Philosophical Issues in Medical Intervention Research

Jesper Jerkert

Licentiate thesis in Philosophy
KTH Royal Institute of Technology
Stockholm, 2015
Abstract


The thesis consists of an introduction and two papers. In the introduction a brief historical survey of empirical investigations into the effectiveness of medicinal interventions is given. Also, the main ideas of the EBM (evidence-based medicine) movement are presented. Both included papers can be viewed as investigations into the reasonableness of EBM and its hierarchies of evidence.

Paper I: Typically, in a clinical trial patients with specified symptoms are given either of two or more predetermined treatments. Health endpoints in these groups are then compared using statistical methods. Concerns have been raised, not least from adherents of so-called alternative medicine, that clinical trials do not offer reliable evidence for some types of treatment, in particular for highly individualized treatments, for example traditional homeopathy. 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 evaluable in a clinical trial, namely (1) the proper distinction of the two treatment groups and (2) the elimination of confounding variables or variations. These are delineated, and a few misunderstandings are corrected. It is concluded that the conditions do not preclude the testing of alternative medicine, whether individualized or not.

Paper II: Traditionally, mechanistic reasoning has been assigned a negligible role in standard EBM literature, although some recent authors have argued for an upgrading. Even so, mechanistic reasoning that has received attention 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, both in the epistemic and in the health-related sense. I distinguish three main types of negative mechanistic reasoning and subsume them under a new definition of mechanistic reasoning in the context of assessing medical interventions. Although this definition is wider than a previous suggestion in the literature, there are still other instances of reasoning that concern mechanisms but do not (and should not) count as mechanistic reasoning. 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. Each negative type corresponds to a range of evidential strengths, and it is argued that there are differences with respect to the typical strengths. The variety of negative mechanistic reasoning should be acknowledged in EBM, and presents a serious challenge to proponents of so-called medical hierarchies of evidence.

Keywords: scientific method, study design, methodology, alternative medicine, medical research, individualized treatments, eligibility, confounders, evidence, evidence-based medicine, mechanistic reasoning, hierarchy of evidence.

Jesper Jerkert, Division of Philosophy, Department of Philosophy and History, KTH Royal Institute of Technology, Brinellvägen 32, SE-100 44 Stockholm, Sweden

© 2015 by Jesper Jerkert.

Thesis composition

This thesis consists of an introduction and the following papers:


II. Jesper Jerkert: “Negative mechanistic reasoning in medical intervention assessments”, submitted manuscript.

There is also a summary in Swedish.

Other works by the author

During the writing of this thesis, several other texts were authored. Some texts that are at least remotely connected to philosophy and/or to the topic of this thesis are listed below.

Popular philosophy & popular science

Philosophical book reviews

Teaching texts
Contents

Abstract 2

Thesis composition 3

Acknowledgments 7

Sammanfattning på svenska (Summary in Swedish) 9

Introduction 17
  1. The questions .............................. 17
  2. Historical outline .......................... 18
  3. What is EBM? ............................. 27
  4. Critique of EBM .......................... 30
  5. Summary of the papers ................... 34
  6. Prospects for future work ............... 38
References .................................. 40

Paper I: Why alternative medicine can be scientifically evaluated 45
  1. Introduction ............................... 45
  2. Clinical trial criticisms .................. 46
  3. Objective ................................ 48
  4. A model situation ........................ 49
  5. The distinguishing criterion ............ 49
  6. The relevance issue: eliminating confounding variables ... 51
  7. Treatments inaccessible to scrutiny? .......... 52
  8. Countering three misunderstandings ....... 53
  9. Deviations from statistical ideality ........ 56
  10. What about blinding, randomization, and other clinical trial features? .......... 57
  11. Conclusion ............................... 59
References .................................. 59
Paper II: Negative mechanistic reasoning
in medical intervention assessment 63
1. Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 64
2. Defining mechanistic reasoning . . . . . . . . . . . . . . . . . . . . 65
3. A typology of negative mechanistic reasoning . . . . . . . . . 71
4. The nature of evidence
   and consequences for hierarchies of evidence . . . . . . . . . . . . . . . . 76
5. Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 78
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 78

Index 81

Theses in Philosophy from the Royal Institute of Technology 85
Acknowledgments

My main supervisor Sven Ove Hansson (professor at the Division of Philosophy, KTH) has provided comments on everything included in this thesis, as well as overall support and encouragement. Participants 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), who also gave useful feedback for Paper II. Tor Sandqvist (lecturer at the Division of Philosophy) also provided useful comments on Paper II. For all these comments and criticisms I am grateful. To the extent that there are aesthetic merits in this work, credit should be shared with Johan E. Gustafsson (Department of Philosophy, University of York), who taught me many things about \LaTeX{} typesetting.

Of course, none of the above-mentioned is in any way responsible for the arguments or conclusions contained in this thesis.
Sammanfattning på svenska
(Summary in Swedish)

Denna licentiatuppsats består av en introduktion och två 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 två 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?

För att förstå frågornas relevans och sätta in dem i sitt sammanhang behövs en kort bakgrundsteckning. Därefter sammanfattar jag var och en av artiklarna.

Bakgrund


Trots historiska exempel som Linds, är det ett fakta 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 medicinsens 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 fysiolo-
giskt 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 (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
- 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 studeras på det sättet. Inte minst från alternativmedicinskt håll har sådana påståenden luftats, 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. Min artikel II, som handlar om mekanistiska resonemang (som kommer på sista plats i exempelhierarkin ovan), rör exakt denna fråga.

Båda de ingående artiklarna i detta arbete kan alltså sägas ha en direkt beröring med EBM:s evidenshierarkier och deras giltighet.

**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 $n = 1$-studie deltar endast en försöksperson, som utsätts för olika experimentella betingelser (behandlingar) i en förutbestämd ordning.
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 möjlheten 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 försöket 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 måta 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 alternativmedicineens 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.


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 alternativ medicinska metoder som har någon nämnbar spridning också uppfyller dem. Det föreligger således inget formellt hinder att testa dem vetenskapligt. I den mån metodernas företrädare vill avfärda de resultat som kommit ur vetenskapliga prövningar måste deras argument vara betydligt mer sofistikerade än de föga trovärdiga påståendena om att vetenskaplig utvärdering inte skulle vara möjlig.

²Nödvändiga villkor hade varit till stor nytta i diskussioner med alternativ medicinskens 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 alternativ medicinskens förespråkare som säger precis motsatsen.
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 typ exempel på ett mekanistiskt resonemang kan vara följande:

Läkemedlet *L* innehåller substansen *S*₁, som när det tas upp i 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 upptäckligt 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 huvudbemärkelse: 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 verklig 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 stark evidensstyrka, i synnerhet om man bedömer att det omöjliga 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 måste man ta hänsyn till och försöka inkorporera i EBM.
Introduction

1. The questions

The overall topic of the present thesis is scientific methods and their trustworthiness. More precisely, the thesis includes two papers on methodological issues in medical research. The papers try to answer the following two questions:

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

– What, if any, is the proper role of mechanistic reasoning in a scientific evaluation of medical treatment effectiveness?

Inquiries into the methodology of a particular scientific field can always be claimed to belong to the specific science 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. However, there should be no doubt that they are also philosophical. Typically for much philosophy, each inquiry in this thesis takes an outside look at a practice and investigates the principles that govern (or that should govern) that practice.

We will return to the questions and, of course, to their answers (skip to section 5 for a summary). To understand their significance and to put them in context, some background material will be presented in this introduction. However, it ought be said that it is not necessary to read this introduction in order to understand the included papers; these are supposed to be self-contained.¹ This introduction, then, will probably be of greatest value to the

¹On the other hand, this introduction is self-contained, too. It includes summaries of the papers (Section 5), so that anyone who is not interested in details may read the introduction only.
general reader with an interest in medical and/or experimental methodology, and not necessarily to the philosopher.

Since the 1990’s, medical evaluation research has been debated particularly in relation to a school of thought known as evidence-based medicine (EBM). But of course, medical evaluation research has a history before the advent of EBM. A historical sketch of medical intervention research will be provided (section 2), after which the ideas of EBM are presented and discussed (sections 3–4). As will be shown, both questions above can be considered to be closely related to the validity of so-called hierarchies of evidence, which are generally acknowledged to be crucial for the understanding and application of EBM. Section 5 summarizes the papers, and section 6 provides brief suggestions for further inquiries in the field.

A terminological note first. The questions above talk about effectiveness. Often the term efficiency is considered as synonymous; sometimes this is true for effectivity, too. Whichever term is used, it is often contrasted to efficacy. In a medical setting, the difference may be explained as follows. Efficacy is established for a medical treatment if it is shown to be beneficial for patients under ideal circumstances, like in a clinical trial. (Often, one is not content with the treatment simply being beneficial, but requires it to be better than some alternative treatment that is already being used in routine practice.) Effectiveness, on the other hand, would be demonstrated if the treatment is shown to be beneficial in clinical reality. The relation between efficacy and effectiveness, and in particular how to move from claims of efficacy to claims of effectiveness, is a recurrent theme in philosophically informed debates concerning methodology in medicine and in the social sciences (e.g., Cartwright, 2009, 2011). However, this relation is not investigated in the present thesis. I therefore take the liberty of using the term effectiveness in this introduction, although sometimes efficacy would have been more correct.

2. Historical outline

2.1 Medical intervention research

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 that 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, and their integration, were unknown to followers of ancient medicine. 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, 1967:9f). Louis founded the Société d’Observation Médicale, and his thus practised médicine d’observation has been identified as a historical predecessor of EBM (Vandenbroucke, 1996).

Empirically oriented schools such as médicine 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, to 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 particularly 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 disinterested 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 in 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 and for about 2000 years, the dominant theory of the workings of the body stated that there are four fluids, also known as *humors*, that supposedly fill 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. Given that the humoral 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.

What explains the reluctance against testing treatments empirically? Perhaps simply the staunch reliance on erroneous theories such as humoral pathology, and perhaps also connections between health and spiritual/religious beliefs, that were not so easy to challenge (Hansson, in press). On the other hand, Wootton (2007:144) has maintained that it is strange that medical treatments were not being put to test:

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?

The reluctance to test 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. Only in certain delimited quarters is a generally skeptical attitude towards clinical trials upheld. Alternative medicine is the prime example. Arguments to the effect that alternative medicine, due to its highly individualized treatments, could not be tested in clinical trials are scrutinized and found wanting in Paper I.

In a clinical trial, patients with a specific disease (or with a specified collection of symptoms) are divided into groups. One group gets the treatment of interest, whereas the others are given other treatments (or no treatment). The outcomes are compared at group level with statistical methods. For a trial (or an experiment in general) to be reliable, it has to be carefully controlled. In a controlled experiment all variables believed to influence the outcome are either held constant or are being manipulated by the experimenter.² (Often only one variable is being manipulated and all others are being held constant.) Two methodological features that are contributing to control have become particularly associated with good clinical trials: blinding (masking) and randomization. The historical developments of blinding and randomization may therefore be of some interest and are reviewed in the subsections below. The phrase “randomized controlled trial” is ubiquitous in medical intervention research and is abbreviated RCT.

2.2 The history of blinding

Experimental blinding amounts to intentional ignorance on some part. In the context of evaluating medical interventions, a participant is blinded (masked) if he does not know to which treatment group 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 a scientific setting, 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

²There 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 an 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.
or pharmacology. The earliest phase concerned the medical establishment’s response to threatening unconventional healing methods such as mesmerism and homeopathy. Mesmerism – named after its inventor Franz Anton Mesmer (1734–1815) – seems to have been the first unconventional healing system to be combated with an armory that included blinding. The earliest instances featured physical blindfolding with bandages in experiments performed under Benjamin Franklin’s supervision. 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” (ibid.: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 should be conducted. 100 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 which five had received the homeopathic dilution and three had received snow water. Since Reuter had predicted that most participants that received the

---

3 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.

4 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. For an amusing account of Franklin’s investigation into mesmerism, see Lopez (1993).
homeopathic dilution would have unusual experiences, it was concluded that he was wrong (Stolberg, 2006). (Perhaps the most remarkable thing about this test is that it was not only double blind, but also randomized.)

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 incorporate it into clinical research. Blinding as a methodological feature of controlled trials constitutes the last phase in Kaptchuk’s review (1998:422ff).

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” (ibid., 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.

2.3 The history of randomization

In the context of medical intervention assessments, randomization amounts to the allocation of subjects to different treatments using some random mechanism. Put more generally, randomization amounts to the selection of experimental conditions by using a random mechanism. Randomization can be used at more than one stage in an experiment. As an example, consider the following telepathy experiment (Hacking, 1988:447f). A “sender” is instructed to stare at a card and make an inner image of it, trying to transmit it to a “receiver” in an adjacent room who is expected to guess the suit of the
card. The card is drawn at random in each trial (randomization 1). Then it is decided at random whether the sender is really given the card to stare at or not (randomization 2). The cases where the receiver guesses the suit without the sender actually staring at any card are used as controls. The main point of using controls in this situation would be that significantly positive results in the control trials indicate hitherto unnoticed design weaknesses.

Randomization is a younger methodological feature than blinding, at least if we consider the time at which it had become a fairly widespread practice. A trial published in 1948 on the use of streptomycin in pulmonary tuberculosis has been dubbed a “watershed” due to the careful use of randomization (Doll, 1998). Not until then, after World War II, did randomization 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 randomization at all, but that is not true. For example, already in the 1920’s and 1930’s randomization had an influential promoter in Ronald A. Fisher, whose textbooks (1925; 1935) went through several updated editions. Fisher, renowned as a statistician and as an evolutionary biologist, initially worked empirically with agricultural field trials, where different seeds were grown in matched plots.

But Fisher was not first either, nor was agricultural science the first discipline in which randomization 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 the first to have been randomized will obviously depend on what procedures we accept for yielding (sufficient) randomness. Today, one would hardly use the day of admittance as randomization variable. Perhaps one would expect the history of experimental randomization 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 J. B. van Helmont (1580–1644) suggested a trial (which, however, did not actually take place) to test humoral 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, 1648:526f; quoted from the translation in James Lind Library [2012]).

Hence, van Helmont is sometimes credited as the originator of experimental randomization. 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.

Van Helmont’s suggestion published in 1648 did not establish any practice of randomized trials. Ian Hacking (1988) has traced the systematic use of modern randomization (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. Could the subject 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 randomization first became more extensively discussed after its introduction. Almost simultaneously with the Peirce–Jastrow experiment, Charles
Richet (1850–1935), a French physiologist with a great interest in para-
psychology, 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 randomizer seems to have been that it enabled
inferences about the existence of hypothesized 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 well-publicized within the
parapsychological community; hence his practice of random drawing surely
became known.

A few years later, randomization was debated in relation to telepathy
trials in the form of number-guessing (not performed by Richet). Critics
pointed out that the selection of numerals to be transferred was crucial. For
if people tend to think alike and to share some favorite integers among those
available for the test, then we would expect performances above chance
level although no telepathy had occurred. Randomized numerals could be
used to remove this bias (ibid.:442).

As noted in the preceding section, historically there have been two
main reasons offered for why blinding should be applied: to eliminate
suggestion and to increase the probability that participants would remain in
the study. Both justifications can be combined non-controversially. As for
randomization, the situation is a little more complex. One could distinguish
at least three main justifications for experimental randomization (Worrall,
2002), but they are rather different from one another, and whether they
could be combined or not will depend on one’s epistemological outlook (in
particular, it seems to be dependent on whether one is a Bayesian or not).
The three main justifications are as follows:

1. Randomization facilitates blinding.
2. Randomization provides the theoretical basis for performing statistical
   hypothesis testing of the traditional “Fisherian” type.
3. Randomization controls for unknown confounders.

It may seem strange that both for blinding and for randomization, different
motivations are given. However, this only shows that neither blinding nor
randomization are by themselves methodological necessities. Rather, they
are sometimes solutions to more fundamental problems that ought to be
spelled out. (This is also emphasized in Section 10 of Paper I.)
3. What is EBM?

3.1 The origin of EBM

As was noted above (Section 2.1), 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 somewhat earlier, a JAMA article by a body called the Evidence-Based Medicine Working Group (1992) counts as the founding document. It 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)

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 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

---

5 There were 31 signatories that comprised the Evidence-Based Medicine Working Group, with Gordon Guyatt designated as chair. 24 out of 31 were affiliated with departments at McMaster University, Hamilton, Ontario (Canada). The EBM movement thus also has a rather precise geographical origin.

6 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 claim that the emergence of EBM constitutes a Kuhnian paradigm shift has no bearing on the present thesis, it will not be further discussed here.
patients. The practice of evidence based medicine means 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 good and well, but one also needs to explain what should count as clinical evidence, and how evidence from many studies is supposed to be aggregated (especially evidence from different study types). Within EBM, this is accounted for in so-called hierarchies of evidence, 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).

3.2 Hierarchies of evidence

A hierarchy of evidence is a list of study types (research designs). The higher up in the list, the stronger evidence is obtained from the study type (for answering a specified research question), it is claimed. Hierarchies of evidence were suggested before the 1992 seminal EBM article. One of the earliest, though not called a “hierarchy of evidence” at the time of publication, is found in an article by the Canadian Task Force on the Periodic Health Examination (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.

Since most lists consist of study types, the term “hierarchy of evidence” could be seen as a misnomer; “hierarchy of study types” would be more correct. Nonetheless, “hierarchy of evidence” is an established term and is retained here.

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.
Opinions of respected authorities, based on clinical experience, descriptive studies or reports of expert committees.

Perhaps the most well-known hierarchy of evidence today is the following, published by OCEBM Levels of Evidence Working Group (2011):

- Systematic review of randomized trials or n-of-1 trials
- Randomized trial or observational study with dramatic effect
- Non-randomized controlled cohort/follow-up study
- Case-series, case–control studies, or historically controlled studies
- Mechanism-based reasoning.

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
- 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.

All hierarchies above are presented, explicitly or implicitly, as being useful for assessments of whether an intervention helps or not. (For other research questions, the hierarchy might have to be changed.)

---

9This 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).
Both of the research questions in this thesis can be viewed as closely related to hierarchies of evidence. In Paper I, the claims of some alternative medicine proponents to the effect that their methods cannot be evaluated in clinical trials are scrutinized. Since high-quality clinical trials are found at the top of EBM’s hierarchies of evidence, what is at stake is also the applicability of EBM ways of thinking to alternative medicine. In Paper I, there is no discussion about whether RCT’s really should be at the top of the hierarchies. Section 4.1 below includes a brief discussion of this issue.

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. 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.

4. Critique of EBM

EBM has been scrutinized and criticized from various viewpoints. There is, of course, a general skepticism against empirical investigations from proponents of theories and schools in which critical thinking is not highly held. These critics will often feel that EBM’s demand for empirical evidence collected from clinical trials is unfair. In Paper I this topic is addressed with reference to alternative medicine, and it will not be further discussed here. Some other criticisms and debates will be reviewed below. In Section 4.2 the most serious concerns that are related to Paper II will be discussed. First, in Section 4.1, a few other points will be raised. These points are, however, not central to the present thesis.

4.1 The meaning and correctness of hierarchies of evidence

Since hierarchies of evidence are at the centre of EBM, a fair amount of effort has been invested into arguing that they are problematic. Just by looking at the example hierarchies above, we may note, for example, that the “all or none” item in the Straus et al. (2005) hierarchy is not a study type but an outcome from a study. So that’s an inconsistency. A few other possible inconsistencies will also emerge when the hierarchies are compared. So one possible line of criticism could be to ask: Why are the hierarchies different, although they are claimed to be useful for the same research question? Personally, however, I suspect that this criticism can be countered. (For an outline, see Section 6.3.)
Another line of criticism concerns the interpretation of a hierarchy of evidence. To the extent that it is interpreted in a certain categorical manner, it seems to be untenable. For example, Borgerson (2009:218) writes that “certain research methods in medicine are thought to be categorically better than others”, but she gives only one quotation from an EBM supporter to this effect (also, she does not explain the exact meaning of “categorically better”). La Caze asserts that “[t]he popular EBM guidebooks recommend applying the hierarchy categorically” (2011:84), but he gives only one example, viz. the same as Borgerson. Grossman and Mackenzie (2005:516) write that “most researchers consider that RCTs are always superior to all other types of evidence”; yet they do not present any quotations to this effect.

This criticism seems to be based on a straw man characterization of how hierarchies of evidence are actually used within EBM. Certainly, one can find a few quotations from EBM supporters that seem to endorse a categorical interpretation; the problem is that one can find many more quotations that reject such an interpretation. Examples of the latter include Guyatt & Rennie (2002:7), Nordenstrom (2007:40), Howick (2011:56), and OCEBM Levels of Evidence Working Group (2011). The standard EBM view today is that a study type higher up in a hierarchy is better in some ceteris paribus sense than a study type further down, but for each type there is room for upgrading or downgrading the evidential strength using study quality assessments (Gugiu & Gugiu, 2010).

A third line of criticism is simply that the hierarchies are not correctly designed. Two issues of controversy emerge here. One is whether the balance between empirical investigations and theoretical deliberations has been correctly struck. This is of particular interest in relation to my Paper II and is discussed in Section 4.2, below. The other issue is whether randomized controlled trials are superior to observational studies (as claimed in the hierarchies). Many papers are devoted to this latter question. Two examples follow. Concato et al. (2000) have claimed 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 were compared, the average results were very similar. In response to this and similar claims, EBM defender Howick (2011:39ff) has insisted that

---

10 The single most cited statement to this effect – the one supplied by Borgerson and (though not exactly verbatim) by La Caze – is the following from Straus et al. (2005:118): “If the study wasn’t randomized, we’d suggest that you stop reading it and go on to the next article in your search”.

11 It is also often rather unclear exactly what is meant by a “categorical interpretation”. Several meanings could in principle be distinguished.
randomization reduces more biases than observational studies. In addition, he denies that it is a general truth that RCTs and observational studies produce similar results.\textsuperscript{12}

4.2 The favoritism of empirical evidence in EBM

As is clear from the exposition in Section 3.2, EBM supporters are very skeptical towards theoretical deliberations (“opinion […] based on physiology, bench research or ‘first principles’ ”; “mechanism-based reasoning”) and towards authorities (“opinions of respected authorities”, ”reports of expert committees”). This skepticism, I would argue, is very much in line with the dominant perspective in EBM. This perspective is finding out \textit{what} works; and finding this out is supposed to be \textit{practically feasible}. Let me discuss briefly these two ideas, with special reference to mechanistic reasoning.

First I must explain 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 supposed to be 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. Also, generally speaking the very notion of a mechanism signals that there is no complete understanding of exactly what is going on. Hence, there is always some uncertainty as to whether the chain is correct.

Within EBM there is a strong reliance on empirical and experimental methods, methods that tell us \textit{what} works. Research methods that tell us what works in clinical practice typically do this by finding correlations between intervention changes (such as administration of a drug vs. administration of another drug) and pre-specified outcome measures (such as relief of certain symptoms). What happens in between – how exactly any change in outcome is brought about – is irrelevant to the presence or absence of

\textsuperscript{12}The debate on the merits of RCTs vs. observational studies is of course closely related to the more general question of how to justify randomization; see above Section 2.3 and below Section 6.4.
a statistical correlation. Mechanistic reasoning, on the other hand, is relevant almost exclusively to what happens in between. Hence, mechanistic reasoning is 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 (2011:154ff) presents a list of erroneous conclusions based on mechanistic reasoning from which the following examples are sampled: antiarrhythmic drugs will reduce mortality due to sudden cardiac death; hormone replacement therapy will reduce menopausal symptoms; vitamin E reduces the risk of coronary heart disease and atherosclerosis.

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) generalizations from large amounts of empirical data. (Below, I call this “theory-knowledge”.) According to the EBM literature, normal practitioners of medicine (doctors, nurses, etc.) should be able to apply the EBM approach in their everyday clinical reality. This was stressed in the founding text of EBM (Evidence-Based Medicine Working Group, 1992), and is continuously repeated in current literature on EBM which often comes in a handbook format, written for use in clinical practice. Mechanistic reasoning, being too specialized, does not fit this practically oriented perspective.

Mechanistic reasoning also has an undeniable connection to expert opinion, which is met with equal skepticism in the EBM literature. Already in the JAMA 1992 founding article of the EBM movement, the traditional paradigm that EBM was intended to replace was described as putting “a high value on traditional scientific authority”, where “answers are frequently sought from direct contact with local experts or reference to the writings of international experts” (Evidence-Based Medicine Working Group, 1992:2420). This would all be changed with EBM, it was asserted.

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 is really helpful for finding out what works, but practitioners’ ignorance impedes its use. It would also be unconvincing to claim that since EBM’s main objective is to

---

13For example, Straus et al. (2005), Guyatt & Rennie (2002), and Nordenstrom (2007) are all in handbook format.
find out *what* works (and not *why* it works), mechanistic reasoning should be assigned a low strength of evidence. Surely mechanistic reasoning could also be invoked to support claims *that* something works (or does not work). To this, a critic of mechanistic reasoning could reply: yes, mechanistic reasoning could support such claims, but the support is always weaker than that provided by empirical studies. This seems to be the position taken by EBM supporters Howick, Glasziou & 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”.

The question would then be: why is support from theory always (or, at least, generally) weaker than support from empirical investigations? Such a claim is not self-evident. After all, what we call theories are statements about the workings of the world that normally synthesize and are supported by large amounts of empirical data. Skepticism against theory-knowledge therefore needs to be supplemented with arguments why we should rely on “direct” experimental results but not on generalizations from experimental results.

But maybe I am exaggerating the theoretical depth of the EBM adherents’ argument here. As indicated above, an important argument against mechanistic reasoning – perhaps the argument – is that mechanistic reasoning has gone wrong on several occasions. However, to demonstrate that this argument is too simple, it would be enough to show that under specifiable circumstances, mechanistic reasoning is very likely to be correct. And that is one of the things I try to do in Paper II.

5. Summary of the 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 behavior, 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
introduction

The first requirement says that for each participant it must be possible to tell which treatment has been given. This *distinguishing criterion* (DC) is, a little 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 un-scientific views and theories. Another example is the claim that individually adjusted, and hence non-standardized, 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 it fulfills these two requirements. 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 scientific testing. To the extent that the proponents of alternative treatment methods dismiss the results of scientific tests of effectiveness, their arguments have to be more sophisticated than simply claiming that a scientific scrutiny is not possible.

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

---

14 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.
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 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 characterize different types of negative mechanistic reasoning, which have been rather neglected in the literature. Mechanistic reasoning can be 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 can also be negative in an epistemic sense: if it does not include a mechanistic chain. Three main types of negative mechanistic reasoning, NegA, NegB and NegC, are presented and are characterized 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 characterizations 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 skeptical 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.

6. Prospects for future work

In this final section, a few possible prospects for further studies will be briefly outlined.

6.1 The nature of evidence in EBM

As indicated in section 4.3 above, and also in Paper II below, one could criticize EBM supporters for using a concept of evidence that is too narrow. This is, however, but one question regarding the scope and nature of evidence relevant to EBM. One could ask, for example, whether there are, in any clear and interesting sense, different kinds of evidence, and if so, how they could relate to different ranges of evidential strength. One recent article using the phrase “kind of evidence” is Illari (2011).
6.2 The fate of hierarchies of evidence

One could also focus more strictly on the hierarchies of evidence. As I note in Paper II, if my arguments regarding negative mechanistic reasoning are correct, those that defend the hierarchies have a serious challenge to tackle. Also, Stegenga (2014) has recently argued that hierarchies of evidence should be expelled from EBM altogether. On the other hand, used as rules of thumb (or as “heuristics”), hierarchies of evidence may still make a lot of sense. An analysis of the criticism from Stegenga and others could be performed to give a verdict on the use of hierarchies of evidence within EBM.

6.3 Consistent hierarchies of evidence?

An even more specific question about hierarchies of evidence was touched upon in Section 4.1: since the hierarchies suggested in the literature are not exactly identical, are they really consistent? I suspect that if one carefully answers the following questions for each of the suggested hierarchies (by scrutinizing what the designers themselves say about them), one will probably get answers that are different enough to justify the differences in design:

1. Which are the patients of interest? (E.g., a single patient in front of the practicing doctor or all those that have a specified collection of symptoms?)
2. What is the nature of the outcome measures (are they well-defined or open-ended)?
3. How do we investigate the outcome measure?

As for question 2, I tentatively disagree with Vandenbroucke (2008) who suggests that the important distinction for assessing the validity of a hierarchy of evidence is whether one is interested in risks or benefits of the intervention; and I also tentatively disagree with Osimani (2014) who suggests that it is the distinction intended vs. unintended outcomes that is crucial. I believe that the most informative question to be asked about the outcome measures is whether they are well-defined or open-ended. Of course, a careful argument would be needed to sustain this claim.

6.4 The justification of randomization

As noted in section 2.2 above, there are three main justifications offered in the literature on the use of experimental randomization. Some critics
have rejected all three justifications and have concluded that randomization should not be ascribed any epistemic privilege (Worrall, 2007). Defenders of randomization within EBM have dismissed the critics (e.g., La Caze et al., 2012). It is far from obvious that the matter is settled. Hence, an investigation into the justification of randomization in clinical trials could be considered.

References


42 PHILOSOPHICAL ISSUES IN MEDICAL RESEARCH


