation of medical treatments: it is not difficult to fulfill them, and my impression is that the vast majority of alternative medicine treatments that have any measurable popularity do fulfill them. There is thus no formal reason that precludes RCT testing. To the extent that the proponents of alternative treatment methods dismiss the results of RCT

[32] The latter message would be useful if alternative medicine supporters that claim to use scientifically evaluable methods. The argumentation in Paper I is more useful in relation to the opposite attitude, which seems to me to be much more frequent.

tests of effectiveness, their arguments have to be more sophisticated than simply claiming that a scientific scrutiny is not possible.[33]

5.2. Paper II

Paper II is about mechanistic reasoning and its evidential value in the evaluation of medical treatments. In this paper, I try to do several things. First, I discuss how to define mechanistic reasoning in the context of interest, i.e., medical intervention research. Departing from a definition given by Howick, Glasziou, and Aronson (2010) I try to justify several changes and propose the following:

> Mechanistic reasoning is reasoning that involves either an inference from mechanistic chains to claims concerning specified intervention outcomes, or an inference from an investigation of whether there are plausible mechanistic chains to claims concerning specified intervention outcomes.

This means that mechanistic reasoning either includes a mechanistic chain or includes an argument concerning whether there can be a mechanistic chain. This definition is wider than a definition that requires a mechanistic chain to be included. For example, the following is mechanistic reasoning according to my definition, although no description of a mechanistic chain is included:

> A headset with light-emitting diodes is to be used for "channeling bright light directly to photosensitive regions of the brain through the ear canal" for a few minutes per day. This is claimed to be an efficient treatment of mood swings due to, e.g., seasonal daylight variations. However, according to current physiological knowledge there are no photoreceptors in the ear canal, and hence any light entered there can only be mediated as heat to other parts of the head. There is no conceivable way in which

[33] Two afterthoughts on Paper I shall be offered here. (1) In Section 1 of the paper, there is a formulation to the effect that we will investigate what conditions medical treatments "must fulfil" in order to be scientifically evaluable. In Section 4, it is claimed that treatments "must" be tested on a large number of people. Taken literally, both formulations are exaggerated, for one could imagine a situation where the result from a case study would allow (a sufficient degree of) generalisation to other patients. (2) In Section 10, it is suggested that blinding and randomisation are not necessary methodological steps in scientific investigations but are "recommendable practices to the extent that they contribute to satisfying DC+EMT and ECV+CS". One could get the impression that DC+EMT and ECV+CS are ends that have to be satisfied in any empirical scientific investigation (although I also mention, in the same paragraph, the possibility of "other implicit methodological conditions"). I just would like to emphasise here that it is difficult to establish the most fundamental ends of an empirical scientific investigation.

> small amounts of heat in the ear canal could causally decrease the frequency or severity of mood swings. In conclusion, there cannot be a mechanistic connection between intervention and postulated outcome in this case.

Secondly, I characterise different types of *negative* mechanistic reasoning, which have been rather neglected in the literature. Mechanistic reasoning is negative in a health-related sense if it suggests an outcome that is bad to the patient (or at least that is not positive although a positive outcome was expected). Mechanistic reasoning is negative in an epistemic sense if it does not include a mechanistic chain. Three main types of negative mechanistic reasoning, labelled NegA, NegB, and NegC, are presented and are characterised as follows:

NegA	Reasoning which includes a mechanistic chain, suggesting a negative outcome to the patient (or a neutral outcome against a background expectation of a positive outcome).
NegB	Reasoning that constitutes a serious but failed attempt to find a mechanistic chain connecting the intervention and the outcome.
NegC	Reasoning in which meta-mechanistic arguments suggest that there cannot be a mechanistic connection between intervention and outcome.

NegA is negative in the health-related sense. NegB and NegC are epistemically negative, but NegB includes an attempt to find a mechanistic chain, whereas NegC is "meta-mechanistic", i.e., one does not look for mechanistic chains but investigates whether such chains are possible and concludes that they are not.

Each type is associated with a range of evidential strength. Some general differences emerge. NegA is a mirror image of positive mechanistic reasoning. Just as positive mechanistic reasoning can be, under certain circumstances, very reliable or, under other circumstances, very unreliable, the evidential strength of NegA reasoning is highly variable. NegB typically carries low strength of evidence, the main reason being that failing to find a mechanistic chain does not at all guarantee that there is none. In NegC reasoning, one appeals to previous knowledge which in many cases may be considered to have a high degree of certainty. Therefore, the strength of evidence associated with NegC reasoning is rather great or even very great. I present several

examples of negative mechanistic reasoning according to types NegA, NegB, and NegC.

Thirdly, I use my definition of mechanistic reasoning and the characterisations of its negative variants to argue that proponents of EBM have dismissed mechanistic reasoning too quickly. Sometimes mechanistic reasoning carries great evidential strength. In certain EBM literature one will even find that "evidence" has been defined in such a way that mechanistic reasoning does not count as evidence at all. That is unreasonable. Mechanistic reasoning can be unreliable. That is a good reason for being sceptical as a default attitude, but my analysis shows that some types of mechanistic reasoning are reasonably reliable, and this should be acknowledged and incorporated into EBM.[34]

5.3. Paper III

Evidence hierarchies seem to be almost unique to EBM: they are not used in many other fields, and have probably never been adopted in any natural science. An evidence hierarchy is an ordered list of "investigative strategies" (as I suggest we call them, since not all items found in real evidence hierarchies are "study types" or "study designs", even though most of them are). A curiously under-researched question is what the order relation of such a hierarchy means. I distinguish four main interpretations.

The first is called (NON-OVERLAPPING STRONGER). According to this interpretation, if the investigative strategy S_1 is above strategy S_2 in the hierarchy, then any investigation using S_1 provides stronger evidence than any investigation using S_2. In other words, their evidential ranges do not overlap; hence the suggested name of the interpretation. This is a very strong interpretation, and therefore an implausible one in the context of evaluating medical treatments against some disease or collection of symptoms. For example, if RCTs are above observational studies in the hierarchy (which is normally the case in EBM), then according to the (NON-OVERLAPPING STRONGER)

[34] Two afterthoughts to Paper II are the following. (1) In Section 2, Scenario LED is presented as an instance of meta-mechanistic reasoning to the effect that there cannot be a mechanism present. This is correct, but the certainty of such an assertion is variable. One link of the Scenario LED reasoning is that "current physiological knowledge" says that there are no photoreceptors in the ear canal. I am not aware of any principle or "law" of physiology that would prohibit the existence of such photoreceptors. Had there been such a principle, then, arguably, the reasoning would have been even more certain. Meta-mechanistic reasoning directed against homeopathy (mentioned in Section 4 of Paper II) often makes reference to principles that are taken to be more certain than just empirical investigations of some bodily structure. This variability of meta-mechanistic reasoning could have been more clearly acknowledged. (2) In Section 2, there is a formulation about "purely empirical" investigations. The phrase is unfortunate since, arguably, investigations cannot be exclusively empirical (i.e., involving no theory). However, the phrase is not crucial for the reasoning in which it appears; in fact, it could be omitted.

interpretation no observational study, no matter how well performed, can provide stronger evidence than an RCT that tests the same treatment–disease coupling. The common understanding of different investigative strategies (e.g., RCTs vs. observational studies) is, however, that they can be more or less credible, more or less prone to bias, and this makes it difficult to argue conclusively that the ranges of evidential strengths cannot overlap. Nonetheless, I give several examples where scholars (mainly critics of EBM) seem to take for granted that this interpretation must be what EBM defenders support.

The second interpretation is called (STRONGER CP), meaning "stronger *ceteris paribus*". It means this: if S_1 is above S_2 in the hierarchy, then an investigation using S_1 yields stronger evidence than an investigation using S_2 if they are "alike" or "comparable" in all relevant respects except for the fact that they instantiate different strategies. The problem with this interpretation is that it is difficult to make sense of the likeness or comparability just mentioned. Let us assume that RCTs are above mechanistic reasoning in the hierarchy. These two investigative strategies are so different from one another that it is difficult to understand what a *ceteris paribus* condition would even mean.

The third interpretation is (TYPICALLY STRONGER). According to this interpretation, S_1 being above S_2 in the hierarchy means that the typical (or average, or median) investigation using S_1 provides stronger evidence than the typical (average, median) investigation using S_2. The fourth interpretation is (IDEALLY STRONGER): S_1 being above S_2 in the hierarchy means that an ideally performed S_1 investigation (i.e., an investigation of perfect quality) provides stronger evidence than an ideally performed S_2 investigation. A weaker but perhaps more useful variant of the latter is (STRONGER IF GOOD ENOUGH): a high-quality (but not necessarily perfect) S_1 investigation provides stronger evidence than any S_2 investigation. The (TYPICALLY STRONGER) and (IDEALLY STRONGER) interpretations—including the (STRONGER IF GOOD ENOUGH) variant of the latter—are more credible in the context of evaluating medical treatments than the (NON-OVERLAPPING STRONGER) or (STRONGER CP) interpretations. The GRADE framework, in which there is a two-item hierarchy with RCTs above observational studies—does not say which interpretation is the intended one, but I argue that the (TYPICALLY STRONGER) interpretation is the best fit for GRADE, as GRADE writings imply that RCTs are taken to provide stronger evidence than observational studies in a general way, not just restricted to some qualitative top tier of investigations.

Even if it would be established which interpretation is correct, unfortunately very little follows that is helpful for the complex task of aggregating evidence from different tiers in the hierarchy. This still does not entail that a hierarchy is completely unhelpful in the practice of evaluating the overall evi-

dence in some matter. It could be appealed to for assessing the need for more research using particular strategies, for example. And generally, I believe that specifying which interpretation is taken to be correct could contribute to progress in some debates surrounding EBM. Obviously, specifications should be supplied both by defenders and by critics of EBM and of evidence hierarchies.

5.4. Paper IV

Arguments for randomisation in clinical trials fall into two categories: theoretical and practical arguments. Those that I call theoretical are inseparably connected to theoretical concepts. The practical ones are about the decrease of one or more biases, without any necessary connection to theoretical concepts. In this paper, the theoretical arguments are assessed. There are three main theoretical arguments for randomisation:

(1) Randomisation is needed for performing statistical null hypothesis testing.
(2) Randomisation is needed for drawing (plausible) causal inferences from treatment to effect.
(3) Randomisation is an acceptable and convenient way of achieving prior distributions of covariates in a Bayesian framework.

These arguments are theoretical because they refer to the concepts of null hypothesis testing, causal inference, and Bayesian reasoning, respectively. In assessing these arguments, I try to use two guiding principles, namely (*a*) imagining a concrete setting of a planned, realistic trial, and (*b*) keeping in mind that the randomness inherent in randomisation ought to matter for the argument.

Spelling out (1) in greater detail, there seem to be two sub-arguments that are entirely theoretical, and that could lead to (1). Both are found in Ronald Fisher, the foremost pioneer of null hypothesis testing. The first sub-argument is that randomisation makes valid, even after any stratification has been performed, a model according to which all remaining errors (i.e., those errors that remain after any balancing measures have been carried out) emanate from a single probability distribution. The other sub-argument is that randomisation transforms systematic errors into random errors. I approve of the first argument but disapprove of the second: randomisation could make some systematic errors less likely to be present, for example by circumscribing the possibility of human pattern recognition, but yields no guarantee that there are no disturbing variables present. Critique from

Isaac Levi, Teddy Seidenfeld, and others is discussed but is found not to be decisive against Fisher. However, (1) is worded a little too strongly. I suggest the following modified version: Randomisation contributes to making null hypothesis testing assumptions valid.

Argument (2) has received support from Nancy Cartwright and David Papineau, among others. I find (2) to be a weak argument, if (*a*) and (*b*) are kept in mind: though it cannot be denied generally that randomisation may facilitate causal inferences, in realistic clinical trials, randomisation does not convey anything near a guarantee that changes in the outcome must have been caused by the treatment. I also argue—contrary to what Cartwright has claimed—that RCTs typically are not performed in order to establish a causal connection (an epistemic end) but rather to investigate whether the treatment works (a practical or action-guiding end).

As for (3), most Bayesians have been very sceptical of randomisation. Still, some have said that randomisation can be accepted in the Bayesian paradigm. Even so, full and formal Bayesian analyses of clinical trial outcomes are rarely asked for, which makes it rather difficult to come up with a generally supportable formulation of (3) in the context of clinical trials.

In sum, only the modified version of (1) survives the present scrutiny of the theoretical arguments in favour of randomisation in the context of clinical trials.

5.5. Paper V

Practical arguments for randomisation in clinical trials are arguments according to which randomisation reduces (or even removes) some unwanted bias. I have identified four main arguments of this type:

(1) Randomisation contributes to allocation concealment.
(2) Randomisation contributes to the baseline balance of groups.
(3) Randomisation removes self-selection bias, i.e., bias created from trial participants making, or at least affecting, choices of treatment.
(4) Randomisation removes allocation bias, i.e., bias created when those that run a trial make, or at least affect, choices of treatment for the participants.

I scrutinise arguments (1)–(4) using the same guiding principles (*a*) and (*b*) as in Paper IV.

As for (1), allocation concealment denotes the process and precautions for keeping secret which patients are to be allocated to which treatment. The secrecy is normally extended both to the participants and to the clinicians that

will interact with the participants. I find this argument for randomisation to be weak, the main reason being that the randomness in randomisation is not crucial: even if there may be situations where the fact that some sequence was generated randomly gives some protection from illegitimate dissemination (for example by being more difficult to memorise), it is, generally speaking, much more important for the integrity of the sequence that it is handled according to a worked-out instructions than that it has been created by a random process. Such instructions could be identical no matter how the sequence was generated.

Argument (2) seems to be very common. In discussing it, one has to consider both simple randomisation and stratified randomisation. By simple randomisation we mean a process where each participant is assigned to a treatment group using a random process, without having been ordered or grouped in any way beforehand. In such a scheme, every allocation is probabilistically independent from any other. Stratified randomisation, on the other hand, is a two-step process. First, the participants are subject to some grouping with the intent that those groups (*strata*) are similar with respect to one or several confounders. This step is the stratification, and the confounders selected for grouping in this step may be called stratification variables. The second step is the randomisation proper, in which the participants are assigned to the final treatment groups using some random mechanism in such a way that fixed proportions of participants are assigned to each treatment group from each stratum. The proportions are fixed in the sense that they are identical across strata.

When will the balance in the treatment groups of an arbitrary confounding variable V be better from randomisation than from no randomisation? It depends on whether we are looking at simple or stratified randomisation; and the feasibility and effectiveness of the latter depend, in turn, on whether V is known not only generally but in each participant; and, in case V is not known in each participant, the feasibility and effectiveness of stratified randomisation depend also on whether V is correlated with some other confounder W that *is* known individually (and hence could be used for stratification). In this rather thorny matter, I reason as follows:

- If V is known so that stratification is possible, then stratification gives a balance that is guaranteed to be better than what can be expected from simple randomisation.

- If V is known generally but not in every participant, then if V is correlated with some other confounder W that is known in the individual participants, then stratification with respect to W (whether or not

followed by randomisation within strata) is likely to give better balance with respect to V than would have been achieved through simple randomisation.

– If V is not known, or if V is known but not individually and V is not correlated with some other confounder that can be stratified, then we cannot balance V better than what is expected from simple randomisation.

All of this leads to the conclusion that (2) is a rather weak argument for randomisation, as compared to, or combined with, stratification. Simple randomisation certainly contributes to the balancing of treatment groups compared to non-random allocation procedures, but only, as far as I can see, by way of decreasing allocation bias and/or self-selection bias.

Argument (3) is judged to be tenable: the randomness in randomisation ensures that no self-selection bias can occur in the treatment allocation process. It has been argued that the performance of an *interventional* study—as opposed to a non-interventional study—is the important factor that rules out self-selection bias to occur, rather than the factor of randomisation (as opposed to non-randomisation). I do not agree. While I grant that self-selection bias is typically small or absent in an interventional study, the bias is not *guaranteed* to be absent just because the study is interventional.

The allocation bias argument for randomisation (4) appears to be strong: it is true that randomisation eliminates any possibility of allocation bias in a trial, and it is also true that realistic alternatives to randomised allocation typically involve a substantial risk of allocation bias.

Arguments (3) and (4) are thus approved. Out of these, (4) appears to be the strongest. Arguments (1) and (2) are rejected the way they have been delineated here.

References

Achinstein, Peter (2001). *The Book of Evidence*. Oxford: Oxford University Press.

Achinstein, Peter (2004). A Challenge to Positive Relevance Theorists: Reply to Roush, *Philosophy of Science* 71, 521–524.

Ackerknecht, Erwin H. (1967). *Medicine at the Paris Hospital 1794–1848*, Baltimore, MD: Johns Hopkins University Press.

Andersen, Holly (2012). Mechanisms: what are they evidence for in evidence-based medicine? *Journal of Evaluation in Clinical Practice* 18(5), 992–999.

Archard, David (2017). The Methodology of Applied Philosophy. In Kasper Lippert-Rasmussen, Kimberly Brownlee, and David Coady (eds), *A Companion to Ap-*

plied Philosophy, 1st edition, Chichester, UK and Hoboken, NJ: John Wiley & Sons, 18–33.

Baron, Jeremy Hugh (2009). Sailors' scurvy before and after James Lind—a reassessment, *Nutrition Reviews* 67(6), 315–332.

Belsey, Andrew (1995). Hypothesis. In Ted Honderich (ed.), *The Oxford Companion to Philosophy*, Oxford: Oxford University Press, 385.

Borgerson, Kirstin (2009). Valuing evidence: Bias and the evidence hierarchy of evidence-based medicine, *Perspectives in Biology and Medicine* 52(2), 218–233.

Boring, Edwin G. (1954). The nature and history of experimental control, *American Journal of Psychology* 67(4), 573–589.

Boylston, Arthur W. (2010). Thomas Nettleton and the dawn of quantitative assessments of the effects of medical interventions, *Journal of the Royal Society of Medicine* 103, 335–339.

Broadbent, Alex (2019). *Philosophy of Medicine*. Oxford: Oxford University Press.

Canadian Task Force on the Periodic Health Examination (1979). Task Force Report: The periodic health examination, *Canadian Medical Association Journal* 121, 1193–1254.

Carnap, Rudolf (1947). On the application of inductive logic, *Philosophy and Phenomenological Research* 8, 133–148.

Chalmers, Iain, Estela Dukan, Scott H. Podolsky, and George Davey Smith (2012). The advent of fair treatment allocation schedules in clinical trials during the 19th and early 20th centuries, *Journal of the Royal Society of Medicine* 105(5), 221–227.

Concato, John, Nirav Shah, and Ralph I. Horwitz (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs, *The New England Journal of Medicine* 342(25), 1887–1892.

Crupi, Vincenzo (2020). Confirmation. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). Available at https://plato.stanford.edu/archives/spr2020/entries/confirmation/.

Curd, Martin and J. A. Cover (1998). *Philosophy of Science: The Central Issues*. New York: W. W. Norton.

Doll, Richard (1998). Controlled trials: the 1948 watershed, *British Medical Journal* 317, 1217–1220.

Donaldson, I. M. L. (2016). van Helmont's proposal for a randomised comparison of treating fevers with or without bloodletting and purging. *JLL Bulletin: Commentaries on the history of treatment evaluation*. Available at www.jameslindlibrary.org/articles/. Accessed December 23, 2020.

Drake, Stillman (1992). Music and Philosophy in Early Modern Science. In Victor Coelho (ed.), *Music and Science in the Age of Galileo*, Dordrecht: Kluwer, 3–16.

Eckhoff, Torstein (1989). Lotteries in allocative situations, *Social Science Information* 28(1), 5–22.

Evidence-Based Medicine Working Group (1992). Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine, *Journal of the American Medical Association* 268(17), 2420–2425.

Fisher, Ronald A. (1925). *Statistical Methods for Research Workers*, 1st edition. Edinburgh: Oliver & Boyd.

Fisher, Ronald A. (1935). *The Design of Experiments*, 1st edition. Edinburgh: Oliver & Boyd.

Franklin, Allan (1990). *Experiment, Right or Wrong*. Cambridge: Cambridge University Press.

Glymour, Clark (1980). *Theory and Evidence*. Princeton, NJ: Princeton University Press.

Goldenberg, Maya J. (2006). On evidence and evidence-based medicine: Lessons from the philosophy of science, *Social Science & Medicine* 62, 2621–2632.

Grossman, Jason and Fiona J. MacKenzie (2005). The Randomized Controlled Trial: gold standard, or merely standard? *Perspectives in Biology and Medicine* 48(4), 516–534.

Guyatt, Gordon and Drummond Rennie (eds) (2002). *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Chicago: American Medical Association Press.

Hacking, Ian (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

Hacking, Ian (1988). Telepathy: origins of randomization in experimental design, *Isis* 79, 427–451.

Hájek, Alan (2008). Arguments for—or against—Probabilism? *British Journal for the Philosophy of Science* 59, 793–819.

Hájek, Alan (2019). Interpretations of probability. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition). Available at https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/.

Hansson, Sven Ove (2015). Experiments Before Science: What Science Learned from Technological Experiments. In Sven Ove Hansson (ed.), *The Role of Technology in Science: Philosophical Perspectives*. Berlin/Heidelberg: Springer, 81–110.

Helmont, Johannes Baptista van (1648). *Ortus medicinæ: Id est Initia physicæ inaudita. Progressus medicinae novus, in morborum ultionem, ad vitam longam*. Amsterdam: Apud Ludovicum Elzevirium.

Howick, Jeremy (2011). *The Philosophy of Evidence-Based Medicine*. Chichester: Wiley Blackwell & BMJ Books.

Howick, Jeremy, Paul Glasziou, and Jeffrey K. Aronson (2009). The evolution of evidence hierarchies: what can Bradford Hill's 'guidelines for causation' contribute? *Journal of the Royal Society of Medicine* 102, 186–194.

Hróbjartsson, Asbjørn, Peter C. Gøtzsche, and Christian Gluud (1998). The controlled clinical trial turns 100 years: Fibiger's trial of serum treatment of diphtheria, *British Medical Journal* 317, 1243–1245.

Huth, Edward (2006). Quantitative evidence for judgments on the efficacy of inoculation for the prevention of smallpox: England and New England in the 1700s, *Journal of Royal Society of Medicine* 99, 262–266.

Ichikawa, Jonathan Jenkins and Matthias Steup (2018). The Analysis of Knowledge. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), available at https://plato.stanford.edu/archives/sum2018/entries/knowledge-analysis/.

James Lind Library (no date). Facsimile and translation (by I. M. L. Donaldson) of extract from Helmont (1648), www.jameslindlibrary.org/van-helmont-jb-1648/. Accessed December 23, 2020.

Jerkert, Jesper (2019). *Science in Theory and Practice: An Introductory Survey*. Stockholm: Division of Philosophy, KTH Royal Institute of Technology.

Kaptchuk, Ted J. (1998). Intentional Ignorance: A History of Blind Assessment and Placebo Controls in Medicine, *Bulletin of the History of Medicine* 72(3), 389–433.

Kelly, Michael P. (2018). The need for a rationalist turn in evidence-based medicine, *Journal of Evaluation in Clinical Practice* 24, 1158–1165.

Kelly, Thomas (2016). Evidence. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). Available at https://plato.stanford.edu/archives/win2016/entries/evidence/.

Keynes, John Maynard (1921). *A Treatise on Probability*. London: Macmillan.

Liddle, Jeannine, Margaret Williamson and Les Irwig (1996). *Method for Evaluating Research and Guideline Evidence*. Sydney: New South Wales Department of Health.

Lipton, Peter (2004). *Inference to the Best Explanation*, 2nd edition. London and New York: Routledge.

Lopez, Claude-Anne (1993). Franklin and Mesmer: An Encounter, *Yale Journal of Biology and Medicine* 66, 325–331.

McEntegart, Damian (2014). Block Randomization. In N. Balakrishnan (ed.), *Methods and Applications of Statistics in Clinical Trials: Concepts, Principles, Trials, and Design, Volume 1*. Hoboken, NJ: John Wiley & Sons, 125–138.

Medicare Services Advisory Committee (2000). *Funding for New Medical Technologies and Procedures: Application and Assessment Guidelines*. Canberra: AusInfo.

Naylor, R. H. (1980). The Role of Experiment in Galileo's Early Work on the Law of Fall, *Annals of Science* 37, 363–378.

OCEBM Levels of Evidence Working Group (2011). The Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence, available at www.cebm.net/wp-content/uploads/2014/06/CEBM-Levels-of-Evidence-2.1.pdf. (Accessed August 29, 2020.)

Porter, Roy (1998). *The Greatest Benefit to Mankind: A Medical History of Humanity*, 1st American edition. New York: W. W. Norton.

Preventive Services Task Force (1996). Guide to Clinical Preventive Services: Report of the US Preventive Services Task Force, 2nd edition. Baltimore: Williams & Wilkins.

Riess, Falk, Peter Heering, and Dennis Nawrath (2005). Reconstructing Galileo's Inclined Plane Experiments for Teaching Purposes, *Proceedings of the International History, Philosophy, Sociology and Science Teaching Conference*, Leeds.

Roush, Sherrilyn (2004). Discussion Note: Positive Relevance Defended, *Philosophy of Science* 71, 110–116.

Sackett, David L., William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes, and W. Scott Richardson (1996). Evidence based medicine: what it is and what it isn't, *British Medical Journal* 312, 71–72.

Snyder, Laura J. (1994). Is Evidence Historical? In Peter Achinstein and Laura J. Snyder (eds.), *Scientific Methods: Conceptual and Historical Problems*, Malabar, FL: Krieger Publishing Company, 95–117.

Stolberg, Michael (2006). Inventing the randomized double-blind trial: the Nuremberg salt test of 1835, *Journal of the Royal Society of Medicine* 99, 642–643.

Straus, Sharon E., W. Scott Richardson, Paul Glasziou, and R. Brian Haynes (2005). *Evidence-Based Medicine: How to Practice and Teach EBM*, 3rd edition. Edinburgh: Elsevier.

Tonelli, Mark R. (1998). The philosophical limits of evidence based medicine, *Academic Medicine* 73(12), 1234–1240.

Uddenberg, Nils (2015). *Lidande & läkedom. I: Medicinens historia fram till 1800*. Lidingö: Fri tanke.

Vandenbroucke, Jan P. (1996). Evidence-Based Medicine and "Médecine d'Observation", *Journal of Clinical Epidemiology* 49(12), 1335–1338.

Vandenbroucke, Jan P. (1998). Clinical investigation in the 20th century: the ascendancy of numerical reasoning, *The Lancet* 352 (Suppl.), s12–s16.

Winsberg, Eric (2010). *Science in the Age of Computer Simulation*. Chicago: Chicago University Press.

Wootton, David (2007). *Bad Medicine: Doctors Doing Harm Since Hippocrates*. Oxford: Oxford University Press.

Sammanfattning på svenska (Summary in Swedish)

Denna avhandling består av en introduktion och fem artiklar. Det övergripande ämnet kan sägas vara vetenskapliga metoders tillförlitlighet. Mer specifikt handlar arbetet om vissa metodanknutna frågor inom klinisk medicinsk forskning. Huvudfrågorna i de fem artiklarna är:

- Vilka villkor måste uppfyllas för att en medicinsk behandlingsmetods effektivitet ska kunna utvärderas vetenskapligt?
- Vilken är mekanistiska resonemangs rättmätiga roll i vetenskapliga utvärderingar av medicinska behandlingars effektivitet?
- Vad betyder ordningsrelationen i evidenshierarkier som används vid bedömning av evidensläget för medicinska behandlingsmetoder?
- Vilka teoretiska argument för randomisering i kliniska prövningar är hållbara?
- Vilka praktiska argument för randomisering i kliniska prövningar är hållbara?

För att förstå frågornas relevans och sätta in dem i sitt sammanhang kan en kort bakgrundsteckning behövas. Därefter sammanfattar jag var och en av artiklarna. En betydligt fylligare bakgrundsteckning finns i den engelskspråkiga introduktionen.

Bakgrund

Empiriska undersökningar av medicinska behandlingsmetoders effektivitet har förekommit under flera hundra år. Det kanske mest berömda historiska exemplet är James Linds (1716–1794) skörbjuggsexperiment ombord på *HMS Salisbury* 1747. Lind valde ut tolv sjömän som var drabbade av sjukdomen. De delades in i sex grupper à två personer. Grupperna fick olika behandlingar. Gruppen med sjömän som fick två apelsiner och en citron vardera per dag återhämtade sig påtagligt från sjukdomen.

Trots historiska exempel som Linds, är det ett faktum att medicinska behandlingsmetoder under lång tid i mycket ringa utsträckning utsattes för empiriska tester, trots att det inte är metodologiskt särskilt svårt att åstadkomma tester av acceptabel kvalité. Historikern David Wootton har rentav hävdat att avsaknaden av tester utgör en stor gåta i medicinens historia.

Nuförtiden, och grovt sett ungefär sedan andra världskriget, råder det dock stor enighet om att det är viktigt att empiriskt testa medicinska behandlingsmetoders effektivitet, och att det bästa sättet att göra detta på är i form av s.k. kliniska prövningar. En klinisk prövning är ett experiment. Patienter med samma sjukdom (eller åtminstone samma symptom) delas in i två eller flera grupper. En grupp får den nya behandling vars effekt man är särskilt intresserad av. Den andra gruppen (eller de andra grupperna) får någon annan (etablerad) behandling, ingen behandling eller en behandling som till det yttre liknar den intressanta behandlingen men som är fysiologiskt verkningsfri (placebo). Man mäter förutbestämda hälsoutfall hos alla deltagare och jämför utfallen på gruppnivå med statistiska metoder.

Evidensbaserad medicin (*evidence-based medicine*, EBM) är en rörelse som sedan huvudlanseringen 1992 uppnått en dominerande ställning när det gäller utvärdering av medicinska behandlingsmetoder. Inom EBM sätter man stor tilltro till empiriska metoder och liten tilltro till teoretiska överväganden och till auktoriteter. Dessa attityder framgår tydligt i s.k. evidenshierarkier (*evidence hierarchies* eller *hierarchies of evidence*) inom EBM. En evidenshierarki är en lista över undersökningsmetoder upptagna i fallande trovärdighetsordning när det gäller att utvärdera effekten hos en ny behandlingsmetod. En av de mest spridda evidenshierarkierna, framlagd av OCEBM Levels of Evidence Working Group, ser ut så här (min översättning):

- Systematisk översikt över randomiserade studier eller $n = 1$-studier[1]
- Randomiserad studie eller observationsstudie med dramatisk effekt
- Icke-randomiserad kontrollerad kohort- eller uppföljningsstudie
- Fallserier, fall–kontroll-studier eller historiskt kontrollerade studier
- Mekanismbaserade resonemang.

Mycket av den kritik som har riktats mot EBM har handlat om evidenshierarkierna, vilket är naturligt då dessa uppfattas som centrala för förståelsen av EBM. En typ av kritik går ut på att experimentella studier inte bör sättas kategoriskt överst eftersom inte alla medicinska behandlingsmetoder kan

[1] I en $n = 1$-studie deltar endast en försöksperson, som utsätts för olika experimentella betingelser (behandlingar) i en förutbestämd ordning.

studeras på det sättet. Inte minst från alternativmedicinskt håll har sådan kritik framförts, vilket närmare behandlas i artikel I.

En annan kritik rör balansen mellan empiriska studier och teoretiska överväganden, där EBM hittills tämligen kategoriskt tar ställning för de förra och mot de senare. Artikel II, som handlar om mekanistiska resonemang, har mycket med denna fråga att göra. Mekanistiska resonemang har ofta kommit långt ned, eller saknats helt, i hierarkierna (se exempelhierarkin ovan, där mekanistiska resonemang nämns längst ned). Men detta verkar inte vara helt rimligt. Frågans bedömning hänger dock i viss mån på vad man menar ska räknas som ett mekanistiskt resonemang.

En överraskande negligerad fråga i forskningslitteraturen kring EBM är vad ordningen i evidenshierarkier kan tänkas betyda. När det påstås i en evidenshierarki att en viss undersökningstyp T_1 är bättre än en annan undersökningstyp T_2, vad betyder det närmare bestämt? Att varje undersökning av typ T_1 ger starkare evidens än varje undersökning av typ T_2? Att en enstaka undersökning av typ T_1 ger starkare evidens än hur många undersökningar som helst av typ T_2? Att en T_1-undersökning i regel ger starkare evidens än en T_2-undersökning? Det finns en rad tänkbara tolkningar av vad ordningen betyder, men förespråkarna för evidenshierarkier har sällan varit tydliga med vilken tolkning som är riktig (och samma kritik kan riktas mot hierarkiernas motståndare). Artikel III reder ut vilka tolkningar som finns och vilka som är mer eller mindre rimliga.

Artiklarna I–III kan alltså sägas ha en mycket direkt anknytning med EBM:s evidenshierarkier och deras giltighet. Artiklarna IV och V har också anknytning, men något mer indirekt. Randomiserade studier placeras regelmässigt ovanför icke-randomiserade (observationella) studier i evidenshierarkierna. Och det råder ganska stor enighet om att randomisering är en metodologiskt smart sak. Men *varför* är det smart? Förvånande nog är enigheten betydligt mindre så fort man börjar fråga om sådana detaljer. I artiklarna IV och V görs en noggrann genomgång och bedömning av de argument för randomisering i kliniska prövningar som förekommer i litteraturen.

Artikel I

Företrädare för alternativmedicin hävdar ibland att deras metoder inte alls på samma sätt som den gängse vårdens metoder kan utvärderas vetenskapligt. I artikel I försöker jag därför utreda vilka krav som måste ställas på en behandling för att den alls ska kunna utvärderas vetenskapligt (i kliniska prövningar eller dylikt). Jag begränsar mina resonemang till krav som kan ställas på själva *behandlingen*, till skillnad från krav som (i och för sig rätteligen) skulle

kunna ställas på försökspersoner och deras uppträdande, på insamlade datas korrekta behandling, på använda mätinstruments tillförlitlighet, med mera.

Två krav presenteras och diskuteras utifrån en modellsituation där två behandlingar *A* och *B* ska jämföras i en klinisk prövning. Det första kravet gäller möjligheten att kunna säga vilken försöksperson som fått vilken behandling. Kravet kallas DC (eng. *distinguishing criterion*) och lyder något förenklat enligt följande:

> För varje patient som deltar i försöket måste man, med hjälp av ett kriterium som har formulerats före försökets start och med information som insamlats före eller under försöket, kunna avgöra om behandling *A* eller *B* har givits.

Dessutom krävs att ingen försöksdeltagare erhållit både behandling *A* och behandling *B*.

Det andra kravet är att störande variabler elimineras. Man säger att det finns en störande variabel (eng. *confounding variable*) när en faktor skiljer sig mellan de grupper man vill jämföra utan att det varit meningen, och när denna faktor skulle kunna förklara en skillnad i det man velat mäta. Mer utförligt men ändå något förenklat har jag formulerat detta krav, som jag kallar ECV (eng. *elimination of confounding variables*), sålunda:

> Ingen variabel får finnas i försöket sådan att (i) det finns en systematisk skillnad i denna variabel mellan grupperna som erhållit behandling *A* respektive *B*, (ii) det som man verkligen vill mäta i försöket har påverkats i betydande omfattning av variabeln, och (iii) variabeln ingår inte i kriteriet som omtalas i DC.

Dessutom får behandlingen inte ha egenskapen att dess eventuella effekt försvinner bara för att man testar behandlingen vetenskapligt.

Flera missförstånd kring alternativmedicinens förhållande till vetenskaplig prövning, eller mer allmänt kring hur man lämpligen utvärderar medicinska behandlingsmetoders effekter, kan förstås som brott mot de två kraven. Ett exempel kan vara uppfattningen att det inte skulle gå att vetenskapligt utvärdera behandlingsmetoder som bygger på »andliga« eller på annat sätt ovetenskapliga synsätt och teorier. Ett annat exempel är de ibland framförda påståendena att individuellt anpassade, och därmed icke-standardiserade, behandlingar inte kan testas vetenskapligt. Inte i något fall finns det ringaste stöd för dessa uppfattningar att hämta i de krav jag ställt upp.

De presenterade kraven är tänkta att utgöra *tillräckliga* villkor för vetenskaplig prövning. Det vill säga, om de båda kraven är uppfyllda för en given

behandlingsmetod så kan metodens effektivitet prövas i ett vetenskapligt godtagbart behandlingsexperiment (en klinisk prövning). Däremot vill jag inte hävda att kraven är *nödvändiga* att uppfylla: man kan med andra ord tänka sig att vissa behandlingsmetoder är vetenskapligt utvärderingsbara utan att uppfylla kraven. Det kan kanske i förstone verka ointressant att ställa upp tillräckliga istället för nödvändiga villkor. Men en närmare eftertanke visar att så inte är fallet. Visserligen kan det finnas behandlingsmetoder som inte uppfyller de två kraven men ändå är vetenskapligt testbara, och det är synd att dessa inte kan diagnosticeras av kraven, en situation som inte hade uppstått med nödvändiga villkor. Men det finns två fördelar med tillräckliga istället för nödvändiga villkor. För det första är tillräckliga villkor mer praktiskt användbara. Vi kan nu säga till alternativmedicinens förespråkare (i den mån de alls är intresserade av vetenskaplig prövning): se här, om din behandlingsmetod uppfyller dessa två krav så är den vetenskapligt utvärderingsbar! Nödvändiga villkor hade istället motsvarat en om–så-sats i andra riktningen: om din behandlingsmetod är vetenskapligt utvärderingsbar så uppfyller den dessa två krav.[2] För det andra kan man förmoda, att den som inte håller med mig att de krav jag ställt upp är tillräckliga, anser att de är för slappa och måste skärpas. Det betyder att man vill inta en hårdare attityd till alternativmedicinen än den som följer av mina två krav. Genom att ställa upp tillräckliga krav uppvisar man en generös attityd gentemot alternativmedicinen: det räcker med att en oortodox behandlingsmetod uppfyller dessa modesta krav för att den ska vara vetenskapligt utvärderingsbar.

Den övergripande bild som ges genom mina två tillräckliga villkor för vetenskaplig prövning av medicinska behandlingsmetoder är *inklusiv*: det är inte svårt att uppfylla dem, och mitt intryck är att de allra flesta alternativmedicinska metoder som har någon nämnbar spridning också uppfyller dem. Den som vill hävda att testning likväl inte är möjlig måste tillhandahålla en mer sofistikerad argumentation för detta.

Artikel II

Artikel II handlar om mekanistiska resonemang (eng. *mechanistic reasoning*) och frågan om vad dessa kan ha för värde när man försöker utröna en medicinsk behandlingsmetods effekt. Ett typexempel på ett mekanistiskt resonemang kan vara följande:

> Läkemedlet L innehåller substansen S_1, som när det tas upp i

[2] Nödvändiga villkor hade varit till stor nytta i diskussioner med alternativmedicinens förespråkare om dessa gjort sig kända för att påstå sig använda metoder som är vetenskapligt utvärderingsbara. Argumentationen i artikel I är mer användbar i relation till den, som det tycks mig, tämligen stora andel alternativmedicinska förespråkare som säger precis motsatsen.

> kroppen ger upphov till reaktionen R_1, som i sin tur ger upphov till reaktionen R_2, som frisätter substansen S_2, som har en välkänt hämmande inverkan på symptomet Y. Följaktligen kan vi ge läkemedlet L för att lindra symptomet Y.

Resonemanget består typiskt i en kedja där händelser länkar i varandra och till slut leder fram till den önskade effekten.

I artikeln försöker jag göra flera saker. För det första diskuterar jag hur termen »mekanistiskt resonemang« bör definieras i det sammanhang som här är av intresse, alltså medicinsk behandlingsforskning. Jag utgår från en definition som givits av andra författare, men motiverar flera modifieringar och landar till slut i följande:

> Ett mekanistiskt resonemang är ett resonemang som inbegriper antingen en slutledning från mekanismer till påståenden om specifika behandlingsutfall, eller en slutledning utifrån en undersökning av huruvida det finns troliga mekanistiska kedjor till påståenden om specifika behandlingsutfall.

Med andra ord: ett mekanistiskt resonemang antingen inbegriper en mekanistisk kedja, eller består i ett resonemang kring huruvida det kan finnas någon mekanistisk kedja. Denna definition är bredare än en definition som kräver att en mekanistisk kedja preciseras. Exempelvis kvalificerar sig följande som ett mekanistiskt resonemang enligt min definition, trots att ingen mekanistisk kedja beskrivs:

> Ett par »hörlurar« som inte utsänder ljud utan istället ljus från dioder påstås kunna minska humörsvängningar som t.ex. beror på säsongsbunden dagsljusvariation. Enligt aktuell fysiologisk kunskap finns emellertid inga fotoreceptorer i hörselgången, varför ljus som skickas in där endast kan vidarebefordras till andra delar av huvudet som värme. Det finns inget upptänkligt sätt på vilket små mängder värme i hörselgången skulle kunna orsaka en minskning i frekvens eller allvar hos humörsvängningar. Därför kan det inte finnas någon mekanistisk koppling mellan behandling och önskat utfall i detta fall.

För det andra försöker jag karakterisera olika typer av *negativa* mekanistiska resonemang, då de hittills varit tämligen negligerade i litteraturen. Man kan tala om negativa mekanistiska resonemang i två huvudbemärkelser: epistemiskt och hälsorelaterat. Ett epistemisk-negativt mekanistiskt resonemang

skulle vara ett som inte inkluderar en mekanistisk kedja. Ett hälsonegativt mekanistiskt resonemang skulle vara ett som leder till ett utfall som är negativt för patienten (eller som åtminstone inte är positivt trots att ett positivt utfall vore förväntat). Jag presenterar tre huvudtyper av negativa mekanistiska resonemang, NegA, NegB och NegC, som kan karakteriseras på följande sätt:

NegA	Man föreslår en mekanistisk kedja mellan behandling och utfall, men utfallet är hälsonegativt för patienten.
NegB	Man har seriöst försökt men misslyckats med att finna en mekanistisk kedja mellan behandling och utfall.
NegC	Metamekanistiska argument ger vid handen att det inte kan finnas någon mekanistisk koppling mellan behandling och utfall.

NegA är hälsonegativt. NegB och NegC är båda epistemiskt negativa, men NegB innefattar ett konkret försök att finna en mekanistisk kedja, medan NegC är »metamekanistiskt«, dvs. man letar inte efter mekanistiska kedjor utan undersöker istället möjligheten att det alls skulle kunna finnas sådana kedjor och landar i slutsatsen att så inte kan vara fallet.

Även om var och en av de tre typerna kan associeras med ett intervall av evidensstyrkor, finns det några generella skillnader mellan typerna. NegB har i allmänhet låg evidensstyrka: bara för att vi misslyckats med att finna en mekanistisk kedja betyder det inte att ingen skulle kunna finnas. I NegA har man verkligen funnit (eller tror sig ha funnit) en mekanistisk kedja som leder till ett icke-positivt utfall. Detta har generellt lite högre evidensstyrka än NegB, men behöver ändå inte vara särskilt övertygande: bara för att man har funnit en kedja som leder till ett icke-positivt utfall utesluter det inte att det kan finnas andra kedjor som leder till positiva utfall. NegC, däremot, kan ha ganska hög evidensstyrka, i synnerhet om man bedömer att det omöjligen kan finnas någon mekanistisk kedja med hänvisning till kunskap som man anser vara mycket säker. Den kunskap man åberopar kan vara både av mer teoretisk natur (t.ex. naturlagar) eller mer direkt empiriskt grundad (såsom omfattande kliniska studier). Jag ger flera exempel på negativa mekanistiska resonemang enligt typerna NegA, NegB och NegC.

För det tredje använder jag min definition och min karakteristik av olika typer av negativa mekanistiska resonemang för att argumentera att företrädare för EBM har avfärdat mekanistiska resonemang alldeles för lättvindigt. Ibland kan mekanistiska resonemang ha stor evidensstyrka. I viss EBM-litteratur kan man rentav se att »evidens« har definierats på ett sådant sätt att mekanistiska resonemang inte räknas som evidens alls. Det är orimligt. Det faktum att mekanistiska resonemang ibland kan leda fel är naturligtvis ett

gott skäl till att inta en skeptisk hållning, men min analys visar att somliga typer av mekanistiska resonemang har hög trovärdighet, och detta bör man ta hänsyn till och försöka inkorporera i EBM.

Artikel III

Evidenshierarkier är vanliga i EBM. En evidenshierarki är en ordnad lista av sätt på vilka evidens kan tas fram, ofta med det övergripande syftet att vara ett hjälpmedel för att bedöma och sammanväga evidens för eller emot effektiviteten hos en medicinsk behandlingsmetod. Frågan är vad ordningsrelationen i en sådan hierarki kan tänkas betyda. Om man påstår att evidens som har tagits fram enligt strategi S_1 är starkare än evidens som tagits fram med strategi S_2, så måste detta betyda något mer bestämt. (Det som här benämns »strategier« kan t.ex. vara randomiserad kontrollerad studie, mekanistiskt resonemang, observationsstudie, m.m.) Jag urskiljer fyra huvudtolkningar:

(1) Varje S_1-undersökning ger starkare evidens än vilken som helst S_2-undersökning.
(2) En S_1-undersökning ger starkare evidens än den S_2-undersökning som är så lik S_1-undersökningen som möjligt (alltså bortsett från att de är av olika strategityp).
(3) En S_1-undersökning ger typiskt sett starkare evidens än en S_2-undersökning.
(4) En idealt genomförd (dvs. mycket högkvalitativ) S_1-undersökning ger starkare evidens än en idealt genomförd S_2-undersökning.

För tolkningarna (3) och (4) kan man enkelt formulera olika varianter. I (3) kan man således undra exakt vad »typiskt sett« betyder. Kanske syftar det på ett statistiskt typvärde, kanske på ett aritmetiskt medelvärde eller på en median. I (4) kan man försvaga formuleringen något genom att byta ut »idealt genomförd« mot »tillräckligt väl genomförd«.

Jag argumenterar att tolkningarna (1) och (2) är orimliga i ett sammanhang där hierarkier används för bedömning och sammanvägning av evidens för eller emot effektiviteten hos medicinska behandlingsmetoder. Tolkningarna (3) och (4) (med varianter) är rimligare. Jag argumenterar vidare att tolkning (3) verkar stämma bäst in på GRADE, även om inget explicit ställningstagande för tolkning (3) kan återfinnas i GRADE-gruppens skrifter.

Även om man skulle fastställa en viss tolkning som den riktiga i något sammanhang, följer tyvärr ytterst litet som kan underlätta den komplexa uppgiften att väga samman evidens som härrör från olika strategier. Jag tror dock att en specificering av avsedd tolkning av ordningsrelationen i

evidenshierarkier kan bidra till klarhet och framsteg i vissa debatter kring dessa.

Artikel IV

Argument för randomisering i kliniska prövningar kan indelas i teoretiska och praktiska argument. De förra är oupplösligt länkade till teoretiska begrepp eller sammanhang, medan de senare går ut på att randomisering minskar eller eliminerar någon *bias* (oönskad (risk för) förvrängning av resultaten) utan sådan teoretisk koppling. I denna artikel bedöms de teoretiska argumenten. Tre huvudsakliga sådana har identifierats:

(1) Randomisering krävs vid nollhypotesprövning.
(2) Randomisering krävs för att man ska kunna dra (trovärdiga) slutsatser om att behandlingen har orsakat utfallet (t.ex. tillfrisknandet).
(3) Randomisering kan på ett acceptabelt och bekvämt sätt i ett Bayesianskt ramverk tillhandahålla *a priori*-fördelningar hos variabler.

In min bedömning av dessa argument har jag försökt att efterleva två riktlinjer, nämligen (*a*) att sammanhanget är att en realistisk prövning planeras, och (*b*) att slumpmässigheten som finns i randomisering ska spela roll för argumentets hållbarhet.

Det finns två mer specifika påståenden som kan leda fram till (1). Det ena är att randomisering gör att kvarvarande fel, även efter eventuell genomförd stratifiering, kan modelleras såsom samlade i en sannolikhetsfördelning. Det andra är att randomisering omvandlar systematiska fel till slumpfel. Jag stödjer det första men underkänner det andra påståendet. Sammantaget bör ändå (1) formuleras något svagare, till exempel: randomisering medverkar till att göra antaganden som behövs vid nollhypotesprövning giltiga.

Jag anser att (2) är ett svagt argument, om vi håller (*a*) och (*b*) i minnet. Även om randomisering kan sägas underlätta kausala slutledningar i allmänhet i experiment, kan randomisering i kliniska prövningar inte tillhandahålla något som kommer i närheten av en garanti för att utfallsförändringar måste ha orsakats av behandlingen. Jag vill också hävda att randomiserade kliniska prövningar inte i första hand utförs för att man vill fastställa en kausal koppling (ett epistemiskt mål) utan snarare för att man vill undersöka huruvida behandlingen fungerar (ett praktiskt eller handlingsvägledande mål).

Avseende (3) har de flesta bayesianer varit mycket skeptiska mot randomisering, men några har gått med på att randomisering kan godtas inom ett bayesianskt tänkesätt. Då fullständiga och formella bayesianska analyser av resultaten från kliniska prövningar dock sällan efterfrågas, är det svårt att formulera (3) på ett allmänt hållbart sätt avseende kliniska prövningar.

Sammantaget godkänner jag en modifierad version av (1), men förkastar de andra teoretiska argumenten för randomisering.

Artikel V

Enligt praktiska argument för randomisering i kliniska prövningar kan randomisering minska eller eliminera någon bias, utan att det finns en teoretisk koppling. Jag har identifierat fyra huvudargument av denna typ:

(1) Randomisering bidrar till att hemlighålla vilken patient som ska få vilken behandling.
(2) Randomisering bidrar till att balansera behandlingsgrupperna med avseende på störfaktorer.
(3) Randomisering eliminerar *self-selection bias*, det vill säga bias som uppstår när försöksdeltagarna får välja (eller kan påverka valet av) behandling.
(4) Randomisering eliminerar *allocation bias*, det vill säga bias som uppstår när försöksledare och andra icke-deltagare som är involverade i studien väljer (eller påverkar valet av) behandling för deltagarna.

Jag granskar argumenten (1)–(4) enligt samma riktlinjer (*a*) och (*b*) som i artikel IV.

Argument (1) förefaller svagt. Det viktigaste skälet för denna bedömning är att slumpmässigheten i randomisering inte är avgörande. För att hemlighålla allokeringssekvensen (den ordning i vilken deltagarna ska fördelas till de olika behandlingarna) är det, allmänt sett, mycket viktigare att sekvensen hanteras enligt fastställda regler än att den är framtagen med hjälp av en slumpmekanism. Dessa regler kan vara identiska oavsett hur sekvensen har skapats.

Argument (2) verkar vara mycket vanligt. För att kunna diskutera det måste man vara klar över skillnaden mellan enkel och stratifierad randomisering. I enkel randomisering fördelas varje deltagare till en behandling i enlighet med en slumpprocess, utan att dessförinnan ha blivit grupperad eller ordnad på något annat sätt. Stratifierad randomisering sker däremot i två steg. Först grupperas deltagarna så att grupperna (som kan kallas *strata*) är internt likartade med avseende på en eller flera störfaktorer. Detta steg är stratifieringen. Steg två är den riktiga randomiseringen, då deltagarna fördelas till sina slutgiltiga behandlingsgrupper, med hjälp av en slumpmekanism, på så sätt att fixa proportioner av deltagare hamnar i varje behandlingsgrupp från varje stratum. Proportionerna är fixa i den meningen att de är identiska mellan strata.

När blir en godtycklig variabel V bättre balanserad mellan behandlingsgrupperna med randomisering än utan randomisering? Det beror på om vi betraktar enkel eller stratifierad randomisering. Stratifiering med avseende på V kan dock endast genomföras om värdet på V är känt för varje individ; det räcker inte att V är känd på gruppnivå. Om V inte är känd på individnivå så kan en stratifiering med avseende på en annan variabel W förbättra balansen även för V, förutsatt att V och W är korrelerade. Jag kommer fram till följande:

- Om V är känd så att stratifiering är möjlig, så ger stratifiering med avseende på V en balans som är mycket bättre än vad som kan förväntas genom enkel randomisering.
- Om V är känd endast i allmänhet, men inte för varje deltagare, så ger en stratifiering med avseende på en annan variabel W troligen bättre balans med avseende på V än som hade resulterat ur enkel randomisering, förutsatt att V och W är korrelerade. Förfarandet kräver förstås att W är känd på individnivå.
- Om V är okänd, eller om V är känd men inte på individnivå, och om V inte är korrelerad med någon annan störvariabel som kan stratifieras, så kan V inte balanseras bättre än vad som är förväntat genom enkel randomisering.

Allt detta leder till slutsatsen att (2) är ett svagt argument för randomisering, när randomisering jämförs med, eller kombineras med, stratifiering. Om stratifiering inte ska komma ifråga alls, så blir den relevanta jämförelsen istället den mellan enkel randomisering och någon annan gruppfördelningsprocedur som saknar slumpelement. I en sådan jämförelse verkar enkel randomisering ge bättre balans, men endast genom att minska *self-selection bias* och/eller *allocation bias*, dvs. de två centrala faktorerna i argumenten (3) och (4). Jag finner därför inget bra stöd för (2), om (2) betraktas som självständigt från (3) och (4).

Argument (3) är hållbart: slumpmässigheten i randomisering garanterar att ingen *self-selection bias* kan förekomma. Det har påstått att faktumet att man utför en *interventions*studie – snarare än en icke-interventionsstudie – är avgörande för att förhindra *self-selection bias*, inte faktumet att man randomiserar. Jag håller inte med om detta. *Self-selection bias* är typiskt sett liten, ibland obefintlig, i interventionsstudier, men denna bias kan likväl förekomma i icke-randomiserade interventionsstudier – men kan inte förekomma om randomisering används.

Argument (4), som hänvisar till *allocation bias*, förefaller starkt: randomisering utesluter denna bias, och den skulle typiskt sett ha förekommit i

en klinisk prövning om randomisering inte tillämpats. *Allocation bias* skulle typiskt sett ha förväntats vara ett allvarligare problem i avsaknad av randomisering än *self-selection bias*.

Sammantaget godtar jag alltså (3) och (4). Av dessa förefaller (4) att vara starkast. Argument (1) förkastas helt. Argument (2) förkastas också som självständigt argument, men det är ändå sant att randomisering bidrar till balans mellan behandlingsgrupperna, nämligen genom mekanismerna som nämns i (3) och (4).

Index

Theses in Philosophy from KTH Royal Institute of Technology

Below are listed all theses in philosophy that have been presented and defended at KTH Royal Institute of Technology since the inception of its Division of Philosophy in 2000. A licentiate thesis may be defended half-way between a master's degree and a doctor's degree. The content of a licentiate thesis could therefore be partially or wholly incorporated into a subsequent doctoral thesis by the same author.

Introductions to theses are, in general, freely available through the KTH Publication Database DiVA (see http://kth.diva-portal.org/). Fulltexts of complete theses are generally not available, due to copyright restrictions on included papers.

1. Martin Peterson, *Transformative Decision Rules and Axiomatic Arguments for the Principle of Maximizing Expected Utility*, Licentiate thesis, 2001.
2. Per Sandin, *The Precautionary Principle: From Theory to Practice*, Licentiate thesis, 2002.
3. Martin Peterson, *Transformative Decision Rules: Foundation and Applications*, Doctoral thesis, 2003.
4. Anders J. Persson, *Ethical Problems in Work and Working Environment Contexts*, Licentiate thesis, 2004.
5. Per Sandin, *Better Safe than Sorry: Applying Philosophical Methods to the Debate on Risk and the Precautionary Principle*, Doctoral thesis, 2004.
6. Barbro Björkman [now: Barbro Fröding], *Ethical Aspects of Owning Human Biological Material*, Licentiate thesis, 2005.
7. Eva Hedfors, *The Reading of Ludwik Fleck: Sources and Context*, Licentiate thesis, 2005.
8. Rikard Levin, *Uncertainty in Risk Assessment—Contents and Modes of Communication*, Licentiate thesis, 2005.
9. Elin Palm, *Ethical Aspects of Workplace Surveillance*, Licentiate thesis, 2005.
10. Jessica Nihlén Fahlquist, *Moral Responsibility in Traffic Safety and Public Health*, Licentiate thesis, 2005.
11. Karin Edvardsson, *How to Set Rational Environmental Goals: Theory and Applications*, Licentiate thesis, 2006.
12. Niklas Möller, *Safety and Decision-Making*, Licentiate thesis, 2006.
13. Per Wikman Svahn, *Ethical Aspects of Radiation Protection*, Licentiate thesis, 2006.
14. Hélène Hermansson, *Ethical Aspects of Risk Management*, Licentiate thesis, 2006.
15. Madeleine Hayenhjelm, *Trust, Risk and Vulnerability*, Licentiate thesis, 2006.
16. Holger Rosencrantz, *Goal-setting and Goal-achieving in Transport Policy*, Licentiate thesis, 2006.
17. Kalle Grill, *Anti-paternalism*, Licentiate thesis, 2006.
18. Jonas Clausen Mork, *Is it Safe? Safety Factor Reasoning in Policy Making under Uncertainty*, Licentiate thesis, 2006.

19. Anders J. Persson, *Workplace Ethics: Some Practical and Foundational Problems*, Doctoral thesis, 2006.
20. Eva Hedfors, *Reading Fleck: Questions on Philosophy and Science*, Doctoral thesis, 2006.
21. Nicolas Espinoza, *Incomparable Risks, Values and Preferences*, Licentiate thesis, 2006.
22. Mikael Dubois, *Prevention and Social Insurance—Conceptual and Ethical Aspects*, Licentiate thesis, 2007.
23. Birgitte Wandall, *Influences on Toxicological Risk Assessments*, Licentiate thesis, 2007.
24. Madeleine Hayenhjelm, *Trusting and Taking Risks: A Philosophical Inquiry*, Doctoral thesis, 2007.
25. Hélène Hermansson, *Rights at Risk—Ethical Issues in Risk Management*, Doctoral thesis, 2007.
26. Elin Palm, *The Ethics of Workplace Surveillance*, Doctoral thesis, 2008.
27. Jessica Nihlén Fahlquist, *Moral Responsibility and the Ethics of Traffic Safety*, Doctoral Thesis, 2008.
28. Barbro Björkman [now: Barbro Fröding], *Virtue Ethics, Bioethics, and the Ownership of Biological Material*, Doctoral thesis, 2008.
29. Karin Edvardsson Björnberg, *Rational Goal-Setting in Environmental Policy: Foundations and Applications*, Doctoral thesis, 2008.
30. Holger Rosencrantz, *Goal-Setting and the Logic of Transport Policy Decisions*, Doctoral thesis, 2009.
31. Kalle Grill, *Anti-paternalism and Public Health Policy*, Doctoral thesis, 2009.
32. Marion Godman, *Philosophical and Empirical Investigations in Nanoethics*, Licentiate thesis, 2009.
33. Niklas Möller, *Thick Concepts in Practice: Normative Aspects of Risk and Safety*, Doctoral thesis, 2009.
34. Johan E. Gustafsson, *Essays on Value, Preference, and Freedom*, Licentiate thesis, 2009.
35. Lars Lindblom, *The Employment Contract between Ethics and Economics*, Doctoral thesis, 2009.
36. Eduardo Fermé, *On the Logic of Theory Change: Extending the AGM Model*, Doctoral thesis, 2011.
37. Linda Johansson, *Robots and Moral Agency*, Licentiate thesis, 2011.
38. Johan E. Gustafsson, *Preference and Choice*, Doctoral thesis, 2011.
39. Per Norström, *Technology Education and Non-scientific Technological Knowledge*, Licentiate thesis, 2011.
40. Jonas Clausen Mork, *Dealing with Uncertainty*, Doctoral thesis, 2012.
41. Per Wikman-Svahn, *Ethical Aspects of Radiation Risk Management*, Doctoral thesis, 2012.
42. Karim Jebari, *Crucial Considerations: Essays on the Ethics of Emerging Technologies*, Licentiate thesis, 2012.
43. Linda Johansson, *Autonomous Systems in Society and War: Philosophical Inquiries*, Doctoral thesis, 2013.
44. Dan Munter, *Ethics at Work: Two Essays on the Firm's Moral Responsibilities towards Its Employees*, Licentiate thesis, 2013.
45. Sara Belfrage, *In the Name of Research: Essays on the Ethical Treatment of Human Research Subjects*, Doctoral thesis, 2014.
46. Patrik Baard, *Sustainable Goals: Feasible Paths to Desirable Long-Term Futures*, Licentiate thesis, 2014.
47. William Bülow, *Ethics of Imprisonment: Essays in Criminal Justice Ethics*, Licentiate thesis, 2014.
48. Per Norström, *Technological Knowledge and Technology Education*, Doctoral thesis, 2014.
49. Anna Stenkvist, *Pictures, Mathematics and Reality: Essays on Geometrical Representation and Mathematics Education*, Licentiate thesis, 2014.

50. Karim Jebari, *Human Enhancement and Technological Uncertainty: Essays on the Promise and Peril of Emerging Technology*, Doctoral thesis, 2014.
51. Mikael Dubois, *The Justification and Legitimacy of the Active Welfare State: Some Philosophical Aspects*, Doctoral thesis, 2015.
52. Payam Moula, *Ethical Aspects of Crop Biotechnology in Agriculture*, Licentiate thesis, 2015.
53. Jesper Jerkert, *Philosophical Issues in Medical Intervention Research*, Licentiate thesis, 2015.
54. Alexander Mebius, *Philosophical Controversies in the Evaluation of Medical Treatments: With a Focus on the Evidential Roles of Randomization and Mechanisms in Evidence-Based Medicine*, Doctoral thesis, 2015.
55. Patrik Baard, *Cautiously Utopian Goals: Philosophical Analyses of Climate Change Objectives and Sustainability Targets*, Doctoral thesis, 2016.
56. William Bülow, *Unfit to Live among Others: Essays on the Ethics of Imprisonment*, Doctoral thesis, 2017.
57. Björn Lundgren, *Semantic Information and Information Security: Definitional Issues*, Licentiate thesis, 2016.
58. Robert Erdeniz, *Military Operations Planning and Methodology: Thoughts on Military Problem-Solving*, Licentiate thesis, 2017.
59. Jesper Ahlin, *Autonomy and Informed Consent: Conceptual and Normative Analyses*, Licentiate thesis, 2017.
60. Li Zhang, *On Non-Prioritized Multiple Belief Revision*, Doctoral thesis, 2018.
61. Björn Lundgren, *Information, Security, Privacy, and Anonymity: Definitional and Conceptual Issues*, Doctoral thesis, 2018.
62. Jesper Ahlin Marceta, *Authenticity in Bioethics: Bridging the Gap Between Theory and Practice*, Doctoral thesis, 2019.
63. Karl Sörenson, *In Search of Lost Deterrence*, Licentiate thesis, 2019.
64. Maria Nordström, *Is Time Money? Philosophical Perspectives on the Monetary Valuation of Travel Time*, Licentiate thesis, 2020.
65. Jesper Jerkert, *Philosophical Aspects of Evidence and Methodology in Medicine*, Doctoral thesis, 2021.
66. Anna Wedin, *Ethical Adaptation to Sea Level Rise: The Planner's Perspective*, Licentiate thesis, 2021.