The London School of Economics and Political Science

HIERARCHIES OF EVIDENCE IN EVIDENCE-BASED MEDICINE

CHRISTOPHER J BLUNT

A THESIS SUBMITTED TO THE DEPARTMENT OF PHILOSOPHY, LOGIC & SCIENTIFIC METHOD OF THE LONDON SCHOOL OF ECONOMICS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY, LONDON, SEPTEMBER 2015

DECLARATION

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 79,599 words.

ABSTRACT

Hierarchies of evidence are an important and influential tool for appraising evidence in medicine. In recent years, hierarchies have been formally adopted by organizations including the Cochrane Collaboration [1], NICE [2,3], the WHO [4], the US Preventive Services Task Force [5], and the Australian NHMRC [6,7]. The development of such hierarchies has been regarded as a central part of Evidence-Based Medicine (e.g. [8-10]), a movement within healthcare which prioritises the use of epidemiological evidence such as that provided by Randomised Controlled Trials (RCTs).

Philosophical work on the methodology of medicine has so far mostly focused on claims about the superiority of RCTs, and hence has largely neglected the questions of what hierarchies are, what assumptions they require, and how they affect clinical practice.

This thesis shows that there is great variation in the hierarchies defended and in the interpretations they are, and can be, given. The interpretative assumptions made in using hierarchies are crucial to the content and defensibility of the underlying philosophical commitments concerning evidence and medical practice.

Once this variation is been identified, it becomes clear that the little philosophical work that has been done so far affects only *some* hierarchies, under *some* interpretations. Modest interpretations offered by La Caze [11], conditional hierarchies like GRADE [12-14], and heuristic approaches such as that defended by Howick et al. [15,16] all survive previous philosophical criticism.

This thesis extends previous criticisms by arguing that modest interpretations are so weak as to be unhelpful for clinical practice; that GRADE and similar conditional models omit clinically-relevant information, such as information about variation in treatments' effects and the causes of different responses to therapy; and that heuristic approaches lack the necessary empirical support.

The conclusion is that hierarchies in general embed untenable philosophical assumptions: principally that information about *average* treatment effects backed by high-quality evidence can justify strong recommendations, and that the impact of evidence from individual studies can and should be appraised in isolation. Hierarchies are a poor basis for the application of evidence in clinical practice. The Evidence-Based Medicine movement should move beyond them and explore alternative tools for appraising the overall evidence for therapeutic claims.

For Lynn Blunt

(1955-2014)

There was but one question he left unasked, and it vibrated between his lines: if gross miscalculations of a person's value could occur on a baseball field, before a live audience of thirty thousand, and a television audience of millions more, what did that say about the measurement of performance in other lines of work? If professional baseball players could be over- or under-valued, who couldn't? Bad as they may have been, the statistics used to evaluate baseball players were probably far more accurate than anything used to measure the value of people who didn't play baseball for a living.

—Michael Lewis, *Moneyball*, p.72 [17]

Real evidence is usually vague and unsatisfactory. It has to be examined—sifted. But here the whole thing is cut and dried. No, my friend, this evidence has been very cleverly manufactured—so cleverly that it has defeated its own ends.

—Agatha Christie, *The Mysterious Affair at Styles* [18]

ACKNOWLEDGEMENTS

I would like to thank John Worrall for supervising this thesis, for his continual support and feedback, and for giving the lectures which introduced me to Evidence-Based Medicine in the distant past of my undergraduate days. I also want to thank Roman Frigg for acting as secondary supervisor, and for encouraging me to study for the PhD and to do it at LSE. Huge thanks are also due to Luc Bovens for his support and for the opportunities to teach in the department, without which this thesis and my time at LSE would have been much the worse. Thanks are also due to Alex Voorhoeve, Foad Dizadji-Bahmani and Armin Schulz. I also want to thank Stephen John, Hasok Chang and Tim Lewens for their supervision whilst at Cambridge, and in particular Hasok for giving the first philosophy lecture I ever attended back in 2006. He is almost entirely accountable for my decision to study philosophy of science.

My thanks also go to the Arts & Humanities Research Council and the Philosophy department for funding this work, and to Bill Bottriell for his extremely generous support as an undergraduate. Bill is one of the kindest and friendliest people I have been lucky enough to meet.

Invaluable advice came from my friend and colleague Rosa Runhardt, as well as Catherine Greene and Nicolas Wuethrich. I was fortunate enough to teach a great many undergraduate and masters students over the last four years, and cannot stress enough how valuable that experience was.

Unending gratitude goes to my father Robin Blunt and my mother, Lynn, who was probably the only person in the world who cared about this thesis more than me. I only wish she could read it. Finally, and most of all, thank you to my partner Kate Butler, without whose continual distractions and endless patience this thesis would either have been finished much sooner or never at all.

CONTENTS

Table of Figures

CHAPTER ONE:

HIERARCHIES OF EVIDENCE IN EVIDENCE-BASED MEDICINE

CONTENTS

1.1: The Influence of Hierarchies of Evidence

Evidence appraisal is the process of deciding whether, and to what degree, evidence (either from a single study or an evidence base of many studies) supports a claim. In medicine, evidence appraisals are performed at many levels. Individual practitioners appraise the evidence for the claim that a treatment will be effective for their patient, and that they should recommend the treatment to the patient (see e.g. [19-22]). Public health agencies appraise whether evidence supports the claim that a public health intervention is likely to be beneficial (e.g. [5,23-25]). Regulatory agencies such as the National Institute of Health and Care Excellence (NICE) in the UK appraise whether evidence supports licensing a treatment for a particular condition, compensating that treatment on the National Health Service, and recommending that treatment in their guidelines to clinicians (see [2,26]).

In the last 25 years, hierarchies of evidence have become one of the most influential and important tools in evidence appraisal in medicine. They have been formally adopted by a wide range of governmental agencies including NICE [2,26-28] and the Health Development Agency [29] in the UK, the Agency for Healthcare Research & Quality [30] and Preventative Services task force [5,24] in the USA, the influential Canadian Task Forces on the Periodic Health Examination [31] and Preventive Health Care [32], and the Australian National Health and Medical Research Council [6,33]. International agencies such as the Cochrane Collaboration [1], the Institute for Clinical Systems Improvement (ICSI [34]) and the World Health Organization (WHO [4]) use hierarchies, as do professional bodies such as the American College of Chest Physicians [35], the American Academy of Neurology [36] and the American Thoracic Society [37]. The GRADE Working Group, authors of the influential GRADE hierarchy (see [12,14]), currently cite around 90 organizations which use their model [38].

The Evidence-Based Medicine (EBM) movement has urged individual practitioners to use hierarchies of evidence in their practice. Models of clinical practice published by EBM proponents call for clinicians to perform critical appraisal of evidence as part of routine practice, using a hierarchy of evidence to guide that appraisal (e.g. [8,20,22,39-43]). Elsewhere, EBM proponents have called for practitioners to base recommendations to patients on guidelines formulated through hierarchical appraisal tools such as GRADE (e.g. [35,44-50]).

Evidence appraisal affects which causal claims are believed and acted upon in medical practice and policy. Increasingly, governmental bodies, guideline developers and practitioners are using hierarchies of evidence to perform or support their appraisal. Hierarchical approaches to evidence are therefore extremely influential, affecting which treatments are licensed by governments, promoted in guidelines, and recommended by clinicians.

Despite this, relatively little philosophical work scrutinises hierarchies of evidence themselves. Evidence-Based Medicine has attracted much philosophical discussion. In particular, philosophers such as Worrall [51,52], Upshur [53,54], Bluhm [55,56] and Goldenberg [57] have argued that Randomised Controlled Trials (RCTs) are over-valued by EBM, while others such as Howick [58,59] and Russo & Williamson and colleagues [60-63] have argued that EBM under-values or fails to appreciate the important roles of mechanisms and mechanistic evidence in healthcare (see Chapter 4 for further discussion of these arguments). However, no systematic account of hierarchies has been provided. Without detailed analysis of hierarchies as a structure for evidence appraisal and the fundamental assumptions within a hierarchical approach to evidence, it is unclear whether hierarchies really commit users to such philosophically-problematic claims about RCTs and mechanistic reasoning. It is also unclear whether there are other philosophical assumptions about evidence built into hierarchies which could affect evidence appraisal and thus clinical practice.

It is this lacuna in the literature which this thesis addresses. This thesis presents a novel philosophical analysis of hierarchies of evidence in Evidence-Based Medicine, offering a philosophical framework for the analysis of hierarchies, and presenting the first history of hierarchies. It then considers whether the common criticisms of hierarchies and of EBM truly apply to hierarchies in general. It will be demonstrated that these objections apply only to certain interpretations of hierarchies.

The fundamental philosophical assumptions required to use hierarchies as the primary method of evidence appraisal are then considered. Some hierarchies commit users to the untenable assumption that an accurate appraisal of evidence can be performed by appraising the evidence from individual studies in isolation, or straightforwardly aggregating study results. Other hierarchies, including GRADE, focus attention on evidence for estimates of the average treatment effect in some population. These assumptions lead to important, patient-relevant information being overlooked—notably, information about the variation in treatment effects and the causes of this variation. I show that paying attention to this information improves care in a number of cases from across medical specialities. Therefore, a philosophical justification for hierarchical approaches to evidence in clinical practice cannot be provided. The only justification remaining for hierarchies of evidence would be an empirical one—demonstrating that their use does improve the care which patients receive, despite philosophical shortcomings. However, the evidence base for such a justification is currently weak. Therefore, this thesis concludes that hierarchies of evidence should, in the absence of such evidence, be removed from the EBM movement's approach to appraisal—both in clinical practice and guideline development.

This position is not opposed to Evidence-Based Medicine or the aims of that movement. I am broadly sympathetic with their goals. Hierarchical approaches to evidence are not a necessary component of an EBM approach. The EBM movement can and should reject the philosophical assumptions which underpin hierarchical approaches, and devote their efforts to developing alternative methods of evidence appraisal, rather than to tweaking hierarchies.

1.2: Evidence-Based Medicine

Evidence-Based Medicine (EBM) is a social movement within clinical medicine and biomedical science, which arose from the nascent field of "clinical epidemiology" in the early 1990s [39,64]. The central goal of the EBM movement has been to promote and increase the use of the evidence provided by epidemiological studies within clinical medicine. The most famous definition of EBM is given by a major pioneer of EBM, David Sackett, and colleagues:

> *"Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients."* ([22], p.71)

This definition has been repeated across many publications, and returns over 30,000 hits in its entirety on Google. Some of the most prominent examples include the Cochrane Collaboration's website [65], numerous EBM proponents (e.g. [8,43,44,66,67]) and critics alike (e.g. [53,57,68-70]). Unfortunately, this attempted definition is quite uninformative ([57,69-72]). As Worrall has observed, medicine has surely always been based upon evidence of some form ([51], p.316). The key question is what constitutes "*current best evidence*", which this platitudinous definition fails to address.[1]

What differentiates EBM from previous approaches to clinical practice is an emphasis upon evidence from particular sources. EBM has become associated with the view that Randomised Clinical Trials (RCTs) are the strongest and highest-quality form of evidence for use in clinical practice (see e.g. [55,57,74-76]). Sometimes this claim has been expressed very strongly, as in the first edition of Sackett et al.'s textbook *Evidence-Based Medicine*:

> *"If the study wasn't randomised, we suggest that you stop reading it and go on to the next article in your search."* ([22], p.94)

The movement is also associated with the belief that clinical experience and mechanistic reasoning are poor sources of clinical evidence [8,22,77]. In particular, EBM proponents have argued against relying upon the judgments and expertise of senior clinicians in determining whether particular treatments are effective [20-22,40,78]. In its early days, EBM was often presented as a radical repudiation of the authority of senior clinicians—every clinician could and should be challenged to provide clinical trial evidence to support her claims (e.g. [20,39]).

[1] Variations on this definition include Guyatt's "*the application of scientific method in determining the optimal management of the individual patient*" [64]. Daly reports Guyatt's updated definition as: "*Evidence-based medicine is an approach to practicing medicine in which the clinician is aware of the evidence in support of her clinical practice, and the strength of that evidence*" ([73], p.89). These definitions largely suffer the same problems—they are uninformative and fail to distinguish EBM from alternative approaches to evidence in clinical medicine.

1.2.1: EBM AS A NEW PARADIGM

The initial presentation of EBM in 1992 by the EBM Working Group was as a "*new paradigm*" in clinical medicine and medical pedagogy [39]. The term 'paradigm' is ambiguous [79]. However, the features commonly expected of a paradigm include: a set of foundational assumptions agreed upon by all within the discipline; a way of viewing the content of the discipline and the problems it can and cannot solve; and, the methods which are and are not appropriate for use in solving those problems (see [80]). In particular, since Kuhn [80], a paradigm shift suggests quite a radical departure from what proceeds it.

However, it has become clear that early EBM proponents did not intend to endorse a Kuhnian philosophy of science, or to claim that medical practice undergoes Kuhnian revolutions—periods of 'normal science' interrupted by paradigm shifts due (in large part) to the accumulation of anomalies (see [70]). It is more likely that the term 'paradigm shift' was used in a colloquial sense to indicate a major change of ideology or practice [72,81]. A number of prominent EBM proponents subsequently dissociated themselves from the term 'paradigm', and some acknowledged that it was used without reflection [82-86]. A number of critics have addressed the problems of whether clinical practice can be aptly described as paradigm-based, whether Kuhnian history applies to medical practice, and whether the transition to EBM should be described as a Kuhnian paradigm shift (see e.g. [70,75,87-89]). A convoluted interpretation of Kuhn is required to force a fit. With a few exceptions,[2] most EBM proponents have abandoned a 'paradigm' conception of EBM in favour of an account based on a process or a set of principles.

1.2.2: THE PRINCIPLES OF EBM: 'FIVE LINKED IDEAS'

Attempts have been made to define EBM as a set of principles. A very prominent early account comes from Davidoff et al. [21] and Sackett & Rosenberg [20], called the "*five linked ideas*" of EBM:

> *"(1) that our clinical and other health care decisions should be based on the best patient- and population-based as well as laboratory-based evidence;*

[2] Bhandari still defends a paradigm definition [90-92]. Another possible exception is Gordon Guyatt, who continued the paradigm terminology in an interview with Jeanne Daly ([73], pp.88-9).

> *(2) that the problem determines the nature and source of evidence to be sought, rather than our habits, protocols or traditions;*
>
> *(3) that identifying the best evidence calls for the integration of epidemiological and biostatistical ways of thinking with those derived from pathophysiology and our personal experience;*
>
> *(4) that the conclusions of this search and critical appraisal of evidence are worthwhile only if they are translated into actions that affect our patients;*
>
> *(5) that we should continuously evaluate our performance in applying these ideas."*
>
> ([20], p.332)

Several of these principles are uncontroversial and laudable. (2)'s insistence that tradition is not sufficient justification for reliance upon particular evidence-sources is an accepted assumption in philosophy of medicine, while (5) defends performance evaluation, which is a worthy ideal. (4) is a practical concern definitive of the art of clinical practice, as opposed to scientific research (which may be worthwhile even without practical applications). (1) and (3) are the slightly more controversial aspects of this account, suggesting that *"laboratory-based"* evidence and *"ways of thinking ... derived from pathophysiology and our personal experience"* are an insufficient basis for clinical practice. However, in Sackett & Rosenberg's account, the claim is fairly innocuous—epidemiological evidence is to be *integrated with* clinical experience and pathophysiology (though no indication of how this integration should be performed is offered). However, elsewhere, the EBM movement denigrated clinical experience and pathophysiology, and subsequently argued for the subordination of non-epidemiological evidence to epidemiological evidence (e.g. [10,22]).

1.2.3: Epidemiological, Physiological and Pre-Appraised Evidence in EBM

Note that in this context, *'epidemiological and biostatistical'* evidence refers to evidence from controlled clinical studies. In the EBM literature, 'epidemiological evidence' comes from RCTs and observational studies (see e.g. [19,22,78,93]).[3] In an RCT, a trial population is selected and then randomly allocated into two or more study groups. In the simplest design, a two-arm trial, one group is given the 'experimental treatment'—the treatment of interest—and the other is given a control

[3] It would be more accurate to describe these as 'clinical epidemiological' sources. The clinical epidemiology which underpins the EBM movement is distinct from traditional epidemiological research (see Chapter 3, and [73,94-96]).

treatment, either a rival treatment or a placebo. These are labelled the 'experimental group' and 'control group' respectively. Researchers then compare the results on various outcomes of interest between the two groups. Various designs fall under the 'Observational study' label (see [74]). Unlike RCTs, observational studies are non-interventional—researchers do not interfere with patient care, merely observing their outcomes. In a cohort study, two populations who receive different treatments are observed, and their outcomes compared to try to detect differences in the effects of the treatments. In a case-control study, two groups who differ in the outcome of interest are selected, and studied to observe whether their outcomes correlate with some exposure or intervention.[4]

In addition to primary studies, the EBM literature also emphasises "*pre-appraised evidence*" ([10], ch.2). This can refer to reports which independently critically appraise the evidence provided by a study, such as "CATs" (Critically Appraised Topics, a learning method advocated by some EBM proponents [41,98,99]) or reports in summary journals such as the ACP Journal Club [100,101]. It also refers to meta-analyses and systematic reviews, which are often ranked highly in hierarchies (e.g. [6,15,28,43,102-106]). The Cochrane Collaboration distinguishes between systematic reviews and meta-analyses [107]. In a systematic review, researchers apply explicit criteria to find and appraise studies of the treatment of interest, and offer recommendations on the basis of a critical overview of the evidence [1,108]. Meta-analysis is a statistical approach that can be used as part of a systematic review, aggregating or amalgamating evidence from comparable studies to provide estimates of the treatments' effects across the range of studies [107]. A simple form of meta-analysis takes a weighted average of the studies' results, although considerably more complex analyses can be conducted [109,110] (see also Chapter 6).

Evidence that is "non-epidemiological" in the sense intended here includes evidence from clinical experience and expert opinion, and evidence from laboratory-based research. Clinical experience refers to unsystematic observations in clinical practice (see [59]). Expert opinion includes both statements from lecturers and textbooks (see [66,92,111,112]), as well as consensus guidelines from consensus conferences, which have been criticised by numerous EBM proponents (e.g. [92,104,113-115]) and philosophers (e.g. [59,87,116]).

"*Laboratory research*" [20,55,63,117,118] is one of a collection of similar terms used in the EBM literature to refer to a broad range of studies. These include studies of physiological (as opposed to 'clinical') effects (i.e. studies of effects on properties at the cellular or system level measured in

[4] Note that there are non-randomised interventional study designs, including non-randomised controlled trials, in which the experimenters select two groups through some non-random procedure and give one group the experimental and the other the control treatment. Other semi-interventional designs include historically-controlled studies, in which experimenters intervene to give a group of patients an experimental treatment, and use observed data from past patients as the control group (see [51,74,75,97] for further discussion of the uncertain role of non-randomised interventional studies in EBM).

laboratory tests, biomarkers of disease) [119]. Pharmacokinetic studies measure properties of a drug when administered, such as the concentration in the bloodstream or the elimination half-life [43,120]. In vitro and animal models measure effects outside human systems—for instance, studies of the effects of an antibiotic on a colony of bacteria [43]. Genetic studies include tests for association between genetic markers and disease, receptiveness to treatment, and prognosis. Other terms used include "*bench research*" [55,103,121-123], "*basic science*" [8,63,78,121,124], and "*pathophysiologic studies*" [112].[5] These terms are very rarely defined, and no attempt has been made to distinguish them systematically. They are commonly used as synonyms in the EBM literature (see [121]). The phrase "*based on physiology, bench research or first principles*" was introduced into hierarchies by Ball & Phillips [130,131], and has been replicated many times across many hierarchies [2,15,103,132-135], always at the lowest level.

Sometimes, terms such as "*pathophysiologic rationale*" [8,10,85,136] and "*first principles*" [15,103,130,131] are used to refer to clinicians providing a physiological explanation of why a treatment should work, and using this rationale as evidence that it will work. Conflating pathophysiological *evidence* with a hypothesised mechanism should be avoided. However, many hierarchies do not distinguish the two. The process of specifying the mechanism by which a treatment has an effect and providing evidence for each step in that process is sometimes called "mechanistic evidence" or "mechanistic reasoning" [58-63,137]. Chapter 4 discusses challenges to the EBM approach to mechanisms in more detail.

1.2.4: Hierarchies as a Fundamental Principle

While the "*five linked ideas*" illustrate much of EBM's core ideas (the focus upon practice, the insufficiency of laboratory evidence, the importance of epidemiological evidence), they do not express the mechanics of EBM in practice. They offer no indication as to *how* evidence is appraised, integrated or applied.

A second 'principles' approach was very prominently defended by Guyatt et al. in the first chapters of the book compilation of the *Users' Guides to the Medical Literature* series [8].[6] In their

[5] Physiological and laboratory studies should not be confused with physiological and laboratory *outcomes*. Epidemiological studies can include physiological measurements [55]. A number of EBM sources have expressed concerns about the use of physiological and laboratory surrogate outcomes in clinical trials, in which laboratory test values are used as the outcome of interest in trials. For instance, a study of swelling in arthritis patients might use C-reactive protein (CRP) levels as a laboratory outcome (*cf.* [125,126]). Concerns have been raised about inferences from effects demonstrated on these laboratory values to effects on "patient-oriented outcomes" [127,128]—things patients actually care about, like quality of life, mortality and mobility (e.g. [1,8,36,104,129]).

[6] This approach originated in *Users' Guides* XXV [82].

chapter entitled *The Philosophy of Evidence-Based Medicine*, Guyatt et al. summarise EBM through two "*fundamental principles*":

> *"EBM involves 2 fundamental principles. First, EBM posits a hierarchy of evidence to guide clinical decision making. Second, evidence alone is never sufficient to make a clinical decision."*
>
> ([83], p.10)

Here, hierarchies take centre-stage: a hierarchy of evidence guiding clinical decision-making is the foremost principle of EBM. Notably, hierarchies were absent from the "*five linked ideas*". Guyatt et al.'s hierarchy places epidemiological evidence at the top (notably RCTs), with physiologic studies and clinical experience at the bottom (see Fig.1, below). The second principle, the insufficiency of evidence, is a concession to the importance of patients' preferences and input: decisions must be made *with* the patients, rather than *about* them, and their values and preferences should be taken into account. Notably, no further mention is made of this second "fundamental principle": the remaining *Users' Guides* offer no substantive guidance for the interpretation and integration of patients' values [8]. Hierarchies, then, were the most important, defining feature of EBM for Guyatt et al. Moreover, the notion of "integration" of epidemiological and non-epidemiological evidence has been abandoned in this account.

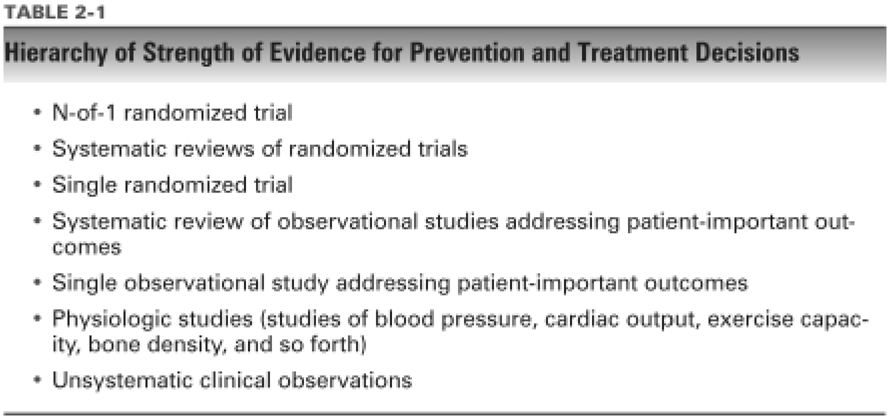


TABLE 2-1

Hierarchy of Strength of Evidence for Prevention and Treatment Decisions

- N-of-1 randomized 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

Figure 1: A hierarchy of evidence described as the first 'fundamental principle' of EBM by Guyatt et al. ([83], p.10)

Karanicolas, Kunz & Guyatt developed these principles:

> *"First, systematic summaries of the highest quality available evidence should inform clinical decisions. Second, wise use of the literature requires a sophisticated hierarchy of evidence. Finally, evidence alone is never sufficient to make clinical decisions; rather, it requires trading off benefits*

> *and risks, inconvenience, and costs, and in doing so considering patients' values and preferences."* ([138], p.1067)

In this iteration, the authors place more emphasis upon individual clinicians using pre-appraised evidence, which in turn should be created using a hierarchy. This shift from conceiving of clinicians as evidence-appraisers to evidence-users is typical of recent developments in EBM approaches (see Chapter 3). Note that here a hierarchy of evidence is necessary for "wise use" of the medical literature.

1.2.5: 'Doing EBM'—EBM as a Process

A final approach to defining EBM describes it as a process rather than a set of principles—EBM is something *done* as opposed to something *believed*. Straus et al. [10] distinguished between three 'modes' of EBM—the "doing mode", the "using mode" and the "replicating mode": respectively, clinicians can appraise evidence themselves, use appraisals performed by others, or replicate decisions made by evidence-based practitioners. The most influential process definition originates from Rosenberg & Donald [40], and Sackett & Rosenberg [20], defining EBM as a four- or five-step process: [7]

1. Setting the Question
2. Finding the Evidence
3. Appraising the Evidence
4. Applying the Evidence
5. Evaluating Performance

This approach was prominently adopted by the so-called "Bible" of EBM [76,141], Sackett et al.'s *Evidence-Based Medicine* [22], which is structured into five parts, each devoted to explicating a step in the process.

Hierarchies were introduced into EBM as a tool for performing step 3, appraising the evidence—they give information about the quality, strength, etc. of the evidence (see Section 2.2). But a second role for hierarchies has also developed in step 4, applying the evidence. These tools attempt to draw a connection between the appraisal of evidence and a recommendation to patients. Often, they straightforwardly convert from high-quality evidence to a strong recommendation. The primary purpose of hierarchies within this role is to appraise the evidence for predictions about the likely effects of treatments on individuals. One example of this straightforward correspondence comes from Chesson et al.'s [142] AASM hierarchy (see Figure 2, below).

[7] Other EBM processes have been presented, including six-step extensions of the process (e.g. [139,140]).

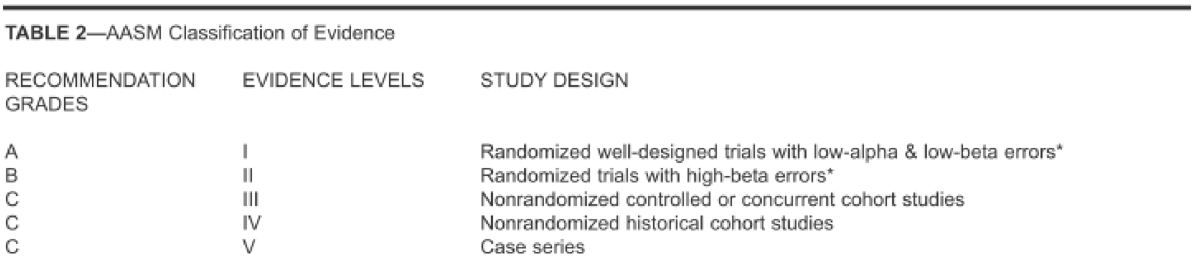


TABLE 2—AASM Classification of Evidence

RECOMMENDATION GRADES	EVIDENCE LEVELS	STUDY DESIGN
A	I	Randomized well-designed trials with low-alpha & low-beta errors*
B	II	Randomized trials with high-beta errors*
C	III	Nonrandomized controlled or concurrent cohort studies
C	IV	Nonrandomized historical cohort studies
C	V	Case series

Figure 2: A hierarchy of evidence from Chesson et al. [142] in 1999, in which there is a clear correspondence between the evidence level and the strength of a recommendation that can be made on that basis.

Hierarchies also gained a role in step 2—finding the evidence—as part of a search strategy [10,16]. According to most interpretations of hierarchies, higher-ranked studies provide evidence with some epistemically-desirable property (or to a higher degree than lower-ranked studies). Therefore, where time and resources are limited, clinicians may prioritise reading reports from high-ranked studies. In particular, Brian Haynes (see [101,143-147]) has worked on developing optimal literature search strategies to identify the likely best evidence, and developed hierarchies which rank sources of pre-appraised and aggregated evidence such as the '4S' system (see Figure 3, below).

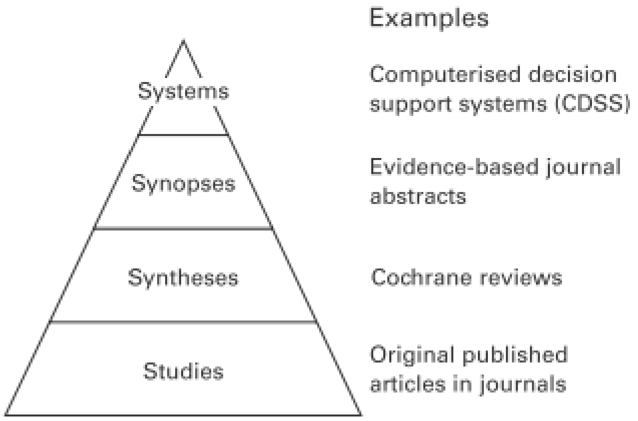
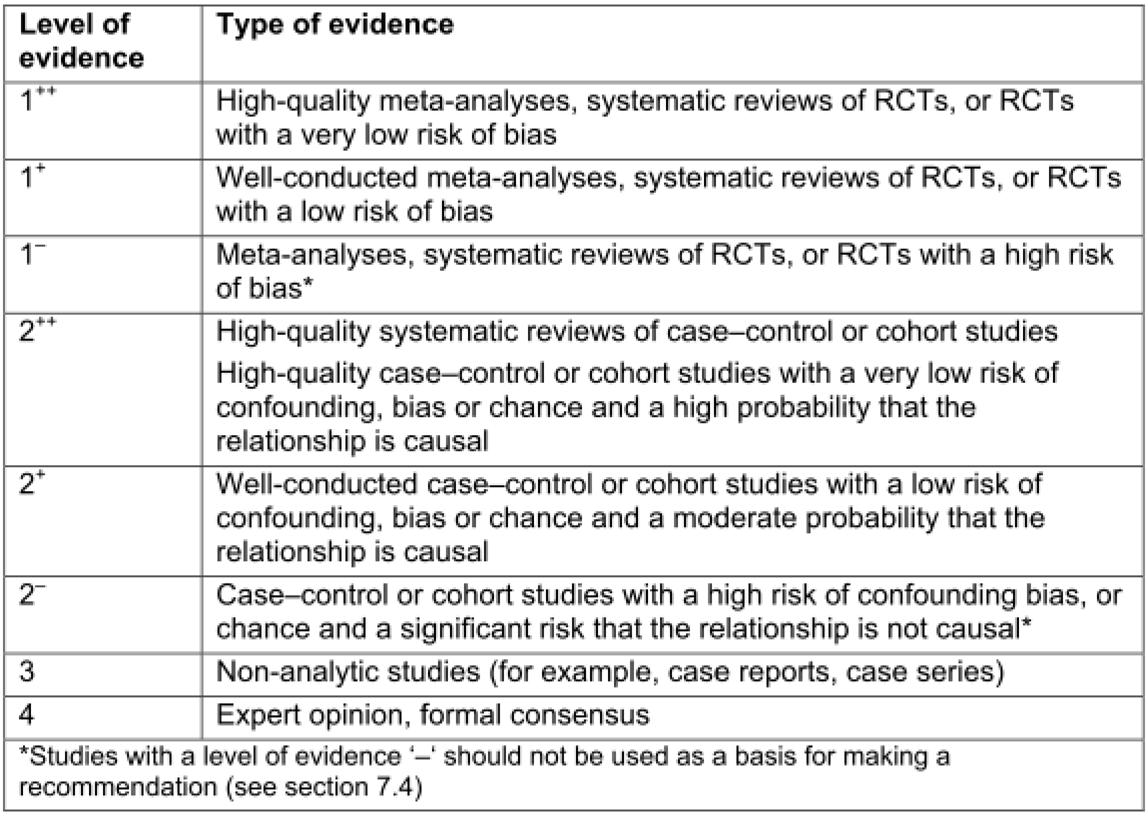

Examples
Systems
Computerised decision support systems (CDSS)
Synopses
Evidence-based journal abstracts
Syntheses
Cochrane reviews
Studies
Original published articles in journals

Figure "4S" levels of organisation of evidence from research.

Figure 3: The '4S' system from Haynes 2001 [145], a hierarchy which primarily ranks sources of pre-appraised and aggregated evidence, designed to facilitate optimal search strategy. Subsequently, Haynes and colleagues have developed the system to 5S and 6S iterations [101,146].

Hierarchies are also used by guideline development agencies and government bodies to standardise the way evidence is appraised. If hierarchies can be used in a replicable way, they could prove a useful tool in providing a standard approach to evidence which is not affected by individual researchers' or policy-makers' preferences, offering a measure of objectivity in guideline development. Hierarchies here again play a role at two stages—in appraising the quality or strength of the evidence from studies, and in deciding how strong a recommendation can be justified on the basis of some evidence. One example of a hierarchy used within guideline development is the SIGN (Scottish Intercollegiate Guidelines Network) hierarchy [134], which has been used by NICE (e.g. [2,28]) amongst others (see Figure 4, below).



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

*Studies with a level of evidence '–' should not be used as a basis for making a recommendation (see section 7.4)

Figure 4: The SIGN hierarchy as used by NICE in the 2004 Guideline Development Methods manual [2].

Hierarchies of evidence, then, play several roles within EBM processes, and have been enshrined in some accounts of the fundamental principles of EBM. Whether applied by a guideline development agency or individual clinicians, the most important role of hierarchies, and the focus here, is in appraising the evidence for predictions of likely effects of treatments on patients. However, EBM certainly should not be equated with a hierarchy of evidence. The goals of EBM include integrating epidemiological evidence into clinical practice and decreasing reliance on clinical experience, authority and laboratory-based research. These goals can be pursued without a hierarchy of evidence. Chapter 2 will show that there is a great deal of variation in the ways that hierarchies can be interpreted. Some interpretations may be at odds with the ideas of many EBM proponents. Chapter 6 will argue that hierarchies of evidence introduce certain assumptions into the evidence appraisal process which EBM proponents need not accept.

1.3: Hierarchies of Evidence

To date there have been no detailed studies of hierarchies of evidence. Several philosophers have criticized hierarchies of evidence, but most focus upon particular choices made in formulating hierarchies—for instance the decision to rank RCTs above observational studies, or meta-analyses above RCTs—and whether a successful philosophical justification for these decisions can be provided (see e.g. [11,57,72,75,76,124,148-150]).

In order to investigate the philosophical assumptions and consequences of hierarchical approaches to evidence, some more basic—but equally un-researched—questions must be answered. What is a hierarchy of evidence? How have they been used and interpreted, and how could they be interpreted differently by different users? Is there a single hierarchy of evidence defended by the EBM movement? This chapter addresses the first question, offering a definition of hierarchies of evidence, while the latter questions will be the focus of Chapter 2.

The Users' Guides defines a hierarchy of evidence as:

> *"A system of classifying and organizing types of evidence, typically for questions of treatment and prevention."* ([8], p.787)

This definition does not distinguish hierarchies from other non-hierarchical approaches to evidence such as the early EBM checklists (e.g. [22,151]) and QATs (Quality Assessment Tools—see [152]) such as CONSORT [153,154], which classify the evidence from studies according to certain properties. The unique features of a hierarchy which differentiate them from other appraisal tools are: hierarchies rank the evidence provided by different sources, do so primarily according to the methodology used, and are concerned with making epistemic judgments about that evidence, not merely grouping them according to design or purpose (cf. [155,156]). For the purposes of this thesis, a hierarchy of evidence is defined as:

> A procedure for making an inference about some epistemic property of the evidence for some clinically-important claim provided by studies, or some epistemic relation which holds between the evidence for some clinically-important claim provided by two or more studies, primarily based on a ranking of the design or methodology used in the studies.

It is important to clarify several points in this definition. First, the term “primarily” is used in the sense of ‘first and foremost’: a procedure *primarily* bases an inference on a ranking of study designs if and only if the ranking of the study’s design is the first consideration *and/or* the most important consideration in making that inference. So a procedure which makes an initial attribution of a value to some property of some evidence on the basis of the study-design, and then tweaks this value on the basis of some non-methodological factors, uses information about design in the relevant sense of ‘primarily’. So does a procedure which uses several criteria, some methodological, some not, and which assigns the greatest weight to the methodological concerns in making that attribution.

The question of whether the procedure results in a claim about the *properties* of the evidence, or about a *relation* between two or more pieces of evidence is left open for now (see Section 2.3). The properties or relations must be ‘epistemic’ in that they should be relevant to knowledge-generation. The properties currently used in hierarchies include ‘quality’, ‘strength’, ‘weight’, ‘validity’, ‘accuracy’, ‘bias’, ‘risk of bias’, ‘level of evidence’ and ‘precision’ (see Section 2.2 for discussion). The relations are comparative relations over these absolute properties—e.g. ‘is higher quality than’, ‘is stronger evidence than’, and ‘has a lower risk of bias than’.

The procedure need only be used to make an inference *about* some epistemic property or relation. This inference may be that the evidence *has* or *lacks* some property or relation (e.g. ‘is valid’, ‘is invalid’), or that it has a specific degree of that property (e.g. ‘high’ quality). However, it also might be a more qualified inference about the property or relation—for instance, that a relation *probably* holds, or that *given certain circumstances*, a property or relation will hold (or hold to a certain degree). Crucially, it need not be the case that the inference assigns a definite value to the property or relation to be an inference *about* that property or relation. It need only give *information* about the property or relation with respect to the evidence.

1.4: "EVIDENCE" IN EVIDENCE-BASED MEDICINE

There has been some debate about the meaning of "evidence" in EBM (see e.g. [53,57,86,149,157-160]). Few EBM proponents have offered clear accounts of what they mean by evidence or engaged with philosophical literature concerning evidence. There are two related questions here: (1) what does it mean for E to be evidence for/against a hypothesis H, and (2) what kind of things are evidence?

1.4.1: PHILOSOPHICAL ACCOUNTS OF EVIDENCE

One may distinguish between 'evidence' simplicter, and 'evidence for/against' some hypothesis [161]. Suppose a judge is presented with the evidence that the defendant's fingerprints were found on a gun, and that bullet casings from the same ammunition used by that gun were present at a murder scene. Colloquially, this might be termed "evidence", without reference to any hypothesis. These are simply items that have been laid before the judge. But this evidence is also *evidence for* the hypothesis that the defendant committed the murder (albeit weak evidence for that hypothesis).[8]

Clarke identified technical senses of 'data' and 'evidence' in the EBM literature [162]. 'Data' are factual claims about individuals. For instance, the report that 'The patient is experiencing pain in her foot' is a data point, as is 'The patient has 45 repeats of glutamine beginning at the 18th amino acid in her Huntingtin gene'[9]. A set of data about a specific individual is a 'patient profile', and a set of data-points about the same property of different individuals is a 'data set'. According to Clarke, 'evidence' in the EBM literature refers to the results of subjecting data sets to some statistical test. The most common way of obtaining evidence is to aggregate some data via some statistic and test for a correlation between certain variables.

Clarke rejects this account of evidence.[10] This account has more in common with the account of evidence simpliciter, rather than evidence for or against hypotheses. To avoid confusion, this thesis will follow Clarke in holding that an aggregation of data subjected to a statistical test results not in

[8] The 'evidence' is also evidence for and against many other hypotheses—it is, for instance, quite good evidence that the defendant once held the gun in question. Evidence simpliciter can provide evidence for or against many hypotheses. This point is discussed in more detail in Chapter 5.

[9] This patient will develop Huntington's disease. Here the data point about her mutant form of the Huntingtin gene is evidence that the patient will develop Huntington's.

[10] Note that this is not Clarke's account of evidence, but rather the account of evidence implicit in the EBM literature which Clarke has identified. Clarke argues against this account on other grounds to those given here, in particular a lack of generalisability to evidence-based policy (see also [163]).

evidence, but in descriptions of *phenomena*[11], or *research findings*. For instance, in the Huntington's case, one phenomenon is: 'that all patients with >30 repeats of glutamine in the Huntingtin gene develop Huntington's disease.' This phenomenon is not termed *evidence* on its own, or referred to as 'evidence' in the abstract. Where phenomena are evidence, the phenomena must affect belief in some hypothesis. This account of evidence coheres better with other terminology used by the EBM movement. As discussed in Chapter 2, hierarchies of evidence attribute properties such as strength and quality to evidence. These properties relate to effects on beliefs. Therefore, an account of evidence as a relationship between phenomena and some belief is preferable.

The commonplace minimal definition of evidence is 'that which justifies beliefs' [161]. This provides an initial answer to question (1): saying 'E is evidence for H' means that E helps to justify believing H. This account appears in the EBM literature. Goldenberg follows Goodman [165] in defining evidence as *"some conceptual warrant for belief or action"* ([57], p.2621). Similarly, Upshur defines evidence as:

> *"an observation, fact or organized body of information, offered to support or justify inferences or beliefs in the demonstration of some proposition or matter at issue"* ([158], p.93).

A further position, known as evidentialism [161], holds that for a reasonable agent, their body of evidence (plus the range of hypotheses which she considers) determines the range of hypotheses that the agent may rationally believe (see [166]). According to this view, two purely rational agents with access to the same evidence base, and considering the same hypotheses, are constrained to believe hypotheses from the same set. One may formulate a strong version of evidentialism, according to which an agent's evidence (plus the range of hypotheses they consider) uniquely determines what they may rationally believe. According to this strong evidentialist position, two rational agents with the same evidence, considering the same set of hypotheses, would have identical beliefs.

Alternative accounts of evidence include Bayesian confirmation theory. According to Bayesians, evidence is that which confirms or disconfirms hypotheses (*cf.* [167]). Belief in a hypothesis H is quantified as P(H), and a rational agent's beliefs must be consistent with the axioms of probability theory [168]. Beliefs are updated in the light of evidence according to a rule of conditionalisation. The rule of conditionalisation states that if $P_T(H)$ represents one's belief in H at time T, and one comes to believe new information E, then one must conditionalise one's belief in H on E—i.e. $P_{T+1}(H) = P_T(H|E)$. E provides evidence for/against H just when one's belief in H conditional on E is different to one's prior belief in H (i.e. $P_T(H) \neq P_{T+1}(H)$). Where the degree of belief in H increases, E is evidence *for* H, and where it decreases, E is evidence *against* H.

[11] 'Phenomena' is used here in the sense defined by Bogen & Woodward [164].

Bayes' Theorem circumscribes how this conditionalisation is performed. Bayes' Theorem states (where P(E)≠0):

$$P(H|E) = {P(H) \times P(E|H)}\Big/{P(E)}$$

A corollary of Bayes' Theorem and the conditionalisation rule is (again, where $P_T(E)$≠0):

$$P_{T+1}(H) = {(P_T(H) \times P_T(E|H))}\Big/{P_T(E)}$$

In Bayesian confirmation theory, $P_{T+1}(H)$ is referred to as the posterior probability of H, $P_T(H)$ as the prior probability of H, $P_T(E|H)$ as the likelihood of E given H, and $P_T(E)$ as the expectedness of the evidence.

Notably, many Bayesian accounts do not require that two purely-rational agents have the same degree of belief in hypotheses given the same evidence base (*cf.* [167,168]). As the corollary to the conditionalisation rule shows, $P_{T+1}(H)$ in part depends on $P_T(H)$, the prior probability ascribed to H. How these priors are set is a matter of considerable debate in Bayesian theory, but some approaches including contemporary subjective Bayesianism allow different rational agents to begin with different priors (see e.g. [167,169,170]). Subjective Bayesian confirmation theory is compatible with a form of weak evidentialism. However, it is inconsistent with the strong evidentialist view described above. Subjective Bayesianism allows for rational agents to reach different beliefs given the same evidence—their priors also affect beliefs.

Other accounts of evidence have been developed. For instance, Popper's [171,172] falsificationist approach rejects confirmation. Evidence for H, or corroboration of H, comes from surviving severe tests of H—attempts to falsify H, i.e. tests which would lead to the rejection of H given a specific outcome. A thorough review of accounts of evidence is beyond the scope of this thesis.

Worrall states a criterion for accounts of evidence which is consistent with each of these approaches. He says:

> *"in order for evidence to count at all strongly in favour of a theory it must not only be accounted for by the theory, it must also be 'otherwise improbable'"* ([76], p.358).

Being otherwise improbable requires that alternative theories do not adequately account for the evidence:

> *"really telling evidence for any claim is evidence that at the same time tells against plausible rival alternative hypotheses"* ([75], p.1015).

Bird (e.g. [173,174]) has similarly argued for this constraint in confirming causal hypotheses:

> *"to know a hypothesis to be true we must at least have evidence that eliminates its plausible competitors"* ([174], p.242)

This intuition is accommodated by each of the theories discussed above. Evidentialism accepts that E will not strongly justify believing H if alternative hypothesis *H'* also accounts for E. Bayesianism incorporates this rule in the updating procedure, as the probability of the evidence $P_T(E)$ decomposes into terms which include $P_T(E|\neg H)$—so, if E is still likely even if H is false, E will have a smaller effect on degree of belief in H. Falsificationism directly embraces this doctrine in requiring a severe test of H.

1.4.2: Accounts of Evidence in EBM

The EBM literature accepts that evidence is that which justifies belief, and appears to accept the constraint identified by Worrall. A few sources present a Bayesian account of evidence (e.g. [86,113,119,175]), but explicit Bayesian reasoning is rare in the EBM literature. The *Users' Guides* use a Bayesian framework to discuss diagnostic reasoning [176,177], but not in the context of treatment decisions. It is more common for EBM texts to suggest that rational beliefs equate to evidence-based beliefs, in sympathy with evidentialism. A falsificationist account has sometimes been suggested (e.g. [86,119]), although this has been more strongly identified with classical statistical hypothesis testing rather than with Popper's account.

The most detailed consideration of notions of evidence by EBM proponents comes from Djulbegovic, Guyatt & Ashcroft [86], who survey three potential accounts of evidence in EBM—falsificationist, inductive and abductive—and argue that EBM is consistent with each. They adopt a neutral position with respect to analyses of evidence, claiming that EBM is compatible with whichever theory of evidence is preferred. They emphasise that "evidence" plays a number of roles within EBM:

> *"there is no epistemological "one size fits all" approach that applies to the entire practice of EBM"* ([86], p.159).

They distinguish three roles of evidence in EBM: evidence as a minimal requirement (i.e. 'the evidence' is a set of incontrovertible observations or facts with which any theories or beliefs that practitioners hold must be consistent); evidence as a justification for belief (i.e. evidence is any information which confirms or disconfirms a hypothesis); and, evidence as a neutral arbiter (i.e. evidence provides an objective way of deciding which of a range of medical practices is superior to the others—it causes different practitioners to converge ultimately on the same beliefs[12]).

However despite this stated neutrality, EBM seems most naturally allied with an evidentialist approach. A falsificationist model is particularly at odds with the overwhelming tendency to cast hierarchies and appraisal in terms of the evidence *for*, rather than against, hypotheses (see e.g. [2,8,10,12,24,134]). There are a number of reasons to believe that EBM favours an evidentialist concept of evidence even to a Bayesian approach. First, as Worrall [75,76] has pointed out (and see [167]), Bayesians see no special role for randomization (see Chapter 4 for further discussion of whether a special epistemic role for randomization is assumed in EBM). Secondly, EBM proponents, like many within medicine, usually assume a classical hypothesis testing framework (see [178,179]). It remains the norm to use Fisherian statistical significance testing and report the rejection or acceptance of the null hypothesis at a specific threshold, or confidence interval, rather than to report the probability of the hypothesis in the light of the evidence, as a Bayesian approach would suggest (see e.g. [1,10,180]).

Finally, and most significantly from the perspective of hierarchies of evidence, many hierarchical approaches ascribe absolute levels of belief to hypotheses given a set of evidence E (e.g. [2,6,12,28,35,133]). This position seems incompatible with subjective Bayesianism—a subjective Bayesian account would require that the prior probability of the hypothesis be taken into account. It is compatible with many Bayesian positions for the same evidence base to entail a high level of belief in a hypothesis given a high prior, but only a moderate level of belief given a low prior. No EBM system appears to take this possibility seriously. Although priors may be washed out after a suitably large range of evidence is considered, each of the hierarchies identified here allows these attributions to made on the basis of a single result. As such, an evidentialist approach seems to be the most compatible with the EBM literature. In particular, the strong evidentialist position, which is incompatible with subjective Bayesianism, seems to be suggested: evidence uniquely determines what practitioners should believe. The definitions given of 'quality of evidence' reinforce this evidentialist interpretation. For instance, the GRADE Working Group state:

> *"the quality of evidence indicates the extent to which we can be confident that an estimate of effect is correct"* ([12], p.1490; see also [14]).

[12] Care should be taken to ensure that this is not interpreted in a simplistic way—EBM proponents do not believe that evidence is prior to or independent of theory. Studies are designed specifically to test and arbitrate between rival hypotheses.

The only thing that explains confidence in an effect hypothesis is high quality evidence—there is a simple correspondence between evidence and belief in the hypothesis, which strongly suggests that EBM proponents accept the evidentialist thesis that evidence determines justifiable belief in hypotheses.

1.4.3: Evidence, Observation and Higher-Order Evidence

The second question above is about what kinds of things 'evidence' can be. What things are eligible to justify beliefs? In the philosophical literature, evidence has been equated with facts, observations, propositions, sense data and mental states, amongst others (see [161]). EBM proponents seem to view evidence as empirical observations. In the *Users' Guides*, for instance, Guyatt et al. state:

> *"any empirical observation about the apparent relationship between events constitutes potential evidence"* ([82], p.1292).

Similarly, Schunemann et al. state:

> *"all relevant clinical studies and observations provide evidence"* ([37], p.609)

Of course, a clinician need not directly make observations herself for them to count as evidence for her beliefs. Although clinicians' personal observations are often included in EBM accounts, the priority is on using observations made through clinical studies. This has suggested to some critics of EBM that the movement sees evidence as consisting of objective mind-independent facts about the world that can be accessed by all. Upshur and Goldenberg both criticize EBM for this. Goldenberg calls it an *"antiquated understanding of evidence as "facts" about the world"* ([57], p.2622), and argues that it neglects the theory-ladenness of observation and the underdetermination of theories by data. Upshur offers a four-part taxonomy of concepts of evidence, and situates the EBM account in the "Quantitative/General" (aka. "Impersonal Mathematical") quadrant [158]. He states that for EBM:

> *"Quantitative data derived through the application of recognized study designs constitute the basis of evidence."* ([158], p.94).

He contrasts this with qualitative approaches, and with 'personal' approaches in which *"Evidence is defined in terms of measurement of personal belief."* (*ibid.*) Upshur identifies Bayesian accounts of evidence as the clearest representation of a 'Quantitive/Personal' approach to evidence, again divorcing Bayesian and EBM accounts.

It is not clear that EBM proponents need accept this identification of evidence with facts. Several EBM sources emphasize the fallibility of evidence (e.g. [8,10,86,165]). An alternative picture could view the challenge of EBM as the problem of individuals using second-hand observations to justify their beliefs. There is a huge literature of reports of systematic observations. EBM instructs practitioners to base their beliefs about medical treatments primarily on these, but it is hard to know how much trust and weight to put in reports of others' observations. In a sense, the 'observations' for users of EBM are observations of evidence of observations—that is, reading study reports gives good reason to believe that certain observations have been made. The belief that these observations have been made can then be used to support hypotheses. On this view, both experiential states of observation and beliefs themselves can be used to justify beliefs.[13] This fits more closely with the evidentialist picture described above.

A number of philosophers have distinguished between first- and second-order evidence (e.g. [161,181-183]). First-order evidence is evidence about some hypothesis, while second-order evidence is evidence about evidence. Alternatively, one could call this meta-evidence. There are at least two distinct but related roles for meta-evidence in medicine.[14] First, X could be evidence that some evidence E for hypothesis H exists (even if E remains unknown). Observations of study reports provide evidence that there is evidence for the hypothesis. An interesting example of second-order evidence is expert opinion (see [181,182]). If a clinician does not have access to data concerning H, but knows that a reliable, evidence-based expert believes H and claims to have strong evidence for H, then this is evidence that some evidence E, the expert's evidence base for H, exists. The expert's beliefs would not be evidence for H for anyone who had already accessed evidence E (including the expert), but surely does lend some credence to H for those who have not accessed or cannot access E. Importantly, those who already know E are unaffected by the expert's belief because this would be double-counting of E, and *not* necessarily because the expert's opinion cannot be evidential. Similarly, disagreement between experts could be taken as evidence that the evidence base is inconclusive (*cf.* [181,182,184]).

Some authors in the EBM movement have rejected this view of second-order evidence as evidential. Schunemann et al. state of GRADE that:

> *"the system also clarifies that expert opinion is not a category of evidence."* ([37], p.609)

Other hierarchy authors disagree—a number of hierarchies include expert opinion as a category, albeit very low-ranked (e.g. [15,23,31,90,102,185]). The reasons for ranking expert opinion at a low

[13] This may suggest a coherentist picture of evidence—coherence with one's beliefs is evidence for a belief. Djulbegovic, Guyatt & Ashcroft [86] hint at this as one of the models of evidence in EBM.

[14] A third and well-discussed role of meta-evidence which is not discussed in detail here is as evidence that previous beliefs were not rationally-justified. Examples include learning that one is at risk of hypoxia, a condition which affects the ability to reason correctly without being noticeable to the person affected [184].

level are understandable—it is proxy evidence that is exposed to a range of potential biases, including misappraisal by the expert, bias or deliberate misleading by the expert, and double-counting where it is unclear what evidence the expert has considered. However, rejecting expert opinion as a source of evidence altogether is harder to justify. It suggests that the GRADE position takes a more attenuated view of evidence than simple evidentialism, and also seems to conflict with Bayesian epistemology.[15] GRADE must offer a more complex evidential picture, in which only certain observations count towards an agent's evidence-base. Restricting the evidence base to first-order evidence may be the intention, although as argued above, it is not obvious that study-reports should count as first-order evidence.

A second form of meta-evidence is evidence about evidence E—for instance, evidence that E is strong, that E is high-quality, etc. The EBM movement clearly endorse this kind of meta-evidence. They believe, for instance, that the fact that a study was randomized is evidence that the evidence from the study is high-quality. It seems to be these kinds of beliefs which hierarchies of evidence primarily codify.

Note that evidence can be both first- and second-order. A new study could provide evidence that H is true, while also providing evidence that the evidence from another study is unreliable. In particular, studies can act as *defeaters* for the evidence from other studies. In an evidentialist framework, a defeater of E undermines the use of E as a justification for a belief H [161,184]. A Bayesian framework can accommodate this intuition too—a defeater makes the evidence E from a study likely even if H is false, and thereby decreases the effect of E in belief in H. Similarly, a Popperian framework can handle defeaters by seeing them as demonstrating that a test T is not a severe test of H. Defeaters could include evidence that belief in H has not been rationally justified, that evidence was misleading, that H fails to account for E, or that there are convincing alternative explanations for the evidence other than H. The role of defeaters in medicine is discussed in more detail in Chapter 6.

In summary, for EBM proponents, evidence is that which justifies belief. Their approaches most resemble an strong evidentialist conception of evidence, in which two rational agents with the same evidence considering the same hypotheses will be justified in believing the same things. For the most part, though, EBM is consistent (potentially given some modifications) with alternative accounts including subjective Bayesianism. There is no clear consensus about what constitutes evidence (i.e. what it is that justifies beliefs), but the literature suggests observations, loosely interpreted to include reports of observations, though not necessarily including statements of others' beliefs. It is clear,

[15] The conflict with subjective Bayesianism results from the fact that expert opinion clearly does affect some rational agents' degrees of belief in a hypothesis, which entails that expert opinion *is* evidence for those agents.

though, that meta-evidence is accepted where E_1 is evidence about E_2. It is these meta-evidential relationships which hierarchies attempt to codify.

This thesis attempts to demonstrate that hierarchies of evidence should be rejected as a component of EBM appraisal. As the EBM literature suggests but does not necessarily endorse particular accounts of evidence, arguments which address particular accounts of evidence may not be successful in undermining the use of hierarchies. Charitability requires that Djulbegovic, Guyatt & Ashcroft's [86] position that EBM could be recast in terms of different philosophical accounts of evidence is taken seriously. Therefore, wherever possible, this discussion will remain independent of particular accounts of evidence. The only assumptions explicitly made about the EBM theories of evidence are that evidence is that which justifies belief, Worrall's condition that strong evidence for H must support H as well as tell against rival hypotheses, and the claim that meta-evidence is admissible as evidence, which the EBM movement must endorse to use hierarchies at all.

For clarificatory purposes, a Bayesian framework is occasionally used to illustrate concepts such as 'strength of evidence'. These cases should be understood as illustrative only, rather than as committing to a particular account of evidence.

1.5: Overview of the Argument

This thesis will argue that hierarchies of evidence are not philosophically defensible, nor are they needed for Evidence-Based Medicine to pursue its goals. Ultimately, the only tenable defence for the use of hierarchies is a pragmatic, empirical one—that practice is actually improved by using hierarchies to appraise evidence.[16] Hierarchies can only be defended as a kind of decision-aid or heuristic, not on the basis of philosophical, theoretical justification. But there is no convincing evidence for the claim that hierarchical appraisal improves practice, although no detailed empirical studies have yet been undertaken.

A number of philosophers have made similar claims, arguing that hierarchies are not theoretically justified. However, their arguments have largely been met by more complex hierarchical approaches and more nuanced interpretations of hierarchies. In order to demonstrate that hierarchies in general are philosophically unjustified, it is necessary to assemble a complete picture of the range of hierarchies which can be defended and the ways in which they can be interpreted. Chapter 2 presents a new framework for the analysis of hierarchies which fulfils this role. There are a number of dimensions upon which interpretative assumptions must be made in order to extract information about evidence from a hierarchy. Sometimes, the way a hierarchy is expressed or the surrounding literature suggests particular interpretative assumptions. However, often users are left with little guidance in making these assumptions. Moreover, there is such great variation in the hierarchies defended that a hierarchy could be found which is compatible with almost any interpretative assumptions.

Chapter 3 presents the first detailed history of hierarchies of evidence and their interpretation. Four phases in the history of hierarchies are identified, from their inception, to their introduction into EBM and subsequent development. The framework for the analysis of hierarchies is used to show how the common interpretations assumptions have changed over time. A number of philosophical interpretations of hierarchies are also presented and analysed. This will allow us to map the space of defended interpretations of hierarchies, and criticise arguments which target straw men, or only apply to certain interpretations.

Given this framework, it is possible to show precisely which interpretations of hierarchies are susceptible to existing criticism. Chapter 4 argues that most existing criticism targets only particular interpretations, most notably from specific periods in the history of EBM. While these arguments are

[16] An important question here will be 'Improved compared to what?'—it is one thing for hierarchical appraisal of evidence to improve practice compared to unsystematic practice which ignores epidemiological evidence, and quite another for it to improve practice compared to alternative models of systematic evidence appraisal.

very important in restricting the space of potential hierarchies and interpretations which may be defensible, they do not undermine a hierarchical approach to evidence itself. Three main groups of interpretations of hierarchies survive existing criticism relatively unscathed: relative rankings, heuristic interpretations, and sophisticated conditional interpretations such as that defended by the GRADE Working Group.

These interpretations are then criticised in Chapters 5, 6 and 7. Chapter 5 focuses on interpretations which restrict hierarchies to relative rankings, such as those defended by La Caze [11], Greenhalgh [43] and Guyatt et al. [8]. Relative rankings provide so little information as to be uninformative for clinical purposes. If a relative ranking is all that remains, then hierarchies have no important role to play in medicine.

Chapter 6 focuses on more sophisticated approaches including conditional rankings, most notably GRADE. Chapter 6 argues that atomistic approaches to evidence, which evaluate the strength or quality of the evidence provided by each study in isolation, are not philosophically defensible. This challenge can be avoided by moving the appraisal to the level of the entire evidence base, as GRADE and numerous others have done. However, this raises a further problem—the range of information which is considered by a GRADE approach is partial. GRADE focuses on information about the average treatment effect in a population. However, it is demonstrated here that in making recommendations to patients, information about average treatment effects is insufficient, no matter how strongly justified by high-quality evidence. A range of case-studies will show the importance of information about the variation in treatment effects and causes of that variation, also known as 'heterogeneous treatment effects'. GRADE's current attempts to accommodate information about heterogeneity are inadequate. A hierarchical approach to heterogeneous treatment effects is not tenable.

Finally, Chapter 7 considers defences of hierarchies as a heuristic or decision-aid. The success of this move will depend purely on empirical evidence that hierarchies are good heuristics—that they achieve the goals of EBM in practice. A survey of the literature shows that this evidence is currently lacking. Therefore, at the present time, neither a theoretical nor an empirical justification for hierarchies of evidence can be successfully provided. Moreover, there are alternative ways to pursue the goals of Evidence-Based Medicine, and in the light of the criticisms raised here, these should be explored.

CHAPTER TWO:

THE MYTHS OF *THE* HIERARCHY OF EVIDENCE

CONTENTS

"The hierarchy of evidence" is a myth—one propounded by both critics and proponents of Evidence-Based Medicine. The myth envisions a single hierarchy of evidence which is defended by all (or at least most) EBM proponents. This is hailed as "*the hierarchy*". The myth is useful for critics and proponents alike. For critics, it facilitates a clear argumentative strategy: identify a hierarchy, claim that it is *the hierarchy*, and criticize its features and properties (*cf.* [11,55,57,69,121,124,186,187]). If the criticism is successful, then a crucial part of Evidence-Based Medicine is undermined. For example, Bluhm, although recognizing that there is variation in the hierarchies defended within the EBM movement, criticizes hierarchies on the grounds that physiological studies and clinical observations are important sources of evidence, contrary to their low ranking in the hierarchy:

> *"EBM urges, however, that whenever possible, the evidence used in clinical decision making should come from as high in the hierarchy as possible."* ([55], p.537)

However, as demonstrated below, the majority of hierarchies, including the extremely influential GRADE, do not include physiological studies or clinical experience in their ranking at all. So this criticism is at best a partial response to hierarchical approaches. Chapter 4 explores more examples of criticism directed at hierarchies which only applies to certain hierarchies interpreted in certain ways.

For proponents, identifying a hierarchy as *the hierarchy* provides implicit authority (e.g. [8,82,188]).[17] If EBM commits to the hierarchy, and the hierarchy presented in a particular paper is identified as the hierarchy, then EBM commits to the hierarchy presented. That paper thus marshals the apparent support of the whole EBM movement for its hierarchy. More significantly from a philosophical perspective, the author no longer appears to need to justify all of the particular features and properties of their hierarchy—it is the consensus position of all EBM proponents, so the justification is in the literature which leads us this far. For instance, Greenhalgh presented a novel hierarchy with many original features in 1997 under the heading "*The traditional hierarchy of evidence*", calling it:

> "*Standard notation for the relative weight carried by the different types of primary study*", aka. "*the hierarchy of evidence*" ([43], p.54).

Greenhalgh cites Guyatt et al.'s 1995 hierarchy [190] as authority for this, despite the two hierarchies being vastly different.

In reality, hierarchies vary in many philosophically significant ways. Different hierarchies imply, support and depend upon different claims about evidence, methodology and practice. For the most part, the individual hierarchies which EBM critics criticize are defended by no more than a

[17] Even authors who explicitly claim that their hierarchy is distinct from those previously formulated usually refer to their hierarchy as distinct from *the hierarchy* (e.g. [33,189]), so use the myth for rhetorical advantage.

couple of authors. Meanwhile, the hierarchies which authors present as *the EBM hierarchy*, and even attribute to previous sources, are often innovations. I have catalogued over 80 distinct hierarchies (see [191]), many of whose authors refer to their version as "the hierarchy of evidence". The variation is so great that arguably no interesting property is common to every hierarchy.[18]

The myth leads to failure to engage critically with hierarchy design. Proponents are not forced to defend the choices they make when formulating their hierarchies, while critics are not exposed as criticizing only certain variants of hierarchy. Moreover, little attention is paid to which hierarchies do and do not survive certain criticisms. There is currently little accepted terminology to discuss variation in hierarchies, making it difficult for critics to point to features of hierarchies which they find problematic, or for proponents to describe and defend the choices and changes they make. The most basic goal of the framework for philosophical analysis of hierarchies presented in this chapter is provide clarifying terminology for this discussion.

What is of interest to philosophers and medical practitioners alike is the information about evidence which follows from hierarchies. As such, this framework is based around the *interpretation* of hierarchies. Interpretation is the process of deriving information about evidence from hierarchies. Interpreting hierarchies requires a range of *interpretative assumptions*. While authors of hierarchies can attempt to guide the user in making these assumptions, in reality users are usually left to make most of these interpretative decisions without guidance. The assumptions made will affect the claims about evidence which the user derives from a hierarchy, and therefore their clinical practice.

The assumptions made when interpreting hierarchies have serious philosophical and practical ramifications. Practically, they affect which evidence practitioners consider, and the weight they give to different sources. Philosophically, the justifications for a hierarchical approach to appraisal depend sensitively upon the interpretative assumptions. In almost all cases, the 'theory of evidence' which supposedly follows from a hierarchy is in fact *read into* that hierarchy according to the interpretative assumptions. Users do not so much read off a theory of evidence from a hierarchy as read a theory *into* the hierarchy. Attempts to isolate an 'Evidence-Based Medicine theory of evidence' based on hierarchies will be unsuccessful, and risk making a straw man. Hierarchies are compatible with a whole range of ideas about evidence.

In summary, this chapter defends two main claims:

1. There is philosophically and practically significant variation between hierarchies of evidence.

[18] Even the property of ranking RCTs above observational studies is not universal, as Evans' 2003 hierarchy demonstrates—both RCTs and observational studies are ranked as "Good" [192].

2. Even if a specific hierarchy is chosen, the practical and philosophical consequences of that hierarchy depend on the interpretative assumptions made, which are not usually clear from the surrounding literature—multiple interpretations are consistent with the same hierarchy.

This chapter identifies six important dimensions upon which interpretative assumptions must be made in order to derive information about evidence. Examples of hierarchies which suggest different assumptions can be found, but many hierarchies leave these interpretative decisions to users.

Very little analysis of hierarchy design has been published to date. The two notable exceptions are Adam La Caze's 'categorical' vs. 'non-categorical' distinction [72] and Robyn Bluhm's distinction between 'hierarchies of evidence' and 'hierarchies of methodology' [55]. Though neither distinction is adopted outright within this analytic framework, the dimensions they consider will be included in modified forms. Bluhm's distinction is adopted with modifications and extensions in Section 2.1. Sections 2.5-2.6 argue that La Caze's distinction, though suggestive of several important properties, is orthogonal to the most important concerns.

2.1: Subject Matter

The definition of a hierarchy stated in Chapter 1 requires that a hierarchy ascribes some epistemic property to evidence or some epistemic relation to two or more pieces of evidence. The evidence is the *subject matter* of a hierarchy—the thing it provides information about. However, not all hierarchies ascribe an epistemic property or relation to the evidence from an individual study. Call hierarchies which do so "Hierarchies of Evidence". However, other hierarchies ascribe a property not to the evidence from a study itself, but to a *study-design* or to an *evidence base* (i.e. a set of evidence from a set of studies) as a whole.

2.1.1: Hierarchies of Methodologies

Bluhm claims that most hierarchies in EBM are not hierarchies of evidence, but *hierarchies of methodologies*. She states:

> "*the term hierarchy of evidence is a misnomer: the hierarchy is actually a hierarchy of methodologies. That is, it focuses not on the actual results of a particular study or group of studies—in other words, on the evidence they provide for the efficacy of a treatment—but on how that evidence was obtained.*" ([55], p.536)

The distinction Bluhm makes is useful, but her claim that most hierarchies focus on the methodology not the evidence is incorrect. Although what hierarchies actually present is a list of methodologies in some ranking, in most cases the intended interpretation is not that one methodology is superior to the other, but that the evidence provided by tokens of that study-type is superior. A hierarchy's input is information about a study's methodology and the output is information about the evidence provided by the study. For instance, Guyatt et al.'s influential *Users' Guides* hierarchy explicitly states that it is ranking the strength of evidence (see Fig.1, above). The basic approach in a hierarchy of evidence is to find an individual study, determine the methodology used, and look up that methodology in a hierarchy, which will then give some information about the evidence that study provides.

But a few hierarchies do ascribe properties or relations to methodologies themselves [193,194]. For instance, David Sackett's 1981 hierarchy [193] (see Fig.5, below) ranks the strength of methods for clinical research. Here, a hierarchy provides a first step in a process for obtaining information about the evidence a study provides—first one gets information about the relative strength of the method, and from there one makes an inference about the evidence a study of that type will provide. The hierarchy does not offer guidance on the second step.

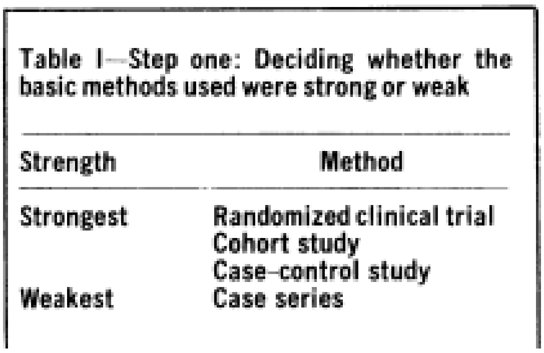


Table I—Step one: Deciding whether the basic methods used were strong or weak

Strength	Method
Strongest	Randomized clinical trial
	Cohort study
	Case–control study
Weakest	Case series

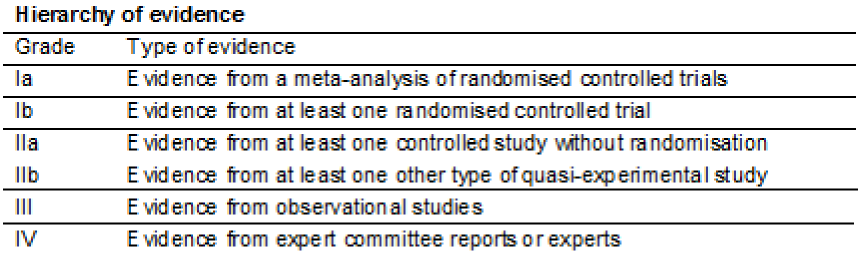
Figure 5: Sackett's [193] originally anonymously-published 1981 hierarchy of methodologies.

This distinction is significant because it affects the information a hierarchy provides. Information about the strength of the underlying design of a study does not necessarily entail information about the strength of the evidence for some hypothesis from the study. This depends on the meaning of 'strength' (see Section 2.2). Presumably, the strength of the methodology is relevant to the strength and quality of the evidence a study provides, but it may not be the only important factor. Bird ([116], p.4) argues that given the importance of effect size in determining the strength and quality of evidence, a hierarchy of methodologies is more philosophically defensible than a hierarchy of evidence. One could envision an evidence appraisal system which involved a hierarchy of methodologies but which did not employ a hierarchy of evidence—the appraisal of the evidence is not made first and foremost on the basis of the methodology, but with an appraisal of the methodology as one component in a multi-factor procedure. Other factors which could be relevant could include how well the methodology was implemented, the size of the effect demonstrated in the study, the clinical importance of the outcomes measured, etc. It is important to note that some hierarchies of evidence do incorporate information about non-methodological features such as effect-size, in the form of *conditions* (see Section 2.5, below). In this framework, these will continue to be considered as hierarchies of evidence because they are used to make an inference about properties of evidence from studies, not studies themselves.

This thesis primarily discusses hierarchies of evidence; the Evidence-Based Medicine movement broadly seek to facilitate the *appraisal of evidence* using a hierarchy (see [8,22,82]), rather than to evaluate methodologies. However, as discussed below, when the links between methodology, study quality and evidence quality are problematic, hierarchies of methodology may escape relatively lightly. Moreover, Bluhm believes that many of the claims which EBM proponents want to make can be supported by hierarchies of methodology alone.

2.1.2: Hierarchies of Evidence Bases

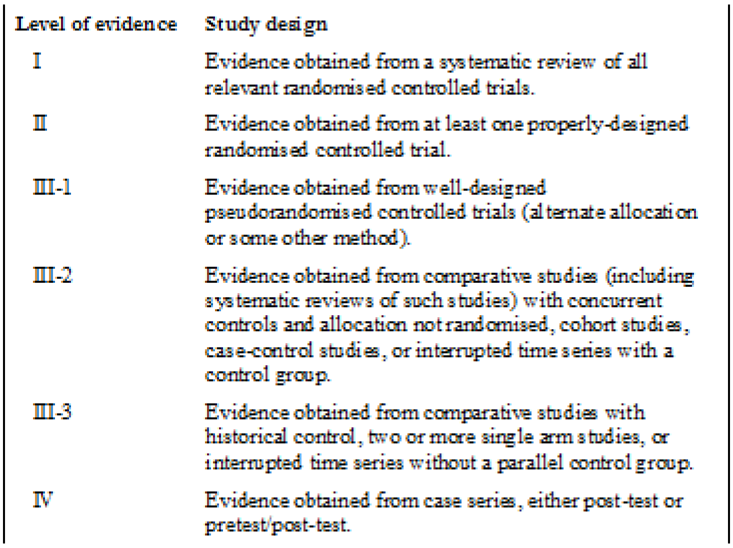
However, some hierarchies of evidence do not ascribe a property or relation to the evidence provided by a single study, but rather to the evidence base for a claim as a whole. Call these “Hierarchies of evidence bases” to distinguish them from hierarchies which apply to singular studies. These hierarchies are particularly popular with guideline-developers and systematic reviewers as they allow the assessment of a number of studies in combination. Examples include NICE’s [28] 2006 hierarchy (see Fig.6) and the Australian National Health and Medical Research Council (NHMRC)’s [6] 1999 hierarchy (see Fig.7). Users first identify the studies which provide information about the clinical claim of interest, and determine what methodologies were used in these studies. They then find the level which best corresponds to the type of evidence which makes up the evidence base in the hierarchy.



Hierarchy of evidence

Grade	Type of evidence
Ia	Evidence from a meta-analysis of randomised controlled trials
Ib	Evidence from at least one randomised controlled trial
IIa	Evidence from at least one controlled study without randomisation
IIb	Evidence from at least one other type of quasi-experimental study
III	Evidence from observational studies
IV	Evidence from expert committee reports or experts

Figure 6: A hierarchy used by NICE in 2006 [28]. The hierarchy allows for the assessment of an evidence-base for a claim—if the evidence base includes *at least one* RCT, it reaches Grade Ib; if it consists only of observational studies, it reaches Grade III, etc.

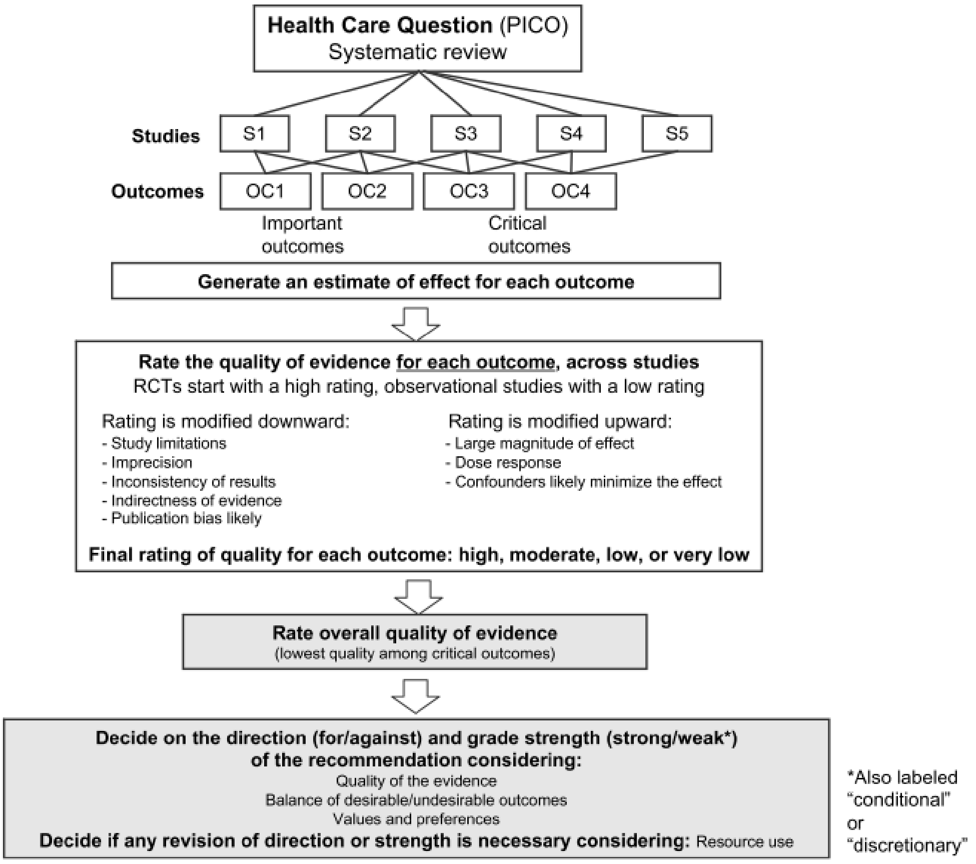


Level of evidence	Study design
I	Evidence obtained from a systematic review of all relevant randomised controlled trials.
II	Evidence obtained from at least one properly-designed randomised controlled trial.
III-1	Evidence obtained from well-designed pseudorandomised controlled trials (alternate allocation or some other method).
III-2	Evidence obtained from comparative studies (including systematic reviews of such studies) with concurrent controls and allocation not randomised, cohort studies, case-control studies, or interrupted time series with a control group.
III-3	Evidence obtained from comparative studies with historical control, two or more single arm studies, or interrupted time series without a parallel control group.
IV	Evidence obtained from case series, either post-test or pretest/post-test.

Figure 7: The Australian NHMRC's [6] 1999 hierarchy also appraises the evidence-base, allowing Level II for *at least one* RCT, with level III-1 for pseudorandomised controlled *trials*, etc.

2.1.3: The GRADE System

The most prominent amongst these approaches is GRADE (see [12-14,195]). GRADE includes a hierarchy of evidence bases, but is a more complex system which involves several stages (see Fig.8, below). The process guides users from formulating a clinical problem all the way to making a recommendation to patients. It consists of six main steps. Users begin with a clinical question of interest [196]—for instance whether a painkiller will be effective as a treatment for migraine for a particular patient. The system can also be used to evaluate evidence for more general recommendations in populations—where this is the case, the GRADE authors are explicit that the recommendation should be stratified to apply to specific populations in specific circumstances [12,48,195] (for more on this point, see Section 6.4). The first step is to identify the relevant outcomes for the patient or population of interest [196]—for migraine, these obviously include pain levels, but will also include measures of possible side-effects. Secondly, users identify studies which provide evidence for an estimate of the treatment's effect, for each outcome of interest.

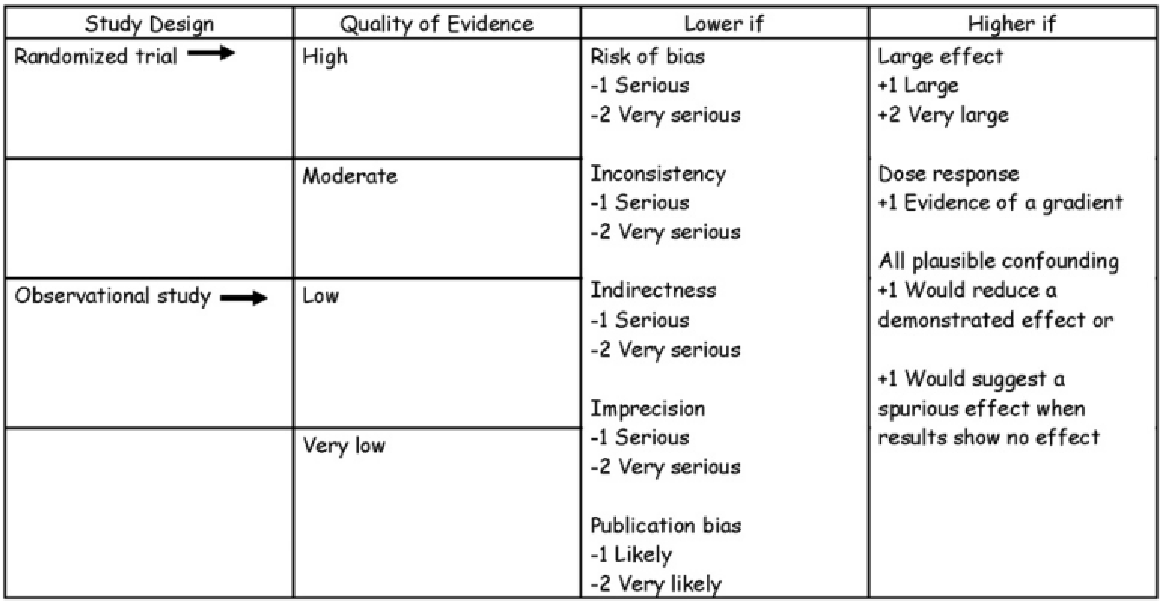

Health Care Question (PICO)
Systematic review
Studies
S1
S2
S3
S4
S5
Outcomes
OC1
OC2
OC3
OC4
Important outcomes
Critical outcomes
Generate an estimate of effect for each outcome
Rate the quality of evidence for each outcome, across studies
RCTs start with a high rating, observational studies with a low rating
Rating is modified downward:
- Study limitations
- Imprecision
- Inconsistency of results
- Indirectness of evidence
- Publication bias likely
Rating is modified upward:
- Large magnitude of effect
- Dose response
- Confounders likely minimize the effect
Final rating of quality for each outcome: high, moderate, low, or very low
Rate overall quality of evidence
(lowest quality among critical outcomes)
Decide on the direction (for/against) and grade strength (strong/weak*) of the recommendation considering:
Quality of the evidence
Balance of desirable/undesirable outcomes
Values and preferences
Decide if any revision of direction or strength is necessary considering: Resource use
*Also labeled "conditional" or "discretionary"

Schematic view of GRADE's process for developing recommendations. *Abbreviation*: RCT, randomized controlled trials.

Figure 8: A summary of the stages in using GRADE from Guyatt et al. ([195], p.385).

Thirdly, users apply the GRADE hierarchy (see Figure 9, below) to assess the evidence base for a hypothesis about each outcome [14]. The hypothesis might be the treatment has a positive, negative or no effect on that outcome, or might specify a precise estimate of the effect. The GRADE hierarchy only includes RCTs and observational studies, and performs its initial ranking on the basis of methodology. If the evidence-base is composed of RCTs, then the initial rating is "high quality". If composed of observational studies, the initial rating is "low quality". GRADE offers four quality levels—high, moderate, low and very low. Unfortunately, GRADE does not offer much information about what to do with mixed evidence-bases: it is not clear whether one RCT amongst a set of observational studies suffices to bring the evidence-base to high-quality initially, or whether a preponderance of RCTs would be necessary, or whether when RCTs are present, the observational studies are disregarded.[19]

[19] For this reason, it seems particularly easy to misunderstand GRADE as presenting a hierarchy of evidence, appraising each RCT or observational study independently. This is not the case, however, as the up- and down-grading criteria demonstrate. They include criteria which clearly apply only to the entire evidence-base such as risk of publication bias and inconsistency amongst the results of studies grade [14,197,198].



Study Design	Quality of Evidence	Lower if	Higher if
Randomized trial →	High	Risk of bias -1 Serious -2 Very serious Inconsistency -1 Serious -2 Very serious Indirectness -1 Serious -2 Very serious Imprecision -1 Serious -2 Very serious Publication bias -1 Likely -2 Very likely	Large effect +1 Large +2 Very large Dose response +1 Evidence of a gradient All plausible confounding +1 Would reduce a demonstrated effect or +1 Would suggest a spurious effect when results show no effect
	Moderate		
Observational study →	Low		
	Very low		

Fig. 2. Quality assessment criteria.

Figure 9: An example of a GRADE hierarchy from Balshem et al. 2011 ([14], p.386). The hierarchy has been presented in many different ways (see e.g. [37,46,199,200]). Here, the initial attribution of quality is indicated by the first column, and potential up- and down-grading criteria are listed in the 4th and 3rd columns respectively.

Once the initial rating is determined, GRADE offers a number of secondary criteria which allow for the quality rating to be up- or down-graded [12,14,129,197,198,201-203] (see Section 2.5 for more on conditionality in hierarchies). Downgrading criteria include risk of bias [203], publication bias [198], and inconsistency amongst the study results [197]. Depending on the severity of these problems, the evidence-base could be downgraded one or two quality levels per problem. The effect of different issues is cumulative, so a high risk of publication bias (-2) plus inconsistency in the results (-1) results in a net downgrading of three levels. So an evidence base of RCTs with high risk of publication bias and inconsistent results would be a “very low quality” evidence base. Upgrading criteria include the studies showing a large treatment effect and evidence of a dose-response gradient[20] [202]. So an evidence-base which consists of observational studies but which shows a very large effect is “high quality”. Up- and down-grading criteria can be offset against one another. So, for instance, if an evidence base of RCTs had a risk of bias, but showed a large effect, the -1 and +1 conditions cancel out, and the evidence base remains “high-quality”.

[20] I.e. the greater the dose, the greater the effect and vice versa.

Fourthly, the user must decide whether to recommend or recommend against using the treatment (or neither), given the balance of effects on the outcomes of interest. As part of this process, users decide which outcomes are and are not of critical importance for the patient or population [204].

Fifthly, the user employs the appraisals of the evidence bases performed in stage 3 to assign an "overall rating" or level of evidence for this recommendation [204]. The GRADE Working Group are explicit that this is equal to the lowest quality-rating for an outcome classed as 'critically important':

> *"the lowest quality of evidence for any of the outcomes that are critical to making a decision should provide the basis for rating overall quality of evidence"* ([12], p.1491).

Finally, users decide how strong the recommendation can be. GRADE limits the options to a strong or weak recommendation [205]. Multiple factors are relevant here, but chief amongst them is the overall quality of evidence as determined in stage 5:

> *"The balance between desirable and undesirable outcomes and the application of patients' values and preferences determine the direction of the recommendation and these factors, along with the quality of the evidence, determine the strength of the recommendation"* ([195], p.386).

Ordinarily, recommendation strength is tied quite closely to quality of evidence. But other relevant factors include the size of the effect and the practical availability of the treatment[21] [12,50,205]. High-quality evidence does not guarantee a strong recommendation where there is a marginal effect size, or the treatment is impractical [48]. In particular, where the treatment has beneficial effects on some outcomes of interest, but detrimental effects on others, weak recommendations are preferred to indicate that patients should be making the decision according to their preferences (see [195,205]).[22]

The GRADE system uses its hierarchy of evidence bases twice over, first appraising the evidence base for hypotheses about treatment effects on each outcome, and then using these

[21] Relevant practicalities discussed include access to the required facilities and the resource use implications (see [14,195]).

[22] Note that Guyatt et al. in 2008 [48] stated that strong recommendations are permissible even when the evidence is low-quality under some circumstances. However, the example they give makes it clear that these are circumstances where the evidence *against* the strong recommendation is low-quality—as they put it: *"Consider the decision to administer aspirin or paracetamol (acetaminophen) to children with chicken pox. Observational studies have observed an association between aspirin administration and Reye's syndrome. (...) the low quality evidence regarding the association between aspirin and Reye's syndrome does not preclude a strong recommendation for paracetamol."* ([48], p.925). Here, the low-quality evidence indicating against paracetamol fails to undermine a strong recommendation, as opposed to justifying it. No positive cases where strong recommendations follow from low-quality evidence of benefit are offered, and this is repeatedly contradicted elsewhere in subsequent publications (e.g. [12,14,195]). Therefore, I will assume that GRADE's position is that strong recommendations require high-quality positive evidence.

appraisals to assess the quality of the evidence for the recommendation overall. Other hierarchies have equated the strength of a recommendation directly with the quality or level of the evidence-base as rated by the hierarchy (see [12,45]), such as the NICE [2,28] and Australian NHMRC [6] hierarchies above.

2.1.4: SUMMARY

Users must determine whether a hierarchy is supposed to appraise the evidence provided by an individual study, the evidence base for a claim as a whole, or a research methodology. In some cases, it can be unclear whether a hierarchy applies to individual studies or to the whole evidence-base. For instance, although the surrounding literature goes some way towards explaining the GRADE process, because the hierarchy does not explain that the terms "RCT" and "Observational study" act as shorthand for "An evidence base consisting primarily of/containing RCTs", and so on, users may be confused and apply GRADE as a simple hierarchy of evidence, evaluating the evidence from individual studies. Indeed, GRADE is often used by guideline developers to evaluate individual studies (*cf.* [4,206]).

There are major philosophical differences between a hierarchy of evidence and a hierarchy of evidence bases. Appraising an evidence base as high-quality does not necessarily entail that any individual study provides high-quality evidence, let alone that all do. These hierarchies also allow for some features of the relationship between different studies to be included in an appraisal. For instance, GRADE allows users to include considerations like inconsistency of results between studies as a factor which downgrades the quality of an evidence base. By contrast, a hierarchy of evidence would only appraise the two contradictory studies in isolation, which could leave practitioners confused when faced with two studies rated as strong or high-quality, which disagree (see Chapter 6 for further discussion). Many of the criticisms aimed at hierarchies of evidence may not apply to hierarchies of evidence bases (see Chapter 4).

2.2: Interpretation Property

A hierarchy ascribes a property or relation, or some degree of that property or relation, to evidence or an evidence-base. This section presents the different properties which hierarchies ascribe to evidence, which will be called "interpretation properties". These include the quality and strength of the evidence or evidence base, the strength of a recommendation which can be justified on that basis, and the validity of evidence. Each concept is defined, and the philosophical consequences of using each as the intended interpretation property are explored. Where hierarchies are used to rank the evidence from different studies against one another, interpretation relations will be used, such as "X provides higher quality evidence than Y" or "X is a stronger evidence-base than Y" (see Section 2.3). Some hierarchies specify the intended interpretation property, while many leave this to the reader.

2.2.1: Evidence Quality

A very common interpretation property in the EBM literature is 'quality of evidence'. Examples of hierarchies interpreting using 'evidence quality' include GRADE [12,14,195] (see Figure 9, above) amongst others (e.g. [37,207-210]). A straightforward example comes from Schunemann et al.'s ATS hierarchy ([37], p.609):

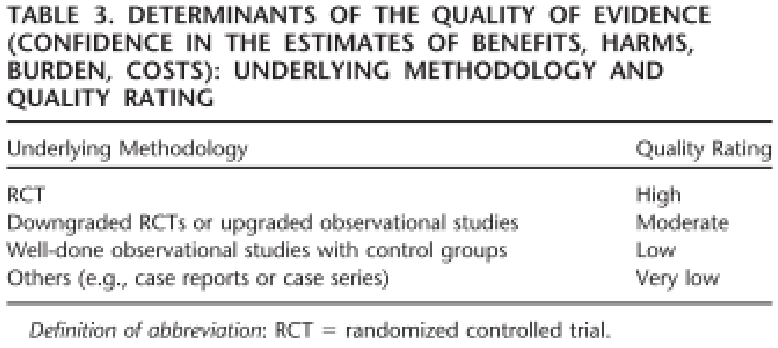


TABLE 3. DETERMINANTS OF THE QUALITY OF EVIDENCE (CONFIDENCE IN THE ESTIMATES OF BENEFITS, HARMS, BURDEN, COSTS): UNDERLYING METHODOLOGY AND QUALITY RATING

Underlying Methodology	Quality Rating
RCT	High
Downgraded RCTs or upgraded observational studies	Moderate
Well-done observational studies with control groups	Low
Others (e.g., case reports or case series)	Very low

Definition of abbreviation: RCT = randomized controlled trial.

Figure 10: Schunemann et al.'s ATS hierarchy ([37], p.609) presents a simplified version of the GRADE ranking, interpreted in terms of quality of evidence.

It can be very difficult to tell what precisely this term means—not least because 'quality' here is a relation between evidence and a purpose or role, which holds to some degree. The evidence from a study could be high-quality as evidence for a claim about the average treatment effect in the study-population, but low-quality as evidence for a broader generalisation, or for a prediction about a specific patient. To determine the quality of evidence one must know (1) for what purpose the evidence is being used, and (2) what makes some evidence good at fulfilling that role.

As outlined in Chapter 1, the focus of EBM is on using evidence to improve clinical outcomes for individual patients in practice. So high-quality evidence is evidence which will allow clinicians to improve patient-care. Evidence does not improve patient-care directly, but equips clinicians with the information needed to make recommendations to patients which would lead to the best (or at least *better*) outcomes for them.

EBM proponents offer some ideas of what the properties of high-quality evidence are. Many refer to the elimination of biases. For instance, the Australian NHMRC refer to:

> *"the degree to which bias has been eliminated by the study design"* ([6], p.2).

The *Users' Guides* provides the following breakdown of quality of evidence:

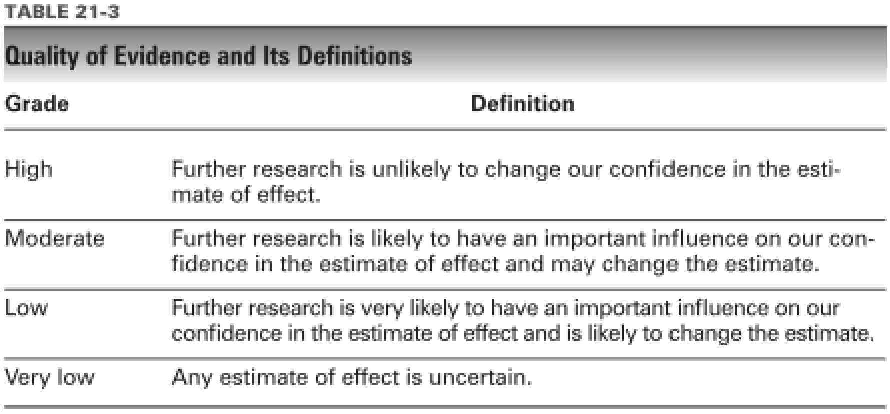


TABLE 21-3

Quality of Evidence and Its Definitions

Grade	Definition
High	Further research is unlikely to change our confidence in the estimate of effect.
Moderate	Further research is likely to have an important influence on our confidence in the estimate of effect and may change the estimate.
Low	Further research is very likely to have an important influence on our confidence in the estimate of effect and is likely to change the estimate.
Very low	Any estimate of effect is uncertain.

Figure 11: A table summarising criteria for each of the four quality levels used by GRADE (see [12,14]), published as part of the *Users' Guides* book ([8], p.610). The definition is in terms of confidence that estimates of the effect will change given further research.

By contrast, GRADE articles are concerned with high-quality *evidence bases*, in which there is evidence which clearly addresses the relevant outcomes. Schunemann et al. state:

> *"High-quality evidence should provide precise estimates of both benefits and downsides, and the balance should be clear"* ([37], p.606).

The GRADE Working Group state:

> *"the quality of evidence indicates the extent to which we can be confident that an estimate of effect is correct"* ([12], p.1490)

This definition was echoed by Balshem et al. ([14], p.403). Quality, then, for EBM proponents seems to be a matter of confidence in the elimination of biases, accurate estimation of an effect, and unlikeliness of beliefs about the effect changing given further research. EBM proponents are also explicit that high-quality evidence must be applicable to practice (see e.g. [9,22,211]).

The desiderata for high-quality evidence, then, include: (a) that the data gathered in the study and the correlations reported as results of correctly applying appropriate statistical tests to the data are accurate; (b) that a causal interpretation of these results is warranted; (c) that there is warrant to generalise the causal hypothesis to a population which includes the patient(s); and, (d) this generalisation of the causal hypothesis confirmed by the evidence is relevant to the process of making recommendations to the patient(s).

The definition of 'quality of evidence' will involve all four of these desiderata. The more confident a rational clinician can be in each of (a)-(d) with respect to some causal hypothesis H, the higher quality the evidence is for H. Desiderata (a) and (b) have both been referred to as 'internal validity' within the clinical epidemiology literature (*cf.* [93,94,212]). (c) is generally termed 'external validity' (see e.g. [213,214]), and (d) 'clinical significance', 'applicability' or 'clinical relevance' (see [22,215-217]).[23] Each property is discussed below (see Sections 2.2.2 and 2.2.3).

High-quality evidence fulfils each of these desiderata to a high degree. Lower quality evidence may have limitations in any one respect, or multiple limitations. Different hierarchies and authors that use evidence quality as the interpretation property suggest different accounts of the relationship between these properties and an overall measure of evidence quality, putting more emphasis on different parts of the definition. In some cases, such as in Guyatt et al.'s 2008 *Users' Guides* hierarchy [83], there is clearly a great deal of emphasis placed upon studies which generalise to the particular patient—they place *n*-of-1 randomised trials at the highest level, ranking them above even systematic reviews of RCTs, because *n*-of-1 trials relate directly to the patient in question and therefore satisfy criterion (c) [218]. Similarly, GRADE's up- and down-grading criteria relate to different desiderata: 'risk of bias' relates to internal validity, while 'indirectness' relates to external validity and clinical significance, for instance (see [129,203]).

[23] In the Bogen & Woodward [164] framework described in Chapter 1, these are rendered as: (a) the data confirm the phenomena, (b) the phenomena confirm the theory, (c) the theory is generalisable to the target population, and (d) the theory is clinically useful.

2.2.2: Internal and External Validity

'Validity' is a technical term within clinical epidemiology; its meaning is very different from *logical validity* [94,219][24]. There are two commonplace ways of understanding this term in the literature, which are often conflated. Some authors define internal validity as the 'accuracy of the results', while other define it as 'establishing a causal relationship in the population studies'. La Caze's definition takes the 'accuracy' approach:

> *""Internal validity" is the degree to which the results of a study are accurate for the sample of patients included in the study"* ([11], p.510).

By contrast, Clarke et al. define internal validity in terms of establishing a causal relationship, i.e. that the treatment is effective, in the population:

> *"internal validity—namely if all goes well, we can establish with confidence that a treatment is effective in the population under examination."* ([63], §2.4.1)

This latter definition is problematic—for one, it overlooks evidence against the claim that a treatment is effective, which can surely also have internal validity. Therefore, the following definition may be more appropriate: 'X is valid evidence for/against causal hypothesis H to the extent that one can be confident that X confirms/disconfirms H.' (see also [212,222]). The former definition corresponds with desideratum (a) for high-quality evidence above, while the latter fits with desideratum (b).

Further confusion enters the literature where a third definition is used—internal validity is *absence of bias*. For instance, the Cochrane Handbook defines internal validity as:

> *"whether it answers its research question 'correctly', that is, in a manner free from bias"*
>
> ([1], p.188)

The *Users' Guides* states:

> "*studies that have higher internal validity have a lower likelihood of bias/systematic error*"
>
> ([8], p.788)

[24] The clinical epidemiological concept of validity is sometimes called "experimental validity" [220] to differentiate the concept from logical validity, and statistical validity concepts such as construct and criterion validity [221,222]. The question of *statistical validity* is within the scope of experimental validity—a conclusion is statistically valid if it is supported only by appropriate, 'reasonable' statistical inferences, avoiding statistical fallacies [220]. The distinctions originate in works by Campbell, Cook and Stanley [155,156,220].

However, biases can affect both accuracy of the results and warrant for a causal interpretation of the findings. Bias could affect whether the difference in outcomes between the groups was reported correctly (i.e. the result matches the actual state of affairs with respect to outcomes), and whether, and to what extent, a difference in outcome is attributable to the treatment. The difference in definitions is really a case of whether one conceptualises internal validity as a question of whether the *correlation* reported is a real correlation, or whether the correlation reported is a *causal* relationship (and in particular, the correct causal relationship, i.e. that the treatment caused the difference in outcome). The WHO definition of internal validity captures this bifurcation neatly—internal validity is:

> "*the extent to which one can be confident that an estimate of the* effect *or* association *is correct.*"
>
> ([4], p.37, emphasis added)

'Bias' refers to any factor which skews the reported result (whether reported as an association or an effect) away from the true state of affairs [1]. In comparative studies, biases include any factor which is unbalanced between the comparison groups which has an effect on the outcomes measured—these imbalances are also referred to as "confounding factors" or "confounders". For instance, suppose a comparative trial of two arthritis treatments, A and B, is performed, measuring the treatments' effects on pain and joint swelling. Potential confounders for this trial include any factor which would affect pain and joint swelling and which could be unbalanced between the groups taking A and B. These include features of the patients—age, gender, race, genetic factors, etc.—and their condition—severity, subtype of the disease, etiology, duration, etc. If these factors can affect pain and swelling, then if they are unbalanced between the groups, differences in the outcomes between the A-group and the B-group may be due (in part or entirely) to these differences. Therefore, the reported correlation may be spurious, and the causal interpretation unwarranted. These biases result from unequal distribution of patient-features and disease-features between the groups, and are therefore sometimes called "allocation biases" [6,223-225]. Where these allocation biases result from the researchers selectively allocating patients to one group or another using (consciously or unconsciously) these features, it is termed "selection bias" [1,43,75,226,227] (see Chapter 4 for further discussion).[25] There can also be selection bias in systematic reviews, where a biased selection of studies is included [43].

But other factors include the way patients are treated—how attentive the clinicians and nurses are, whether the clinicians vary the dosage at all, whether compliance with the treatment is equally assured, etc. Differences in the way patients are treated other than what treatment they are given are called "treatment biases" [187,228] or "performance biases" [43]. Outcomes may also be affected by

[25] Note that another sense of "selection bias" refers to a biased sample being taken from a population [225]. This is not the meaning intended here. Biased sampling will be referred to as "inclusion bias", following the terminology used by Greenhalgh ([43], p.44).

what the patients expect to happen, and what they think the researchers want to happen. Patients who expect a treatment will be very effective tend to experience greater effects than patients who do not think the treatment will be effective [8,36,43]. This may be due in part to greater compliance with the treatment regimen amongst motivated patients. But it may also be due to enhanced placebo effects—placebo effects are strongly associated with expectations of effect [229,230]. Patients may also experience greater effects where they believe the researchers want or expect to find greater effects. The Hawthorne effect is a well-studied phenomenon: when people know they are being observed, and know or believe they know what the observers want to see, they tend to conform to that expectation [231-234]. The Hawthorne effect will be particularly potent where the outcome being measured is within the subjects' control—for instance, self-reported outcomes (e.g. quality of life questionnaires). So, if patients in different groups have different expectations, or think that the observers want to see different things, this may affect their outcomes, and thereby confound the study. These biases are sometimes called "expectation bias" [8,36,43][26] and "observer-expectancy bias" [36] respectively.[27]

There are other biases which affect other kinds of study. Publication bias affects systematic reviews and meta-analyses. Publication bias is the well-confirmed tendency for studies with positive results to be published more often, faster, and in higher-reputation journals than studies with negative results [238-243]. If undetected, publication bias will positively bias the results of a review [1,243]. Relatedly, positive studies are sometimes re-published, often under different names with different authors in different journals [244,245]. If systematic reviews cannot identify these, again reviews can be positively biased [245].

There are many other potential biases, and a full account is beyond the scope of this thesis (see e.g. [237,246-248]). There are many ways to attempt to control for these biases. Some authors claim that random allocation controls for allocation biases (but see Chapter 4). Checking for baseline imbalances in potential confounding factors and reallocating patients where imbalances are found may also solve allocation biases [1,52,75]. It is hoped that blinding patients to their treatment allocation will prevent expectation biases, while blinding researchers and clinicians to treatment allocation will prevent observer-expectancy bias and treatment bias [22,151,249]. There are tests for publication bias such as funnel plot tests which can attempt to identify publication bias amongst sets of studies [198,250,251].

[26] Expectation biases are also sometimes referred to as the Pygmalion effect (when positive expectations lead to better outcomes) [235].

[27] Observer-expectancy bias is a species of experimenter bias—the phenomenon that experimenters are more likely to find the results they expect to find, presumably due to conscious or unconscious manipulations [236,237]. Observer-expectancy bias is one way in which experimenters might produce the desired effect. Others include selective allocation (i.e. selection bias).

Crucially, it is not possible to know for sure whether bias has occurred in a study (*cf.* [1], p.188). This is the well-known calibration problem [252,253]: unless the true effect or association is already known, one cannot tell whether the reported effect is accurate. Therefore, if validity is defined as absence of bias, it is not possible to know the validity of a study. For validity to be a useful construct, it must be defined instead as *justifiable confidence in* the absence of (serious) bias in study-results. That is, validity is a measure not of bias but of *risk of bias*. If a study is exposed to potential biases (e.g. if the comparison groups were imbalanced on some factor which is a potential confounder), then the validity of the study is decreased *even if* that confounding factor did not actually produce bias (*cf.* [11,222]). This can be easily illustrated: suppose there is a very small, poorly-controlled, non-randomized, non-blinded study. Most authors would agree that this study has low validity. But it could be the case that the study-result happens to give a very accurate estimate of the actual treatment effect despite the methodological blunders. Even if this were the case, the study has low validity; validity is a question of reason to believe in the accuracy of the study-results, not simply a measure of accuracy.

The terminology in the literature moves between validity as a measure of confidence in accuracy of results, and as a measure of confidence in a causal interpretation of those results. The definition of validity in terms of risk of bias confuses the two. For the purposes of this thesis, a composite definition is offered similar to that used by the WHO [4]:

> The evidence for hypothesis H from a study S is valid to the extent that confidence is justified that the data collected in S confirm the results of S, and that the results of S confirm/disconfirm H. *(where H is a causal interpretation of the results of S)*

"External validity" refers to a generalisation of the effect found in the population studied to other populations or other settings [213,214,222]. External validity can be defined as justifiable confidence that the evidence confirms a hypothesis *H'* which is a generalisation of H to a broader target population, a different population, or a particularisation of H to apply to specific individuals (who may or may not have been members of the original study-population). Desideratum (c) above is a consideration of external validity in a population which includes the patient(s) of interest.

There are many challenges to generalising results from one population to another. Internal and external validity are sometimes presented as being in tension or conflicting (see e.g. [213,254,255])—studies which have high internal validity have low external validity, and vice versa. It is not clear that this has to be the case. In particular, internal validity seems to be a prerequisite for external validity (*cf.* [10])—if the result is not accurate and does not confirm or disconfirm the causal hypothesis in the study-population, it is hard to see how a generalisation of that hypothesis to another population can be warranted.

It is sometimes argued that RCTs have low external validity, often due to the fact that RCTs are conducted under carefully-controlled and monitored conditions, with expert clinicians, and that trial populations are usually very precisely specified [55,56,178,256-259]. A study by Sokka & Pincus [260] showed that the vast majority of arthritis patients do not meet the inclusion criteria for most RCTs of arthritis medications. Inclusion criteria are often very stringent, particularly excluding older patients, children, patients with unusual presentations of a disease, patients with co-morbidities (i.e. conditions other than the one being treated which may affect how treatments work or how the disease progresses) and pregnant women (see [192,213,261]). Treatments can be more or less effective in patients with different features and when administered under different conditions. Clinicians with less expertise in the use of certain treatments may deliver less effective interventions. Therefore, extrapolating from a study in one population under one set of conditions to different populations and conditions can be difficult. For the purposes of the definition of evidence quality, the similarity of the study population and conditions to the target population and conditions is very important. Pragmatic trials, conducted under normal clinical conditions in broad target populations (see [153,255,262-264]) and *n*-of-1 trials conducted on the patient in question [218] have both been proposed as methods which address problems of external validity.

Internal validity is the intended interpretation property for at least two hierarchies—the US Preventive Services Task Force's (USPSTF) 2008 hierarchy [5] and Eaves' 2011 hierarchy [132].[28] La Caze argues that the most defensible EBM position uses internal validity as the interpretation property ([11,72], and see Chapter 3). As yet, no hierarchy uses external validity as its intended interpretation property, despite some proponents foregrounding external validity in their accounts of EBM (e.g. [22,78,215,266]). Some hierarchies use *applicability* as the interpretation property, which might be used to mean external validity (e.g. [5,82]). Internal and external validity are of most importance as criteria for high-quality evidence, and thereby as part of a justification for a hierarchy interpreted in terms of evidence quality.

[28] Another example is Jonas' [265] evidence hierarchy, which ranks methods from the "most causal" to "least causal", by which he presumably means the providing the most warrant for a causal inference, to the least.

2.2.3: Applicability, Relevance & Usefulness

The applicability, relevance, clinical significance or usefulness of evidence in the EBM context is the extent to which the evidence confirms or disconfirms hypotheses which are relevant to treating patients, i.e. to the process of making a recommendation to the patient. There will be a range of claims relevant to that process, including predictive claims about the likely effects and side-effects of each potential treatment for that patient. Few studies other than *n*-of-1 trials directly confirm these predictions. It is more likely that clinicians will make predictions on the basis of information about the average effect of some treatment on some outcome in some population, information about the variation in treatments' effects, and information about the causes and predictors of different responses to the treatments. Chapter 6 discusses the challenge of applying population-level findings to individuals in more detail.

A sufficiently high level of internal and external validity is a pre-requisite for clinical significance. Evidence is not clinically useful if the clinician cannot be fairly confident that it is accurate and applies to the relevant population. But there are other considerations relevant to clinical significance. These include whether the outcomes measured are the ones that actually interest clinicians and patients. Some studies measure 'surrogate' or 'proxy' outcomes—outcomes which are supposed to provide a measure of the actual outcome of interest, without actually measuring that outcome directly. These may be used because the actual outcome of interest is difficult, expensive or invasive to measure, or because there is no reliable way of measuring it. Examples of surrogate measures include measuring laboratory values such as C-reactive Protein (CRP) levels as a proxy measure for swelling in arthritis [267]. Where the surrogate outcome does not clearly track the outcome of interest, the clinical significance of evidence is affected. Evidence that a surrogate is a reliable indicator of patient-oriented benefits is crucial if evidence relating to surrogate outcomes is to be useable. In the EBM literature, outcome measures which directly relate to those outcomes which interest patients and clinicians—life expectancy, disease burden, quality of life, mobility, etc.—are called "patient-oriented outcomes", and evidence from studies which measure patient-oriented outcomes is sometimes called POEMs (Patient-Oriented Evidence that Matters) [43,104,127,128,268].

2.2.4: Strength of Evidence

Strength of evidence is sometimes used synonymously with evidence quality—in this case, strong evidence is evidence without weaknesses which threaten validity. However, there is another distinct sense of strength—how strongly the evidence affects or can affect a rational agent's belief in a hypothesis. The more of an effect the evidence has on belief in the hypothesis, the stronger that evidence is. Evidence can be strongly positive or strongly negative.

Evidence can be strong without being high-quality. An example often seen in the philosophical literature is the ECMO case. The early ECMO trials were very small and were stopped early [269-271]. Yet it is argued that the evidence is so clear-cut that ECMO is much more effective than conventional therapies as a treatment for pulmonary hypertension in newborns that the evidence was extremely strong—and indeed that further trials were unnecessary (see [52,69,187,272,273]). Similarly, evidence can be high-quality but weak, as in cases where the trials are well-conducted but the treatment effect discovered is marginal or the confidence-interval is wide.

The Bayesian framework described in Section 1.4 facilitates a simple quantification of the strength of evidence. In that view, the strength of evidence E for hypothesis H is simply the change in degree-of-belief in H when conditionalising on E, i.e.

$$S(E,H) = P(H|E\&B) - P(H|B) \qquad \text{where B denotes background knowledge}$$

In other words, the strength of the evidence E for H (S(E,H)) is the difference between the degree of belief in H given only background knowledge B, and the degree of belief in H given both B and the evidence E.

However, this account of strength of evidence is problematic for proponents of hierarchies. Strength of evidence as a measure of change in degree of belief is dependent upon the prior belief in the hypothesis, P(H|B). If a hypothesis is already believed to a high degree, then positive evidence has a lower potential confirmatory strength than had the hypothesis initially been believed to a low degree. Conversely, negative evidence can be stronger where the hypothesis is well-believed, but relatively weak when the prior probability of the hypothesis is already low. In Bayesian terms, the bounds of S(E,H) are set by:

$$(P(H|B) - 1) \leq S(E,H) \leq (1 - P(H|B))$$

That is, evidence against H cannot be stronger than reducing P(H) to 0, so cannot be stronger than P(H|B)–1; evidence for H cannot be stronger than establishing that P(H)=1, so cannot be stronger than 1–P(H|B) (see also [170,274]).

Hierarchies of evidence strength using this definition are not tenable. The strength of evidence depends on when it was obtained (i.e. what is already part of background knowledge B) as well as the methodology of the study. To illustrate, consider two studies X and Y of identical methodologies both of which provide evidence for H. Suppose that prior belief in H is 0.1 (P(H|B)=0.1). Suppose that conditionalising on either study alone would lead to a posterior belief in H of 0.8 given Bayesian updating, i.e. P(H|X&B)=0.8 and P(H|Y&B)=0.8. If X is considered first, then the strength of X for H is 0.7. But the strength of Y is *at most* 0.2, because H can no longer be confirmed by more than 0.2. But any hierarchy would ascribe the same strength to both X and Y.

The concept the hierarchy-makers who use strength as an interpretation property are trying to invoke is more like the *potential strength* of the evidence. Is the evidence compelling enough that it *could* strongly effect on a rational agent's degree of belief in H? Even if belief in H is high, E might still have high potential strength—if, counterfactually, belief in H was low, E would have increased belief in H substantially. So call E's *potential strength* the degree to which it can affect a perfectly-rational agent's degree of belief in H. Eells & Fitelson defend a similar view, identifying the 'degree of evidential support' of E for H with the effect E has on P(H) in the absence of any other evidence [275,276]. Chapter 6 discusses the problems with taking such an approach in more detail, in particular the problem that evidence from several studies taken together can be stronger than the sum of the evidence from the studies in isolation.

'Strength of evidence' is a common interpretation property explicitly stated by authors of hierarchies (e.g. [8,23,105]). However, it is rarely clear whether this is being used as a synonym for 'quality of evidence', or in this latter sense. Examples include Guyatt et al.'s [8] *Users' Guides* hierarchy (see Fig. 1, above), and Edwards, Russell & Stott's [105] hierarchy (see Figure 12, below).

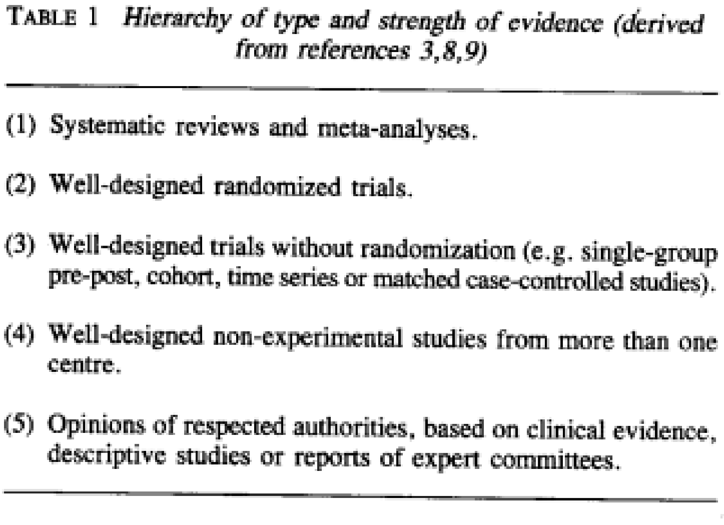


TABLE 1 *Hierarchy of type and strength of evidence (derived from references 3,8,9)*

(1) Systematic reviews and meta-analyses.

(2) Well-designed randomized trials.

(3) Well-designed trials without randomization (e.g. single-group pre-post, cohort, time series or matched case-controlled studies).

(4) Well-designed non-experimental studies from more than one centre.

(5) Opinions of respected authorities, based on clinical evidence, descriptive studies or reports of expert committees.

Figure 12: A hierarchy explicitly using 'strength of evidence' as the intended interpretation property by Edwards, Russell & Stott [105].

A final interpretative assumption is needed in interpreting a hierarchy of strength—whether the hierarchy applies to *positive* or *negative* potential strength, or both. It is plausible that methods which provide strong positive evidence are not able to provide strong negative evidence, or vice versa. For example, EBM proponents often accept that cohort studies provide strong negative evidence: the biases which affect cohort studies usually bias in favour of effectiveness, so a cohort study which apparently shows a treatment to be ineffective is probably strong evidence against the effectiveness of the treatment (e.g. [14,46,202]). But, according to most hierarchies, cohort studies provide considerably weaker evidence *for* the hypothesis that a treatment is effective. Similarly, Howick and colleagues [58,59,137] have argued that mechanistic reasoning provides weak positive evidence (providing a mechanism is no good reason to believe in the effectiveness of a treatment), yet the inability to provide any plausible mechanism could be strong negative evidence (biological implausibility is closely connected with the inability to provide a biological mechanism, and plausibility is a crucial part of Bradford Hill's 'criterion' for causation [137,277]—so mechanistic implausibility is strong evidence against a causal claim). Chapter 6 discusses these issues in more detail.

2.2.5: Level of Evidence

The term 'level of evidence' is particularly ambiguous. It could refer simply to the *level in the hierarchy*. In this sense, the level of evidence is not a useful interpretation property, providing no new information about evidence. But sometimes 'level of evidence' is used to refer to how strongly a hypothesis can justifiably be believed on the basis of some evidence (see e.g. [15,29,103,216,278-280]).

These interpretations are most common amongst hierarchies used by systematic reviewers and guideline developers. They form part of an attempt to standardise how strongly hypotheses are believed. Usually, hierarchies stated in terms of level of evidence are accompanied by a translation to a claim about the *strength of recommendation* (see Section 2.2.6, below). Hierarchies with intended interpretations in terms of level of evidence include SIGN [2,133,134] (see Figure 4, above), and Howick et al.'s '*CEBM Levels of Evidence*' [15,103] (see Figure 14, below). A particularly interesting hybrid design was offered by Cook et al. [151] in 1992, in which levels of evidence are ascribed to both sets of studies and individual studies (in the absence of a systematic review) in the same hierarchy (see Figure 13, below).

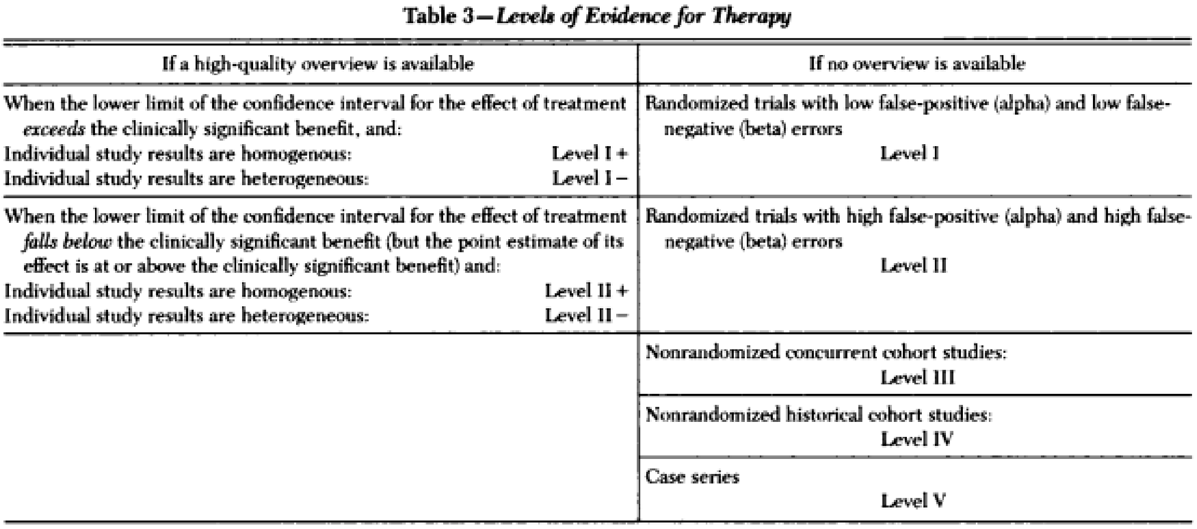


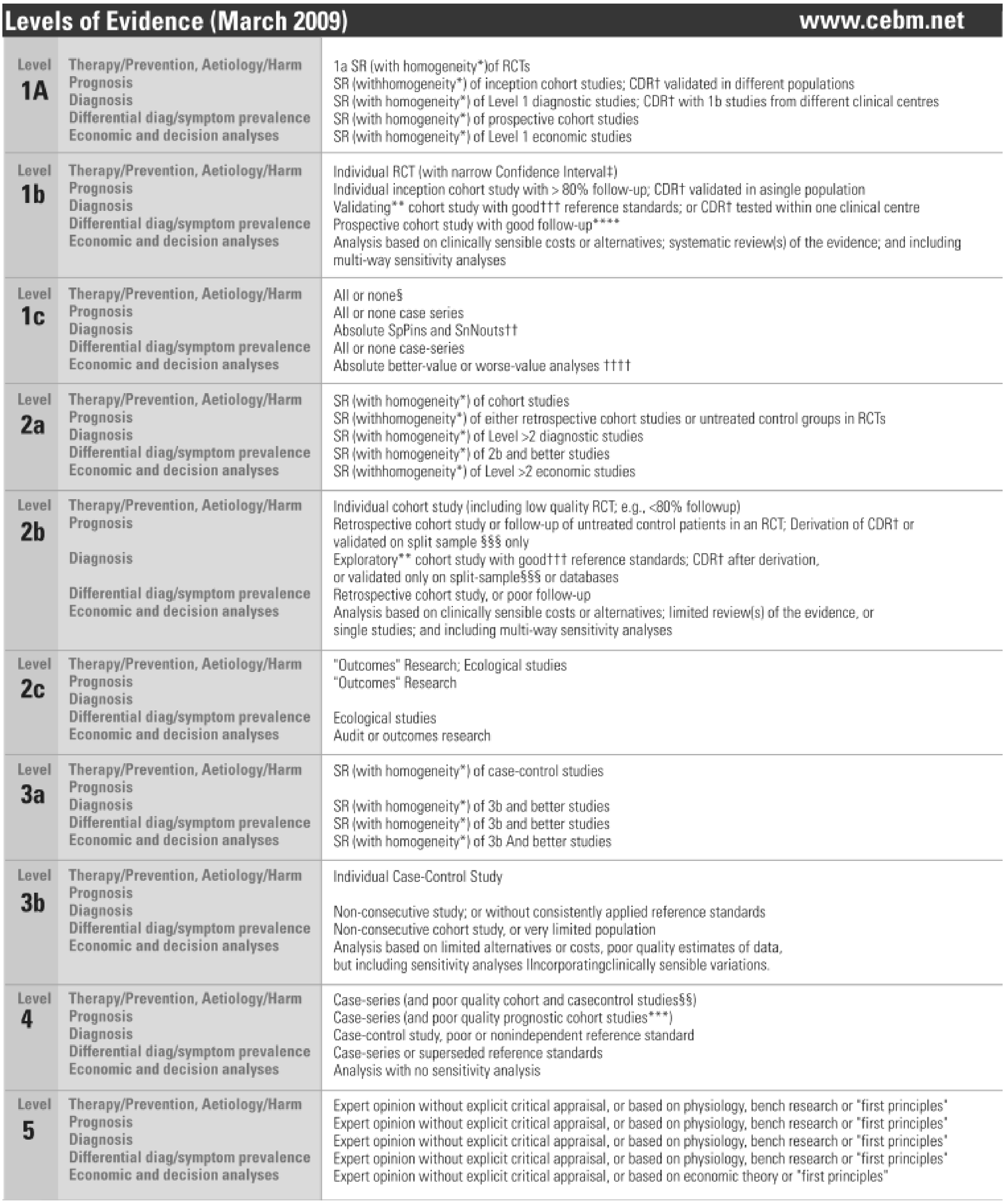
Table 3—*Levels of Evidence for Therapy*

If a high-quality overview is available	If no overview is available
When the lower limit of the confidence interval for the effect of treatment *exceeds* the clinically significant benefit, and: Individual study results are homogenous: Level I + Individual study results are heterogeneous: Level I −	Randomized trials with low false-positive (alpha) and low false-negative (beta) errors Level I
When the lower limit of the confidence interval for the effect of treatment *falls below* the clinically significant benefit (but the point estimate of its effect is at or above the clinically significant benefit) and: Individual study results are homogenous: Level II + Individual study results are heterogeneous: Level II −	Randomized trials with high false-positive (alpha) and high false-negative (beta) errors Level II
	Nonrandomized concurrent cohort studies: Level III
	Nonrandomized historical cohort studies: Level IV
	Case series Level V

Figure 13: Cook et al.'s 1992 *CHEST* hierarchy [151] gives levels of evidence for a therapy (i.e. for the claim that a therapy is effective), appraising the evidence from a set of studies (left) separately from the evidence from individual studies (right).

Unless the interpretation is in terms of a relation (e.g. ‘If there is evidence from RCTs in the evidence-base then a higher degree of belief in the hypothesis is justified than if there are only observational studies in the evidence-base’), interpreting hierarchies in terms of level of evidence seems inconsistent with subjective Bayesianism. Subjective Bayesianism allows for different rational agents to have different degrees of belief in a hypothesis given the same evidence due to different prior probabilities assigned to the hypothesis. As an (absolute) level of evidence interpretation ascribes a particular degree of belief as rationally-justified, it suggests an evidentialist picture, in which any rational agent with the same access to evidence believes hypotheses to the same degree.

In addition to this philosophical consequence, a ranking in terms of ‘level of evidence’ differs from rankings of strength or quality in several important ways. First, ascribing a level of belief justified by the evidence does not necessarily imply any judgments about the quality of the evidence. It is not necessarily the case that evidence must be high-quality in order to entail a high degree of belief in the hypothesis. In particular, where evidence is strong but not high-quality, one would expect a high level of evidence. Similarly, where evidence is weak but high-quality, one might expect a relatively low level of evidence. Secondly, where ‘level of evidence’ is used as the interpretation property for a hierarchy of *evidence bases*, the interpretation does not imply that any individual study within the evidence base provided strong or high-quality evidence. It could be the case that a number of relatively weak or low-quality studies *in combination* justify a high level of evidence for the hypothesis. As Chapter 4 shows, many criticisms of hierarchies will not apply to ‘level of evidence’ interpretations for this reason.



Levels of Evidence (March 2009) www.cebm.net

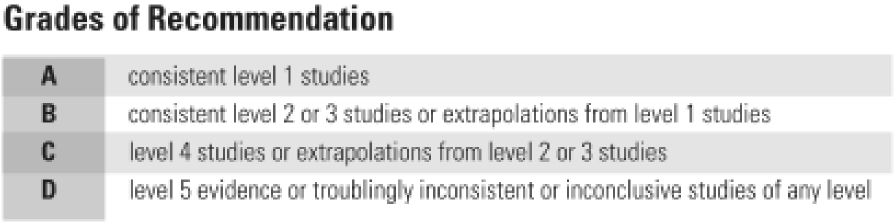
Level 1A	Therapy/Prevention, Aetiology/Harm	1a SR (with homogeneity*)of RCTs
	Prognosis	SR (withhomogeneity*) of inception cohort studies; CDR† validated in different populations
	Diagnosis	SR (with homogeneity*) of Level 1 diagnostic studies; CDR† with 1b studies from different clinical centres
	Differential diag/symptom prevalence	SR (with homogeneity*) of prospective cohort studies
	Economic and decision analyses	SR (with homogeneity*) of Level 1 economic studies
Level 1b	Therapy/Prevention, Aetiology/Harm	Individual RCT (with narrow Confidence Interval‡)
	Prognosis	Individual inception cohort study with > 80% follow-up; CDR† validated in asingle population
	Diagnosis	Validating** cohort study with good††† reference standards; or CDR† tested within one clinical centre
	Differential diag/symptom prevalence	Prospective cohort study with good follow-up****
	Economic and decision analyses	Analysis based on clinically sensible costs or alternatives; systematic review(s) of the evidence; and including multi-way sensitivity analyses
Level 1c	Therapy/Prevention, Aetiology/Harm	All or none§
	Prognosis	All or none case series
	Diagnosis	Absolute SpPins and SnNouts††
	Differential diag/symptom prevalence	All or none case-series
	Economic and decision analyses	Absolute better-value or worse-value analyses ††††
Level 2a	Therapy/Prevention, Aetiology/Harm	SR (with homogeneity*) of cohort studies
	Prognosis	SR (withhomogeneity*) of either retrospective cohort studies or untreated control groups in RCTs
	Diagnosis	SR (with homogeneity*) of Level >2 diagnostic studies
	Differential diag/symptom prevalence	SR (with homogeneity*) of 2b and better studies
	Economic and decision analyses	SR (withhomogeneity*) of Level >2 economic studies
Level 2b	Therapy/Prevention, Aetiology/Harm	Individual cohort study (including low quality RCT; e.g., <80% followup)
	Prognosis	Retrospective cohort study or follow-up of untreated control patients in an RCT; Derivation of CDR† or validated on split sample §§§ only
	Diagnosis	Exploratory** cohort study with good††† reference standards; CDR† after derivation, or validated only on split-sample§§§ or databases
	Differential diag/symptom prevalence	Retrospective cohort study, or poor follow-up
	Economic and decision analyses	Analysis based on clinically sensible costs or alternatives; limited review(s) of the evidence, or single studies; and including multi-way sensitivity analyses
Level 2c	Therapy/Prevention, Aetiology/Harm	"Outcomes" Research; Ecological studies
	Prognosis	"Outcomes" Research
	Diagnosis	
	Differential diag/symptom prevalence	Ecological studies
	Economic and decision analyses	Audit or outcomes research
Level 3a	Therapy/Prevention, Aetiology/Harm	SR (with homogeneity*) of case-control studies
	Prognosis	
	Diagnosis	SR (with homogeneity*) of 3b and better studies
	Differential diag/symptom prevalence	SR (with homogeneity*) of 3b and better studies
	Economic and decision analyses	SR (with homogeneity*) of 3b And better studies
Level 3b	Therapy/Prevention, Aetiology/Harm	Individual Case-Control Study
	Prognosis	
	Diagnosis	Non-consecutive study; or without consistently applied reference standards
	Differential diag/symptom prevalence	Non-consecutive cohort study, or very limited population
	Economic and decision analyses	Analysis based on limited alternatives or costs, poor quality estimates of data, but including sensitivity analyses IIncorporatingclinically sensible variations.
Level 4	Therapy/Prevention, Aetiology/Harm	Case-series (and poor quality cohort and casecontrol studies§§)
	Prognosis	Case-series (and poor quality prognostic cohort studies***)
	Diagnosis	Case-control study, poor or nonindependent reference standard
	Differential diag/symptom prevalence	Case-series or superseded reference standards
	Economic and decision analyses	Analysis with no sensitivity analysis
Level 5	Therapy/Prevention, Aetiology/Harm	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"
	Prognosis	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"
	Diagnosis	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"
	Differential diag/symptom prevalence	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"
	Economic and decision analyses	Expert opinion without explicit critical appraisal, or based on economic theory or "first principles"

Figure 14: Howick et al.'s 2009 *CEBM Levels of Evidence* [103], a complex hierarchy which provides rankings for evidence of therapy, prognosis, diagnosis, differential diagnosis and economic and decision analyses all in the same table. Howick et al.'s Levels are based on hierarchies developed by Ball & Phillips for EBonCall [130,131], and have since been updated to the 2011 *Levels of Evidence* [15,16].

2.2.6: Strength of Recommendation

Some hierarchies focus on recommendations to patients, and state how strong a recommendation can be justified on the basis of some evidence. The normal model is to appraise the level of evidence from a study or studies first, and then use that to set how strongly the treatment can be recommended. Some offer several levels or grades of recommendation, while others, such as GRADE, simply divide recommendations into ‘strong’ and ‘weak’ [50,205,281].

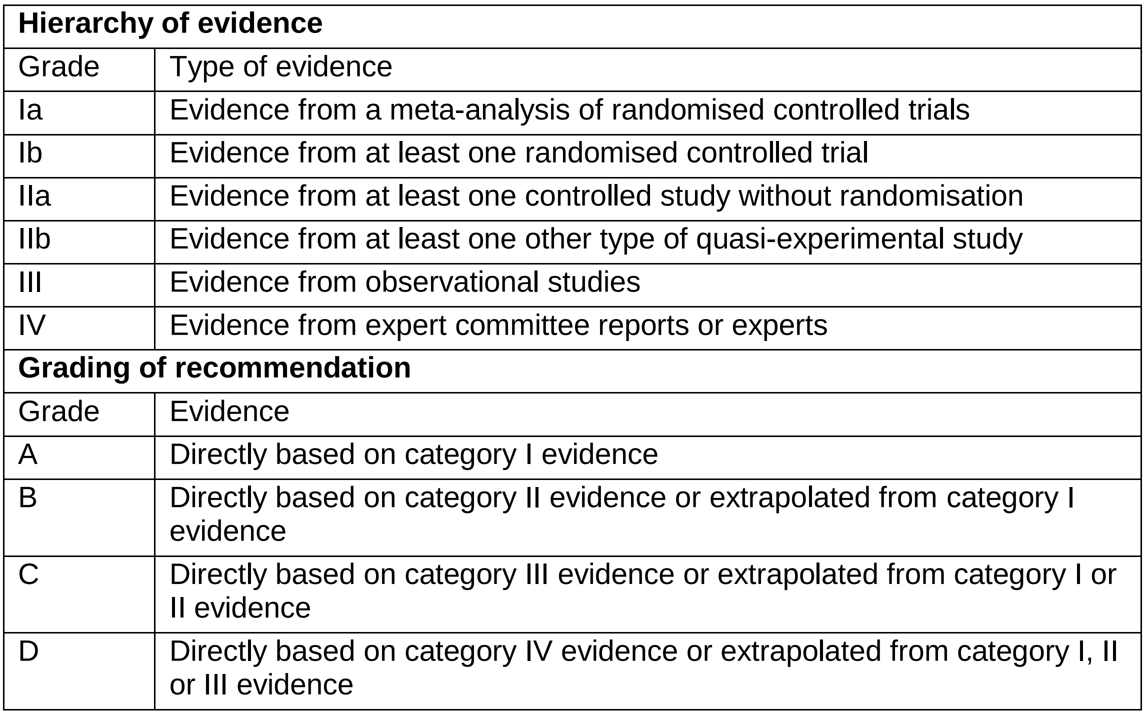
Many hierarchies have assumed a straightforward correspondence between level of evidence and the strength of recommendation. Howick et al.’s *CEBM Levels of Evidence* offer one such approach to translate levels of evidence into grades of recommendation (see Figure 15, below). Sometimes levels of evidence are simply equated to grades of recommendation, as in Chesson et al.’s hierarchy (see Figure 2, above), and in the GREG hierarchy which was used by NICE (see Figure 16, below), amongst others (e.g. [130,131,133,134,282,283]).



Grades of Recommendation

A	consistent level 1 studies
B	consistent level 2 or 3 studies or extrapolations from level 1 studies
C	level 4 studies or extrapolations from level 2 or 3 studies
D	level 5 evidence or troublingly inconsistent or inconclusive studies of any level

Figure 15: The CEBM hierarchy's approach [103] to translating levels of evidence into grades of recommendation. The Levels of Evidence appraises the evidence from individual studies, not evidence-bases, so involves amalgamation criteria at this stage. In effect, this builds a hierarchy of evidence-bases atop a hierarchy of evidence.



Hierarchy of evidence	
Grade	Type of evidence
Ia	Evidence from a meta-analysis of randomised controlled trials
Ib	Evidence from at least one randomised controlled trial
IIa	Evidence from at least one controlled study without randomisation
IIb	Evidence from at least one other type of quasi-experimental study
III	Evidence from observational studies
IV	Evidence from expert committee reports or experts
Grading of recommendation	
Grade	Evidence
A	Directly based on category I evidence
B	Directly based on category II evidence or extrapolated from category I evidence
C	Directly based on category III evidence or extrapolated from category I or II evidence
D	Directly based on category IV evidence or extrapolated from category I, II or III evidence

Figure 16: A version of the GREG hierarchy, developed by Mason & Eccles [284] and used by NICE in some of their guidelines, e.g. [28]. This hierarchy proposes a fairly straightforward correspondence between the grade of the evidence and the grade of the recommendation.

A more nuanced approach to strength of recommendation was used by Guyatt et al. in the *CHEST* hierarchies [35,285-287], building on work by Sackett and colleagues [151,288]. Their approach uses methodology *and* the ratio of benefits and harms to assess the strength of recommendation (see Figure 17, below). This allows for weak recommendations despite strong evidence if the evidence indicates that the treatment is only marginally beneficial. The GRADE approach to strength of recommendations, although not usually expressed outright through tables, resembles that used by the *CHEST* hierarchies. However, their model does assign special weight to quality of evidence, as only high quality evidence can warrant a strong recommendation—though a clear ratio of benefits to harms is also necessary (see [50,205]).

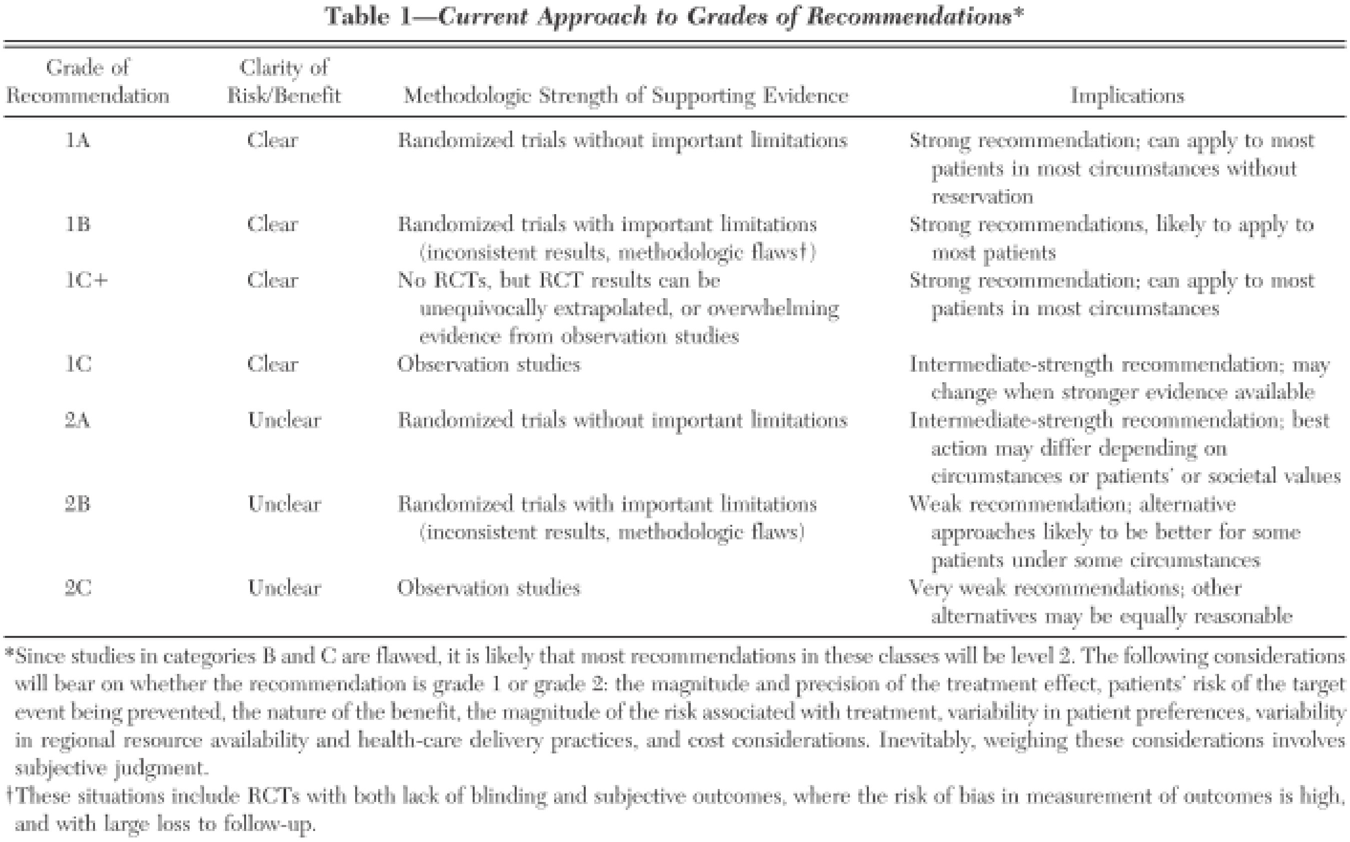


Table 1—*Current Approach to Grades of Recommendations**

Grade of Recommendation	Clarity of Risk/Benefit	Methodologic Strength of Supporting Evidence	Implications
1A	Clear	Randomized trials without important limitations	Strong recommendation; can apply to most patients in most circumstances without reservation
1B	Clear	Randomized trials with important limitations (inconsistent results, methodologic flaws†)	Strong recommendations, likely to apply to most patients
1C+	Clear	No RCTs, but RCT results can be unequivocally extrapolated, or overwhelming evidence from observation studies	Strong recommendation; can apply to most patients in most circumstances
1C	Clear	Observation studies	Intermediate-strength recommendation; may change when stronger evidence available
2A	Unclear	Randomized trials without important limitations	Intermediate-strength recommendation; best action may differ depending on circumstances or patients' or societal values
2B	Unclear	Randomized trials with important limitations (inconsistent results, methodologic flaws)	Weak recommendation; alternative approaches likely to be better for some patients under some circumstances
2C	Unclear	Observation studies	Very weak recommendations; other alternatives may be equally reasonable

*Since studies in categories B and C are flawed, it is likely that most recommendations in these classes will be level 2. The following considerations will bear on whether the recommendation is grade 1 or grade 2: the magnitude and precision of the treatment effect, patients' risk of the target event being prevented, the nature of the benefit, the magnitude of the risk associated with treatment, variability in patient preferences, variability in regional resource availability and health-care delivery practices, and cost considerations. Inevitably, weighing these considerations involves subjective judgment.

†These situations include RCTs with both lack of blinding and subjective outcomes, where the risk of bias in measurement of outcomes is high, and with large loss to follow-up.

Figure 17: Guyatt et al.'s ([285], p.4) *CHEST* hierarchy's approach to grading recommendations, which draws upon both the methodology of the evidence and the clarity of the risk/benefit ratio in assessing recommendations. This might arguably no longer be classed as a hierarchy, as the primary determinant of ranking is the clarity of the risk/benefit ratio, ahead of methodology. A similar approach is taken in other versions of *CHEST*'s hierarchy [35,286,287].

2.2.7: Other Interpretation Properties

Several other properties have been used in isolated instances. Barratt's 2009 hierarchy uses *weight* as its interpretation property—the higher ranked the evidence, the more weight it should be given:

> *"The hierarchy indicates the relative weight that can be attributed to a particular study design."* [106]

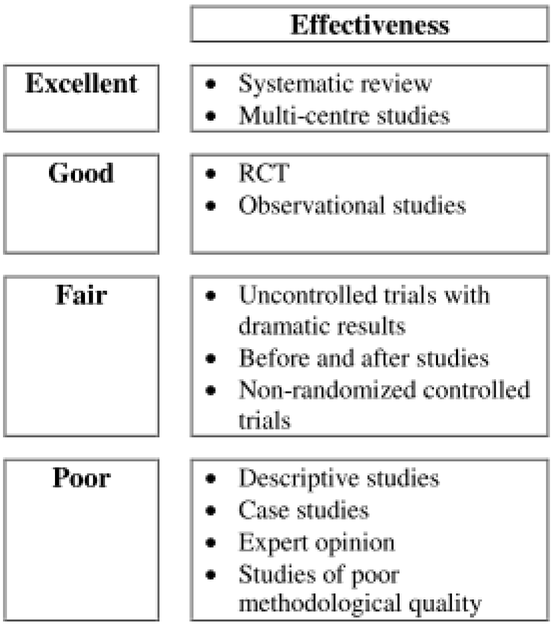
Weight is essentially a normative interpretation of *strength*—that is, to say "Give evidence E considerable weight", means: "Allow evidence E to have a strong effect upon your degree of belief in hypothesis H". Indeed, Barratt draws this connection, using both weight and strength in her presentation of her hierarchy. Barratt's account of weight is in terms of *permissible weight*—that is, the hierarchy tells us how much weight one is *allowed* to give to some evidence (i.e. how strongly we *can* let it affect our beliefs). A stronger version of a weight property could use *necessary weight*—that is, the hierarchy would tell us how much weight one *must* give to the evidence (i.e. how strongly it *should* affect our beliefs).

Barratt also uses robustness in her presentation of her hierarchy—*"the higher up a methodology is ranked, the more robust it is assumed to be*" [106]. In context, Barratt seems to use robustness as a justification for the interpretation in terms of weight. Barratt cites Guyatt et al.'s 2008 hierarchy as authority [83], which uses a similar notion of the likelihood of change in our beliefs as a criterion of evidence quality (see Section 2.2.1, above). In this sense, robustness seems to be merely a synonym for 'high internal validity'—robustness is simply high confidence in the accuracy of the result.

There is another sense of robustness which is alluded to by several hierarchy authors [6,192], but not directly employed in hierarchies to date. Robustness of effect refers to the persistence of the same effect under different conditions, i.e. across different patients, under different circumstances, administered by different clinicians in different settings, etc. Robustness in this sense is related to external validity.

2.2.8: SUMMARY

There are several different properties which hierarchies could give information about. The most common are quality, strength and level of evidence, and strength of recommendation. There has been some variation in the ways each term has been defined or used. When interpreting a hierarchy, users must decide what the intended interpretation property is, and how they are to understand that property. In some cases, there can be very little help from the author on that front. For instance, Evans' 2003 hierarchy ranks evidence from 'Excellent' to 'Poor', but it is not clear whether this refers to strength, quality, validity, etc. (see Figure 18, below).



	Effectiveness
Excellent	• Systematic review • Multi-centre studies
Good	• RCT • Observational studies
Fair	• Uncontrolled trials with dramatic results • Before and after studies • Non-randomized controlled trials
Poor	• Descriptive studies • Case studies • Expert opinion • Studies of poor methodological quality

Figure 18: Evans' 2003 hierarchy offers little guidance in the intended interpretation property. It is not unique in this respect. In particular, recent "evidence pyramids" have commonly left the intended interpretation unclear (e.g. [42,256,289,290]).

The intended interpretation property for a hierarchy is very important from a philosophical perspective. For instance, criticisms often relate to the claims about evidence quality or strength which result from hierarchies. But, as discussed above, hierarchies interpreted in terms of level of evidence or strength of recommendation need not necessarily commit to claims about the quality or strength of evidence (see Chapter 4 for further discussion).

2.3: Absolute and Relative Rankings

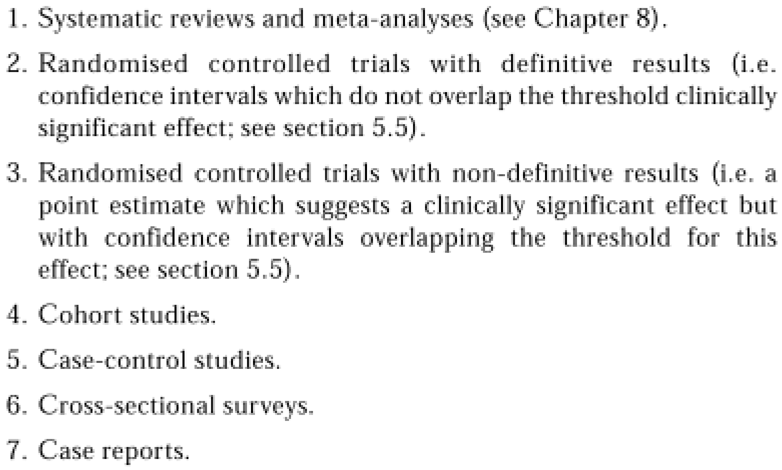
A crucial dimension of interpreting hierarchies is to decide whether they are *absolute* or *relative*. These terms have different meanings across philosophical disciplines, so a clear definition is needed. Relative rankings are *comparative*—they rank evidence from the highest to lowest degree of the interpretation property.[29] From a relative ranking of, say, evidence quality, it does not follow that high-ranked evidence is *high quality* evidence. It only follows that high ranked evidence is *higher quality* than lower ranked evidence. 'A is higher quality than B' is consistent with A being very high quality evidence, and B being high quality evidence, but also with A providing very low quality evidence, and B providing yet lower quality evidence. This is particularly philosophically important because a wide range of criticisms focus on the claim that RCTs provide high-quality evidence, or on the claim that unsystematic observations provide only low-quality evidence, and provide counterexamples to each. But if only a relative ranking is used, a hierarchy does not in fact entail that RCT evidence is high-quality. La Caze has claimed that only a relative interpretation of hierarchies can be justified [11]. Chapter 4 shows how relative rankings avoid many current criticisms, while Chapter 5 will be devoted to exploring relative rankings, arguing that they are uninformative.

A further interpretative assumption is also needed when interpreting relative rankings—whether the ranking is strict. In a strict ranking, a higher ranked study has a *greater* degree of the interpretation property than a lower-ranked study. In a non-strict ranking, it has a *greater-than-or-equal* degree. If the ranking is non-strict, then interpretations such as "RCT-evidence is higher quality than cohort study evidence" will not follow merely from the fact that RCT evidence outranks cohort study evidence. Some hierarchies seem to include strict *and* non-strict rankings, implicitly. For instance, Howick et al.'s CEBM hierarchy [103] has 5 levels, which are presumably strictly ranked, but level 1, 2 and 3 are stratified in 'a', 'b' and 'c' sub-rankings. Presumably 1a is a higher ranking than 1c, but it might be natural to assume the intent is to rank 1a-evidence as greater-or-equal quality to 1c-evidence, while 1c-evidence is strictly greater quality than 2a-evidence.

Despite the limited claims which follow from relative interpretations, they are common in the EBM literature (e.g. [8,32,42,43,194,208,256,265,289-295]). Often, no explicit scale is provided with a hierarchy, which suggests a relative ranking. A user wishing to use such a hierarchy as an absolute ranking would have to devise their own corresponding scale. Only a few hierarchy-makers explicitly state that their hierarchies should be given relative interpretations—Guyatt et al.'s [8] (see Figure 1,

[29] One could equally employ the terms "cardinal" and "ordinal" scale—however, terms like 'relative' are already used (albeit unreflectively) in the literature (e.g. [8,43,82,106]), and fit more neatly with existing medical terminology.

above), Devereaux & Yusuf's [293] and Greenhalgh's (see Figure 19, below) [43,296] hierarchies are rare examples.



1. Systematic reviews and meta-analyses (see Chapter 8).
2. Randomised controlled trials with definitive results (i.e. confidence intervals which do not overlap the threshold clinically significant effect; see section 5.5).
3. Randomised controlled trials with non-definitive results (i.e. a point estimate which suggests a clinically significant effect but with confidence intervals overlapping the threshold for this effect; see section 5.5).
4. Cohort studies.
5. Case-control studies.
6. Cross-sectional surveys.
7. Case reports.

Figure 19: Greenhalgh's 1997 hierarchy ([43], p.54). Greenhalgh published a similar hierarchy in an article series [296]. She explicitly states that this is "Standard notation for the relative weights" ([43], p.54) of evidence.

Absolute rankings ascribe some degree (or absolute range) of the interpretation property to the evidence or evidence-base. For instance, a hierarchy which ranks RCTs as 'High', interpreted in terms of quality, entails that evidence from an RCT is high quality. Comparative claims can follow from absolute ones—if A is high quality and B is low quality, it follows that A is higher quality than B. However, absolute rankings do not follow from relative rankings. In particular, all GRADE hierarchies are absolute rankings of evidence-bases. They ascribe one of 'High', 'Moderate', 'Low' or 'Very low' quality to the evidence base for an estimate of a treatment's effect on an outcome (see [12,14,297]).

2.4: SCOPE

For the purposes of this thesis, the 'scope' of a hierarchy refers to the breadth of types of evidence which are included within it. In addition to questions of whether particular methods are included, there are two ways to classify hierarchies by scope which will be useful here:

(1) Whether the hierarchy's scope is limited to evidence from *individual* studies, *pre-appraised* evidence, or contains both kinds.
(2) Whether the hierarchy's scope is limited to *epidemiological* evidence, *non-epidemiological* evidence[30], or contains both kinds.

2.4.1: PRE-APPRAISED EVIDENCE

Many hierarchies rank only evidence from individual studies, not including systematic reviews and meta-analyses in the ranking. A prominent example is GRADE [12], which includes only RCTs and observational studies as forming the evidence-bases it appraises.[31] This may be in part because GRADE is designed as a tool for the creation of systematic reviews. There are many other examples—many were originally used to perform systematic reviews (e.g. [31,37,46,142,284]), but by no means all (e.g. [90,193,207,216,282,285,298,299]).

Not all hierarchies necessarily omit pre-appraised evidence because they are targeted at systematic reviewers and guideline developers. Some EBM authors have serious concerns about meta-analysis in particular (*cf.* [1,285,300]). The Cochrane Collaboration handbook claims that:

> *"If bias is present in each (or some) of the individual studies, meta-analysis will simply compound the errors, and produce a 'wrong' result that may be interpreted as having more credibility."* ([1], p.247)

One argument for excluding pre-appraised evidence, then, is that to know the quality of a systematic review or meta-analysis, one must appraise the quality of the included studies. Notably, Guyatt et al.'s 1998 hierarchy which does not include pre-appraised evidence advises that:

> *"we should focus on the results of single large trials rather than a meta-analysis"*

[30] Only one hierarchy could be described as not containing epidemiological evidence—Daly et al.'s 2007 hierarchy [189] of qualitative evidence.

[31] Notable exceptions are the original GRADE Working Group paper [12] and the ATS Statement [37], which both include a catch-all "other evidence" category. Interpreting this catch-all category is particularly difficult (see below)

([285], p.4S)

This suggests that in at least some cases, pre-appraised evidence is omitted because it is viewed as unreliable or lower quality than the included evidence. Where pre-appraised evidence is omitted, users must make an assumption about whether these sources are to be omitted from their practice, treated with scepticism, or simply are beyond the scope of that hierarchy and *could* provide very high quality evidence (potentially higher quality than anything included in the hierarchy).

Some sources include pre-appraised evidence such as CATs and reports from journal clubs, in which the source is a pre-appraised single study. In making a CAT, the clinician performs a complete critical appraisal of the evidence provided by a study and makes this available to others to minimize duplication of efforts [301-303]. Large CAT databases have been created. A number of 'evidence pyramids' (e.g. [208,289,294,304]), as well as Haynes' pre-appraised evidence hierarchies [101,145,146], include CATs at a high level. No argument has yet been made as to why other clinicians' critical appraisals of evidence would provide stronger or higher quality evidence than the study being appraised.

2.4.2: Non-Epidemiological Evidence

"Non-epidemiological" evidence, in the sense described in Chapter 1, may or may not be included in the scope of a hierarchy. There are at least four broad classes of non-epidemiological evidence, some of which may be included while others are excluded. These are: expert opinion, clinical experience, laboratory research, and mechanistic reasoning (see Section 1.2.3).

Many hierarchies do not include any non-epidemiological evidence sources. GRADE [12] is limited to RCTs and observational studies. Almost no hierarchies included any non-epidemiological sources until the late 1990s (see Chapter 3 for a full discussion) (e.g. [43,190,193,286]). A crucial interpretative question when the scope is restricted to epidemiological evidence is what users can infer about the evidence from other sources. This is particularly philosophically important where criticisms are addressed at EBM and hierarchies for undervaluing experience, expertise and mechanistic research (see Chapter 4). Early EBM texts emphasized that EBM was about integration of epidemiological evidence with other sources (e.g. [19-22,78])—with this in mind, users may read hierarchies as a ranking of epidemiological evidence, which leaves them to appraise non-epidemiological evidence independently. In other words, such a hierarchy is mute on the topic of the quality or strength of non-epidemiological evidence. However, an alternative interpretation is that any sources not ranked in a hierarchy provide extremely low-quality evidence, or even no evidence at all. Authors are rarely circumspect in delineating the intended scope of their hierarchy.

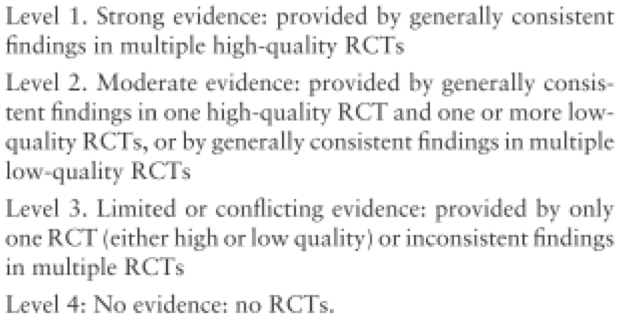
Some hierarchies include some but not all non-epidemiological evidence. For instance, a number include expert opinion and clinical consensus at the lowest level, but do not include laboratory research or mechanistic reasoning of any kind (e.g. [31,90,102,105,133,192,284]). It is difficult to determine whether these hierarchies are supposed to imply that mechanistic evidence is not evidence, is weak evidence, or is appraised separately. Others include laboratory research, but not expert opinion [8,15]. This may reflect statements from the GRADE group that expert opinion does not constitute evidence (see Section 1.4), and be intended to be interpreted that way.

2.4.3: Exhaustiveness and Superlative Interpretations

Some critics have read EBM hierarchies as implying that the highest-ranked evidence is the strongest or highest-quality evidence possible. These are superlative interpretations of hierarchies. Note, however, that only an exhaustive hierarchy—a hierarchy which includes all evidence at some level—licenses such superlative claims. This is particularly clear in those hierarchies which do not include pre-appraised evidence. In every hierarchy in which they appear, systematic reviews and meta-analyses are top-ranked. Yet there are many hierarchies in which they do not feature. Giving a superlative interpretation to those hierarchies is unjustified, unless coupled with the interpretative assumption that anything not explicitly included in a hierarchy is either low-ranked or non-evidential. Despite misgivings about meta-analysis, this seems implausible as an EBM position, especially with respect to systematic reviews of RCTs.

However, the same reasoning also applies to non-epidemiological evidence. Where non-epidemiological evidence is not included in a hierarchy, superlative interpretations cannot be justified (unless an interpretative assumption is made that non-included evidence is implicitly low-quality or not evidence at all). Interpretative assumptions about what to make of un-included evidence are therefore crucial to the philosophical consequences of hierarchies.

Some hierarchies are exhaustive by virtue of “catch-all” categories. These are always at the lowest level of the hierarchy, and include any evidence not assigned to a higher category. Examples include the “other evidence” categories from some versions of GRADE [12,37,46,278] and SORT [104]. A particularly extreme example comes from Van Tulder et al. [305] (see Figure 20, below), whose hierarchy includes only RCT evidence, with a catch-all category for any evidence not from an RCT.



Level 1. Strong evidence: provided by generally consistent findings in multiple high-quality RCTs

Level 2. Moderate evidence: provided by generally consistent findings in one high-quality RCT and one or more low-quality RCTs, or by generally consistent findings in multiple low-quality RCTs

Level 3. Limited or conflicting evidence: provided by only one RCT (either high or low quality) or inconsistent findings in multiple RCTs

Level 4: No evidence: no RCTs.

Figure 20: An extremely RCT-focused hierarchy of evidence bases from Van Tulder et al. [305], featuring an example of a catch-all category.

Without catch-all categories, it is unlikely that a hierarchy will be exhaustive. There are a great many designs for research which are not necessarily captured by terms like "RCT" and "observational study".[32] Some of this variety has been captured by modern hierarchies which include a wider range of evidence. For instance, Roman, Silbersweig & Siu [299] include evidence from observational studies of risk-factors, evidence from inferences from RCTs in other populations, and evidence from RCTs where the treatment was one component of a multi-component intervention. McAlister et al. include evidence from subgroup analyses [216]. It is not clear whether these would fall under the traditional categories of "RCT" and "observational study", in, for instance, the GRADE approach. Systematic reviews of observational studies have been included in some hierarchies (e.g. [29,132])—it is not clear where these would fit elsewhere.

2.4.4: Summary

Hierarchies can include epidemiological, non-epidemiological and pre-appraised evidence sources. The scope of a hierarchy is crucial for the implications about evidence of using a hierarchy—hierarchies with scope limited to epidemiological evidence make no claims about non-epidemiological evidence, for instance. Superlative claims are not usually licensed by hierarchies, as hierarchies are only exhaustive where there are catch-all categories. This means that non-included evidence could be

[32] Bluhm [74] considers a particular instance of non-exhaustiveness due to a conflation between the terms "observational study" and "nonrandomized study". Nonrandomized studies can be experimental where patients are allocated to experimental or control groups by some non-random procedure. Historically-controlled trials may not be observational studies, where the researchers intervene to change the experimental group's treatment.

stronger or higher-quality than that included in a hierarchy. Further interpretative assumptions about what hierarchies imply, if anything, about non-included evidence must be made, but little to no guidance is offered in the literature.

2.5: Conditions and Categoricalness

2.5.1: Categorical Interpretations

Adam La Caze argues that EBM hierarchies should be given *non-categorical* interpretations ([11,72], and see [306]). This section shows that La Caze's notion of 'categoricalness' cross-cuts the interesting properties of hierarchies. There are two aspects of hierarchy interpretation which are blended together in La Caze's notion of categoricalness—conditions and modifiers—which are best separated. In fact, most sophisticated hierarchies could be represented as categorical or non-categorical hierarchies—while conditions and modifiers are not eliminable in this way.

La Caze offers no definition of categoricalness, but does make a claim:

> "*The categorical interpretation of the hierarchy holds that evidence from higher up the hierarchy* trumps *evidence from lower down.*" ([72], p.3, emphasis added)

For him, the most important consequence of categorical interpretations is:

> "All *the results of a randomised study are* always *superior to the results of studies from lower down the hierarchy*" ([72], p.9, La Caze's emphasis).[33]

Indeed, in a categorical interpretation, all evidence from a higher level is always superior to all evidence from a lower one.

This version of categoricalness conflates together three ideas which need untangling. First, La Caze conflates *trumping* with *outranking*. If Evidence A *trumps* Evidence B, then any disagreement between A and B is irrelevant—should they contradict, B is dismissed.[34] By contrast, when A *outranks* B, users do not dismiss B. They put more stock in A, trusting its findings more than B's. But if A and B contradict, the user be more likely to weigh the study's findings against one another, lowering confidence in the estimate of the treatment effect. If A *trumps* B, however, B's findings do not affect belief in the hypothesis. *Lower ranked* evidence may still be taken into account, but *trumped* evidence is not.

[33] Note the judicious use of "superior" to avoid the question of interpretation properties.

[34] The metaphor originates from card games—a trump suit beats a card from a non-trump suit regardless of the values of the cards. If, say, RCTs trump observational studies, then if an RCT and observational study disagree, the RCT result is accepted *no matter* how big, carefully conducted and long the observational study was, *no matter* how clear and large the result of the observational study and how equivocal and small the result of the RCT, and *no matter* whether the RCT was small, well-conducted or short.

La Caze's important consequence of the categorical interpretation does not require trumping, only outranking. 'Superiority' does not require that the inferior study's results are dismissed, just that the superior study's results carry more weight. So I propose an alternate definition:

> A hierarchy is given a *categorical* interpretation just if (for every level X) it is inferred that *all* evidence at level X in the hierarchy *always* possesses property P to a greater degree than *all* evidence at all lower levels. *(where P is the interpretation property)*

It is possible to offer *semi-categorical* and indeed *semi-trumping* hierarchies, in which some but not all levels are given categorical or trumping interpretations. Such semi-categorical interpretations are quite plausible; for instance, lower-level evidence might not *always* beat yet-lower-level evidence, but high-level evidence is always superior to low-level evidence.

Clarity about whether trumping interpretations are defended is important to avoid making a straw man of EBM proponents—whether a hierarchy is interpreted as trumping or not seriously affects which evidence users take into account. Trumping has been roundly criticised, for good reason (see e.g. [72,75,124,163,186])—after all, low-quality evidence is still evidence. But it is also important to be clear that very few serious EBM proponents defend trumping.[35] Trumping is not built into hierarchy systems. By contrast, the categorical interpretation as redefined here seems fairly commonplace in the EBM literature.

La Caze seems to have GRADE hierarchies in mind when he calls for non-categorical interpretations. GRADE allows studies to be up- and down-graded, which means that initially high-ranked evidence like that from an RCT could ultimately be appraised as lower-quality than initially low-ranked evidence from an observational study. But GRADE is actually equivalent to a categorical hierarchy in which certain conditions in addition to methodology are required for some evidence to reach a particular rank. If the sum of up- and down-grading for some evidence is called the 'GRADE

[35] There are very few examples of hierarchy-makers defending trumping interpretations of their hierarchies, with the exception of Niederman & Richards in 2010 [208]. The only example of semi-trumping defended in the secondary literature which I have so far located is Barton's "*The best RCT still trumps the best observational study*" [307]. Jeremy Howick [163] reports from personal communication that Julian Higgins of the Cochrane Collaboration defends the claim that RCTs *either* trump observational studies *or* are not comparable. This is a very different claim to outright trumping, and does not seem reflective of views expressed elsewhere (see e.g. [1]). Few authors have explicitly *rejected* trumping interpretations—but then, few authors discuss the details of their hierarchies at all (exceptions include NICE [2,28], see also [97,308]). It is important not to mistake advice about which evidence to look at *if time is limited* (e.g. [10], p.118; [8], pp.14-5) for the philosophical claim that the rest of the evidence is *trumped* or should be ignored (*cf.*[72,124]).

coefficient' of that evidence, then one could formulate a hierarchy without talk of up and down grading:

Quality of Evidence	
High	Evidence from an RCT with GRADE-coefficient ≥ 0, or from an observational study with GRADE-coefficient ≥ 2
Moderate	Evidence from an RCT with GRADE-coefficient = -1, or from an observational study with GRADE-coefficient = 1
Low	Evidence from an RCT with GRADE-coefficient = -2, or from an observational study with GRADE-coefficient = 0
Very Low	Evidence from an RCT with GRADE-coefficient ≤ -3, or from an observational study with GRADE-coefficient ≤ -1

Figure 21: A putative hierarchy of evidence which is completely equivalent to GRADE, where the GRADE-coefficient is defined as the result of summing the up- and down-grading criteria as applied to the evidence.

This hierarchy is completely equivalent to GRADE, but fits the definition for a categorical interpretation—evidence from each level outranks evidence from those below.

Such an approach has in fact been taken, albeit in a more limited way, by Schunemann et al. [37] in the ATS version of GRADE (see Figure 10, above). They allow for downgraded RCTs or upgraded observational studies to provide moderate quality evidence. This version is not equivalent to the standard GRADE hierarchy, as it does not allow multi-level up- and down-grading.

The important point here is that hierarchies can be given categorical or non-categorical interpretations depending in part on whether there are any *conditions* taken into account other than the methodology which produced the evidence. The GRADE up- and down-grading criteria are examples of conditions which evidence must fulfil other than methodology to attain a particular ranking. The majority of hierarchies include at least some conditions. Conditions are one aspect of what La Caze is trying to highlight in drawing attention to categoricalness—that there are things other than merely being produced by an RCT methodology which affect ranking in a hierarchy.

2.5.2: Conditional Hierarchies

Hierarchies of evidence *primarily* rank evidence or evidence-bases in terms of the methodology which produced that evidence. However, in many cases, there are other criteria which must be met for some evidence to attain a particular rank in a hierarchy. Call these "conditions". Normally, these conditions allow users to discriminate between different levels of evidence from the same methodology. For instance, in Sackett's 1989 hierarchy [298], RCT evidence is top-ranked, but could be Level I or Level II depending on whether the trial was large and whether the results are clearcut (see Figure 22, below). In other cases, such as in GRADE, the methodology is used to set the initial ranking for the evidence, then satisfying certain conditions can move the evidence up or down the ranking.

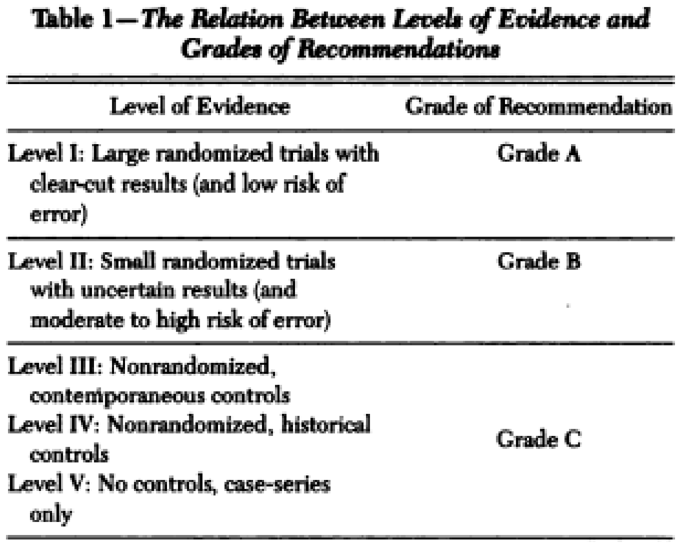


Table 1—*The Relation Between Levels of Evidence and Grades of Recommendations*

Level of Evidence	Grade of Recommendation
Level I: Large randomized trials with clear-cut results (and low risk of error)	Grade A
Level II: Small randomized trials with uncertain results (and moderate to high risk of error)	Grade B
Level III: Nonrandomized, contemporaneous controls Level IV: Nonrandomized, historical controls Level V: No controls, case-series only	Grade C

Figure 22: Sackett's 1989 *CHEST* hierarchy [298] using conditions to discern between two levels of RCT evidence, and corresponding to grades of recommendation.

Hierarchies have used a variety of conditions. Perhaps the most common is some form of *study quality condition*, which requires that the evidence come from a study "of high quality", or "well-performed", etc. This is sufficiently vague to allow sophisticated evidence-appraisers to omit or down-rank evidence from studies with methodological flaws or problems. Quality conditions may be more appropriate where the interpretation property is level of evidence or strength of recommendation. A number of hierarchies require that studies be "*high-quality*" to access higher levels (e.g. [2,29,104,133,134,305,308]). Others highlight particular aspects of a judgment of quality, such as that the study was '*well-conducted'* [5,29,37,207], *'well-designed'* [6,24,31,105,142,185,308], '*without important limitations'* [285], *'properly randomized'* [24,31,309] or *'with low risk of bias'* [2,133,298].

The size of the study is also often a factor. Several hierarchies require that studies be "*of appropriate size*" to detect effects [102,310] or "*properly powered*" [5,132], be "*large*" [293,298], or in one case even have a "*minimum sample size of 30 in each arm*" [32]. Some are concerned that trials should be multi-center, from more than one research group [31,105,132,207].

Many hierarchies refer to the results of the study in addition to the methodology. For instance, some require that results are statistically significant [90,282], with a narrow confidence interval [103,135], or that they are "*clear-cut*" [298], "*definitive*" [43,106], or showing a "*dramatic effect*" [15] in order to access a higher ranking. Sometimes, many conditions are combined together, as illustrated in the range of requirements for top-level evidence in the AAN hierarchy (see Figure 23, below).

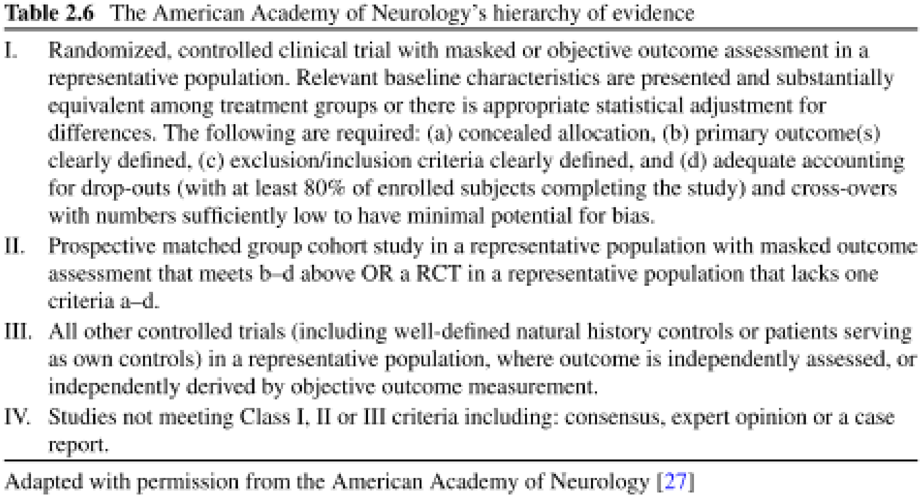


Table 2.6 The American Academy of Neurology's hierarchy of evidence

I.	Randomized, controlled clinical trial with masked or objective outcome assessment in a representative population. Relevant baseline characteristics are presented and substantially equivalent among treatment groups or there is appropriate statistical adjustment for differences. The following are required: (a) concealed allocation, (b) primary outcome(s) clearly defined, (c) exclusion/inclusion criteria clearly defined, and (d) adequate accounting for drop-outs (with at least 80% of enrolled subjects completing the study) and cross-overs with numbers sufficiently low to have minimal potential for bias.
II.	Prospective matched group cohort study in a representative population with masked outcome assessment that meets b–d above OR a RCT in a representative population that lacks one criteria a–d.
III.	All other controlled trials (including well-defined natural history controls or patients serving as own controls) in a representative population, where outcome is independently assessed, or independently derived by objective outcome measurement.
IV.	Studies not meeting Class I, II or III criteria including: consensus, expert opinion or a case report.

Adapted with permission from the American Academy of Neurology [27]

Figure 23: A hierarchy from Elamin & Montori ([311], p.21) by the American Academy of Neurology (AAN) [312], in which a number of conditions, primarily focused on methodology, are required to reach Level I evidence. Conditions are also imposed for Level II and III.

When hierarchies include systematic reviews and meta-analyses, they often differentiate between reviews where the results of included studies were "homogeneous" vs. "heterogeneous" [5,90,103,104,135,151,190], or "consistent" vs. "inconsistent" [197], and rank systematic reviews with homogeneous/consistent results higher. Guyatt et al.'s 1995 *Users' Guides* hierarchy requires that all RCT results have confidence intervals on the same side of the threshold number needed to treat to qualify for the highest ranking [190], while the NHMRC requires that all included trials are clinically relevant [6].

2.5.3: Interpreting Conditions

There are interpretative assumptions which must be made in order to apply a hierarchy with conditions. First, users must decide whether to assume that a condition is fulfilled unless there is evidence to the contrary, or to assume it is not fulfilled unless the study's authors demonstrate that it is. The difference is most acute when there is no evidence or information which tells in either direction. Despite efforts to improve the reporting of medical studies such as CONSORT [154] and STROBE [313], many trial-reports still lack information about, for instance, how blinding was ensured and whether it was checked [314-316], how randomization was performed (if it was performed), which baseline factors were checked for confounding, etc. Moreover, there is a question of how a user needs to look for evidence that a condition is (or is not) fulfilled, and what would count as good evidence for that. In some cases this is relatively easy—one can check the trial report to see how many patients were involved—but in others, such as whether the trial was well-conducted, this is less clear. In cases such as GRADE's downgrading criterion for publication bias [198], finding out whether the problem affects the studies in a systematic review may require extensive study using sophisticated techniques. Should users downgrade all evidence-bases unless there is evidence that publication bias is not affecting the results, or only downgrade when evidence (e.g. a funnel plot [250]) shows reason to suspect publication bias at work?

2.6: MODIFIERS

Adding a modifier to the intended interpretation of a hierarchy allows a more nuanced interpretation. For instance, instead of claiming that a hierarchy which ranks RCT evidence as high-quality implies that "Evidence from an RCT is high quality", a modified interpretation could be: "Evidence from an RCT is *probably* high quality", or "Evidence from an RCT is *usually* high quality". Modifiers are another way of having a non-categorical interpretation of a hierarchy. For instance, suppose a hierarchy which ranks RCT evidence above cohort study evidence is given a modified interpretation: "Evidence from an RCT is *probably* higher quality than evidence from a cohort study". This interpretation does not assert that *all* RCT evidence is higher quality than *all* cohort study evidence. Therefore, the ranking is not categorical. Any hierarchy could be given a non-categorical interpretation by adding suitable modifiers. Modifiers and conditions can also be combined.

There are a number of modifiers which could be used to create nuanced interpretations of hierarchies. Probabilistic modifiers can be used in several ways (e.g. "The evidence is *probably* high-quality"). In one reading, they allow users to express uncertainty about the quality, strength or level of evidence that follows from a hierarchy. For instance, if a hierarchy is interpreted as implying "The evidence from this RCT is probably high-quality", then this can be understand as meaning that one can be justifiably confident that the RCT evidence is high-quality. This is one species of doxastic modifier—a modifier relating to beliefs. These interpretations include, for instance: "One is justified in believing that the evidence is high-quality", or "One can be confident that the evidence is high-quality". Hierarchies may also be given a normative interpretation, such as: "One should believe that the evidence is high-quality" or "One should act according to the assumption that the evidence is high-quality". Note that in each case, the modified hierarchy interpretation does not entail a factual claim about the evidence (that it *is* high-quality evidence), only about justification for beliefs and actions.

Such interpretations are defended in the EBM literature when hierarchies are viewed as *heuristics* or *decision-aids* (see [15,16,97]). Hierarchies as heuristics are not expected to provide infallible information about evidence, but to guide decision-making by telling users to act as if the evidence has certain properties. A heuristic interpretation of a hierarchy could read a particular ranking as: "Acting as if the evidence is high-quality is justified".

Heuristics are sometimes described as short-cuts which omit certain complexities in order to speed up decision-making [317]. They work under certain circumstances (where the omitted complexities have little or no effect on the appraisal), but fail under others (where the complexities have an important effect). Good heuristics work under normal conditions and fail under abnormal ones—preferably ones which users will be able to identify. This suggests another modifier (or alternatively, a condition) for hierarchy interpretation: "*Under normal conditions*, acting as if the evidence is high-quality is justified". Crucially, heuristics are justified not by theoretical argument but by empirical validation. Chapter 7 focuses on heuristic interpretations.

Another reading of a probabilistic modifier understands the probability not as a claim about uncertainty, but as a claim about frequencies. This is particularly relevant for the interpretation of hierarchies of methodologies and hierarchies of evidence bases. A hierarchy of methodology which ranks RCTs highly could be interpreted as: "RCTs *probably* provide high-quality evidence", where this is understood as the claim that a high proportion of RCTs provide high-quality evidence. A hierarchy of evidence bases could be given the reading: "An evidence base which contains multiple RCTs with consistent results *probably* gives a high level of evidence." Again, the intended interpretation could be that, as a matter of fact, most evidence bases which contain multiple consistent RCTs are high level evidence bases.

Probabilistic interpretations are sometimes seen, albeit implicitly, in the EBM literature. The most common claim which justifies ranking RCTs above observational studies, laboratory research and expert opinion is that relying on the lower-ranked sources is *riskier* than relying on higher-ranked sources. The risk is risk of bias, and with it risk of mistaken belief. It is often claimed by EBM proponents that while laboratory research *can* provide strong, high-quality evidence for a hypothesis about treatment effects, clinicians are more likely to form false beliefs if they base beliefs on laboratory research than on epidemiological evidence (e.g. [20,22,58,59,77,93]).

A final type of modifier, defended by La Caze [11], is a *ceteris paribus* clause. *Ceteris paribus*, 'all else being equal', is a modifier which can be applied to relative rankings. Suppose a hierarchy ranks RCT evidence above observational study evidence. A *ceteris paribus* reading of this ranking would be: 'All else being equal, evidence from an RCT is higher-quality than evidence from an observational study'. The *ceteris paribus* clause covers variables such as number of participants, size of the effect shown, successful implementation of the methodology, etc. This modifier limits the claims to like-for-like comparisons, and blocks, for instance, the claim that evidence from a small badly-performed RCT is higher quality than evidence from a large well-controlled cohort study.

La Caze's 'categorical' interpretation has evidence from a study at a particular level in a hierarchy *always* being that level of quality, strength, etc. Modifiers allow hierarchy users to avoid such categorical readings. Probabilistic modifiers allow users to retain uncertainty about the degree to which the evidence has the interpretation property, or to express confidence in the quality, strength or level of evidence, rather than a direct factual claim about the evidence. The claim which worries La

Caze, that “All *the results of a randomised study are* always *superior to the results of studies from lower down the hierarchy*” ([72], p.9, La Caze’s emphasis), is not implied by a probabilistic or doxastic interpretation of hierarchies. A probabilistic relative interpretation asserts only that RCT evidence is probably superior to lower-ranked evidence, while a doxastic interpretation asserts that one can be quite confident that the RCT evidence is superior to lower-ranked evidence, but does not claim certainty. *Ceteris paribus* clauses are consistent with a categorical reading, but only apply where all properties of the evidence other than the underlying methodology are held constant, so avoid many of the problems associated with categorical interpretations which La Caze identified (see Chapter 4 for further discussion).

2.7: Conclusion: The Anatomy of Interpretation

In order to interpret a hierarchy, users must make assumptions on at least six dimensions. What is the subject matter of the ranking? Hierarchies can rank evidence from individual studies, an evidence base potentially formed of evidence from multiple studies, or the methodologies themselves. Which property does this hierarchy ascribe to its subject-matter? Quality, strength, validity, level of evidence and strength of recommendation are all used in the EBM literature. Does this hierarchy ascribe absolute degrees of the property to the subject matter, or does it only provide a relative ordering? Are all kinds of evidence included in the ranking, or is it restricted to epidemiological evidence? What, if anything, should one take the hierarchy to imply about evidence which falls outside of its explicit scope? Are there any conditions other than the underlying methodology which affect the final ranking in the hierarchy? Should one assume that the conditions are fulfilled unless proven otherwise, or does one need evidence that the conditions have been fulfilled? Finally, are any modifiers attached to the intended interpretation of the hierarchy? Does it imply that higher-ranked evidence is always superior to lower-ranked evidence, or only a probabilistic or doxastic claim?

The answers to these questions affect the way users of even the same hierarchy will appraise evidence. The same hierarchy is almost always compatible with multiple interpretations. Few authors are explicit about all, and sometimes any, of these interpretative assumptions. Suppose a hierarchy ranks RCT evidence above observational study evidence, and is explicit that the hierarchy is a hierarchy of evidence ranking quality of evidence. Even with these assumptions specified, and without adding any conditions, this is still consistent with the following very different interpretations:

> 'The evidence from the RCT is high quality, and the evidence from the observational study is low quality.'
>
> 'We are probably justified in believing that the evidence from the RCT is higher quality than the evidence from the observational study.'

Chapter 3 uses this framework to present a history of the development of hierarchies in terms of the changing interpretations hierarchies are given, as well as a summary of the interpretations given to hierarchies by philosophers. The interpretative assumptions made affect the strength of the interpretations, and thus the justifications needed for a hierarchy. Chapter 4 presents criticisms of hierarchies of evidence, and shows that the majority of criticisms only apply to hierarchies given certain interpretations. Some hierarchies may strongly suggest these well-criticised interpretations, but others do not.

Chapter Three:

On the Interpretation of Hierarchies

Contents

Chapter 2 demonstrated that the idea of a single EBM theory of evidence which follows from the widely-endorsed hierarchy of evidence is a double myth. There is no single accepted hierarchy. Even when a particular hierarchy is chosen, there are at least six dimensions of interpretative assumptions which must be made in order to information about evidence from it.

This does not mean that there are no systematic theories of evidence espoused by EBM proponents. In fact, there are many approaches to evidence, varying according to both the design of hierarchy defended and the interpretative assumptions made. Small changes in the appearance of hierarchies can indicate major underlying shifts and differences in interpretative assumptions. This chapter brings these interpretations to the surface.

This chapter takes two approaches: (1) to identify the interpretations of hierarchies accepted by EBM proponents over time, and (2) to identify those which philosophers and practitioners attribute to EBM. Section 3.1 presents the first systematic history of hierarchies, distinguishing four periods in the development of hierarchies from the 1960s to present day. Each of these four periods is particularly associated with a set of interpretative assumptions. The design of hierarchies as well as the surrounding discussion in the EBM literature combines to offer changing pictures of evidence.

The second approach takes a philosophical focus (Section 3.2). There are several accounts of evidence appraisal espoused by proponents of EBM and attributed to EBM by critics in philosophical papers. To facilitate comparison between these and the apparent accounts in the development of hierarchies, each is reformulated here using the framework developed in Chapter 2.

3.1: The Development of Hierarchies

Hierarchies did not appear or rise to prominence simultaneously with Evidence-Based Medicine. They pre-date EBM, yet were not advocated by EBM proponents until the mid-1990s [190], despite the EBM movement originating in 1991 at the latest [39,64]. This section introduces a novel four-part history of hierarchies, identifying the prominent interpretative assumptions being made in each period, and sketching the theory of evidence which results.

3.1.1: Clinical Epidemiology—Hierarchies Before EBM (1979-c.1991)

Hierarchies of evidence pre-date the EBM movement. Some sources (e.g. [318,319]) cite Campbell & Stanley's 1963 *Experimental and Quasi-experimental Designs for Research* [155] as the original evidence hierarchy. Campbell & Stanley present tables in which a range of study-designs are appraised according to vulnerability to twelve potential 'Sources of Invalidity' ([155], pp.8,40,56—see Fig.24, below) They claim that 'true experimental' designs are generally vulnerable to fewer sources of invalidity than 'quasi-experimental' designs, which in turn have fewer potential sources of invalidity than 'pre-experimental' designs ([155], pp.70-1). The conclusion (although expressed with reservations and qualifications) is that using experimental or at least quasi-experimental designs where possible will minimise invalidity.

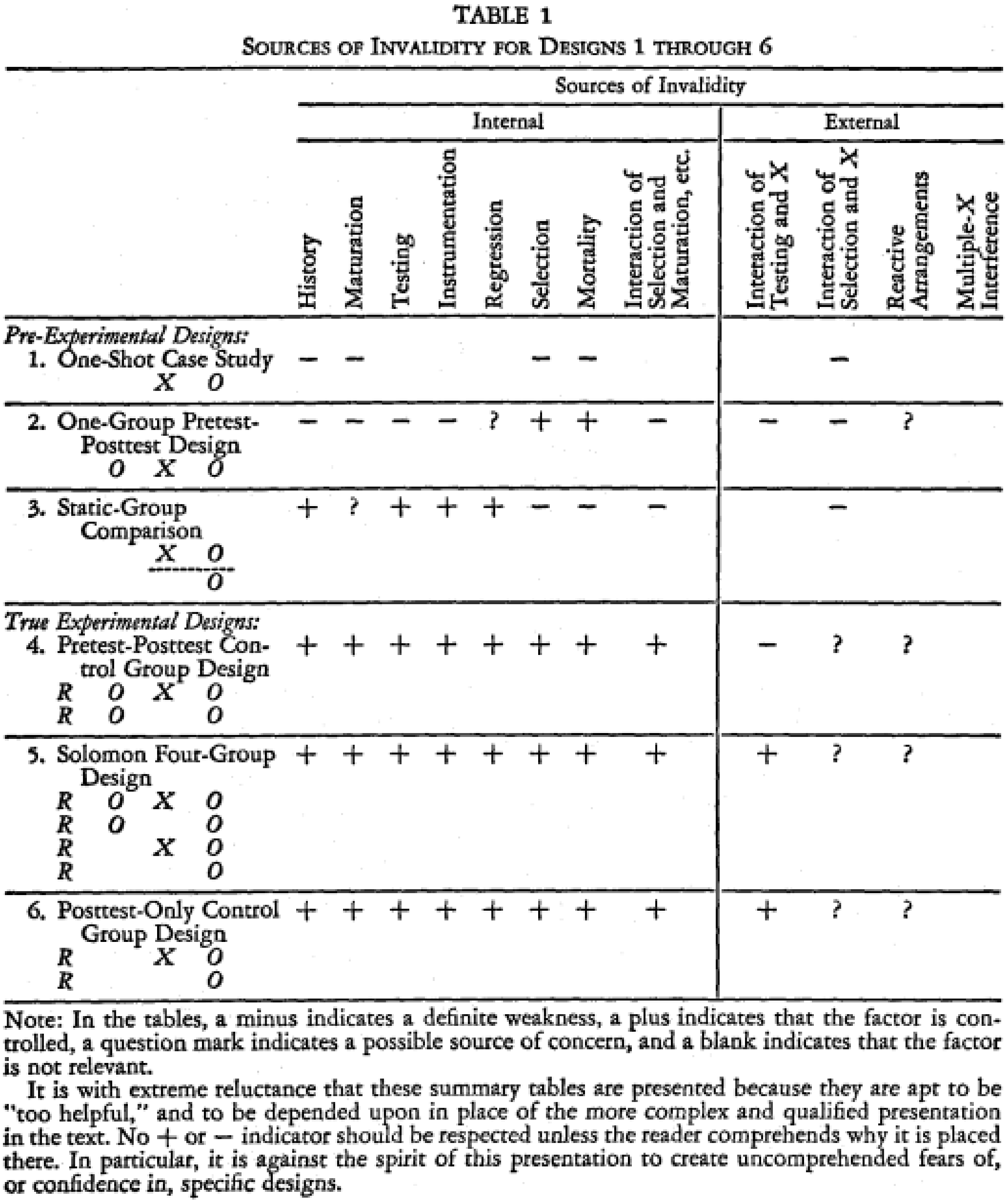


TABLE 1
SOURCES OF INVALIDITY FOR DESIGNS 1 THROUGH 6

	Sources of Invalidity											
	Internal								External			
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Maturation, etc.	Interaction of Testing and X	Interaction of Selection and X	Reactive Arrangements	Multiple-X Interference
Pre-Experimental Designs:												
1. One-Shot Case Study *X O*	−	−				−	−			−		
2. One-Group Pretest-Posttest Design *O X O*	−	−	−	−	?	+	+	−	−	−	?	
3. Static-Group Comparison *X O* - - - - - - *O*	+	?	+	+	+	−	−	−		−		
True Experimental Designs:												
4. Pretest-Posttest Control Group Design *R O X O* *R O O*	+	+	+	+	+	+	+	+	−	?	?	
5. Solomon Four-Group Design *R O X O* *R O O* *R X O* *R O*	+	+	+	+	+	+	+	+	+	?	?	
6. Posttest-Only Control Group Design *R X O* *R O*	+	+	+	+	+	+	+	+	+	?	?	

Note: In the tables, a minus indicates a definite weakness, a plus indicates that the factor is controlled, a question mark indicates a possible source of concern, and a blank indicates that the factor is not relevant.

It is with extreme reluctance that these summary tables are presented because they are apt to be "too helpful," and to be depended upon in place of the more complex and qualified presentation in the text. No + or − indicator should be respected unless the reader comprehends why it is placed there. In particular, it is against the spirit of this presentation to create uncomprehended fears of, or confidence in, specific designs.

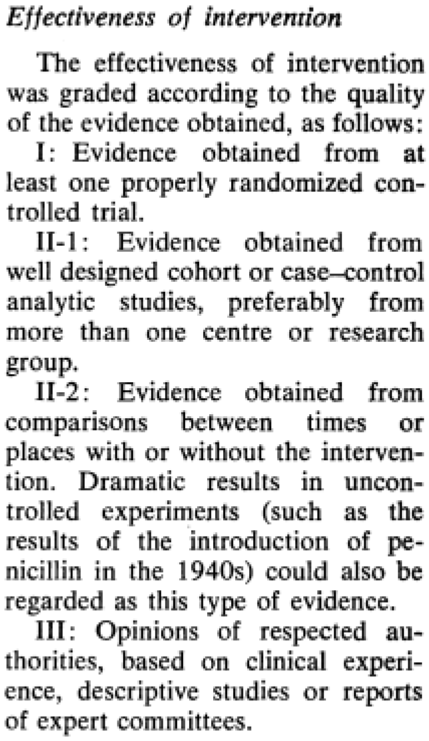
Figure 24: An example of a table from Campbell & Stanley's *Experimental and Quasi-Experimental Designs for Research*, p.8. Note that the numerical values attached to designs do not constitute a ranking, and that again Campbell & Stanley caution against over-reliance on tables in their caption: their "*extreme reluctance*" results from a fear that tables will be misunderstood unless read with the extensive surrounding commentary.

However, despite the superficially similar appearance of their tables to some hierarchies, Campbell & Stanley are explicit that their work should not be interpreted as a *ranking* of evidence. They repeatedly argue that study-designs are chosen according to appropriateness for purpose, and are concerned with cataloguing potential sources of invalidity, rather than making recommendations about the weighting and appraisal of evidence. Campbell & Stanley's work is not focussed on *evidence* but rather *methodology*—its target audience is researchers, not consumers of research. They never produce a ranking; although the study-designs are numbered, this is for convenience, not any kind of ordering ([155], p.8]). In their concluding remarks, they explicitly caution against interpreting their table as a definitive ranking or analysis, and against condensing complex methodological issues into simplified tables:

> *"caution is needed about the tendency to use speciously convenient tables (...) the placing of specific pluses and minuses and question marks has been continually equivocal and usually an inadequate summary of the corresponding discussion"* ([155], p.71)

The first true hierarchy was presented in 1979 by the Canadian Task Force on the Periodic Health Examination [31] (henceforth, CTF), which included EBM pioneer David Sackett. CTF resembles several modern EBM hierarchies (especially [6,28,185,309])—a three/four-tier ranking, emulating the British degree-classification system: I, II-1, II-2 and III (see Figure 25, below). CTF's 1979 hierarchy evaluates the evidence base for the hypothesis that an intervention is effective. CTF is a hierarchy of *evidence bases*, using 'level of evidence' as its interpretation property. Rank I requires *"evidence obtained from at least one properly randomized controlled trial"* ([31], p.1195), while II-1 requires case-control or cohort studies, II-2 historically controlled trials, and III expert opinion and clinical experience.

This is the first instance of the ranking which forms the backbone of many subsequent EBM hierarchies—RCTs first, observational studies next, and expert opinion last. However, the scope is still limited—physiological and pharmacological studies, and mechanistic reasoning are *not* included. It is unclear whether absolute interpretations are endorsed—no specific levels of evidence are provided, so users would have to contribute their own in order to derive an absolute interpretation. Yet there is a hint towards absolute interpretation in the subsequent "*classification of recommendations*", which has the levels 'good', 'fair' and 'poor' evidence for a recommendation. Conditions also apply. Level-I RCTs must be "*properly*" randomized. Level II-1 cohort studies must be "*well-designed*" and "*preferably from more than one centre*". For uncontrolled studies to count as Level II-2 evidence, there must be "*Dramatic results*" ([31], p.1195). CTF gives an account of evidence bases for effectiveness claims, focused on level of evidence, with some conditionality and a fairly narrow scope addressing primarily epidemiological evidence.



Effectiveness of intervention

The effectiveness of intervention was graded according to the quality of the evidence obtained, as follows:

I: Evidence obtained from at least one properly randomized controlled trial.

II-1: Evidence obtained from well designed cohort or case–control analytic studies, preferably from more than one centre or research group.

II-2: 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.

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

Figure 25: CTF's hierarchy of evidence bases, the original hierarchy of evidence published in 1979 ([31], p.1195).

Subsequently, hierarchies became part of 'clinical epidemiology' in the 1980s.[36] Clinical epidemiology is the attempt to use population-level studies and study-designs from epidemiology to meet the needs of clinical practice (see e.g. [94,95]).[37] Foundational texts included *Clinical Epidemiology* [94], and Sackett et al.'s subsequent *Clinical Epidemiology: A Basic Science for Clinical Medicine* [95]. Although neither textbook included hierarchies, the field inherited the appraisal of research evidence from Campbell & Stanley [155], subsequent work by Cook & Campbell [156,220], and epidemiological work by the likes of Cochrane [325] and Feinstein [321-324].

Some clinical epidemiology sources codified this inheritance into hierarchies (e.g. [207,282,309,326]). Unlike CTF, these hierarchies were designed to inform the appraisal of *epidemiological evidence* in clinical practice—they only ranked different epidemiological study-designs (RCTs, observational studies and case reports) against one another. Early examples include Sackett's anonymously published 1981 hierarchy (see Figure 5, above), which introduces strength as an

[36] In his study of the '*what, who and whither*' of clinical epidemiology [96], Sackett traces the roots of clinical epidemiology to John Paul in 1938 [320]. However, the field was truly established as an independent discipline in the 1980s, primarily by Sackett and colleagues working at McMaster University (see [73]). Much of the groundwork for clinical epidemiology was laid by Feinstein [321-324] and Cochrane [325], though neither primarily identified their work as clinical epidemiology [324].

[37] Notably, many of the 1980s clinical epidemiologists had previous medical training, in contrast to most traditional epidemiologists at the time [73]. As a result, the clinical epidemiological community began with considerable independence from epidemiology.

interpretation property, ranking RCTs, cohort studies, case-control studies and case series on a relative scale from “*strongest*” to “*weakest*” ([193], p.986). Sackett’s hierarchy explicitly ranks *methods* by *strength*. Moreover, he clearly intends his hierarchy to be given relative interpretations. He explicitly endorses superlative interpretations—“*strongest*”, “*weakest*”—despite the hierarchy being far from exhaustive. The most natural interpretation is a limited scope of the quantifier in the superlative interpretations: it seems reasonable to presume that Sackett means ‘strongest *epidemiological* method’.

A series of clinical epidemiological hierarchies was produced in the journal *CHEST*, beginning with Sackett’s 1986 hierarchy (see Figure 26, below) [288,298]. This hierarchy has been updated several times from 1992 to 2008 [151,199,285,287]. *CHEST* hierarchies introduced grades of recommendation (A to C) alongside the hierarchy. This reinforces the emphasis upon clinical application—evidence is appraised with a view to making clinical recommendations. The scope of the *CHEST* hierarchy is still limited to epidemiological evidence, but Sackett formulates an explicit hierarchy of *evidence bases*, not methodology. The intended interpretation property is unclear, though he uses the phrase “Level of Evidence”, and the direct correspondence with a grade of recommendation indicates that this would naturally serve as an interpretation property. No indication is given about whether absolute interpretations are permissible, or how to formulate these interpretations. Superlative language has been eliminated. Finally, Sackett introduces clear conditionality—his Level-I RCTs must (a) be large, (b) have clear-cut results, and (c) have low risk of error, while Level-II RCTs must have none of these properties. Users are left uncertain what to do with RCTs with satisfy some but not all of properties (a)-(c)—another matter for interpretation. Presumably, either such midway RCTs are ranked somewhere between I and II (and hence between Grade A and Grade B), or are ranked at Level II, and Sackett’s conditions are meant as a minimum standard.

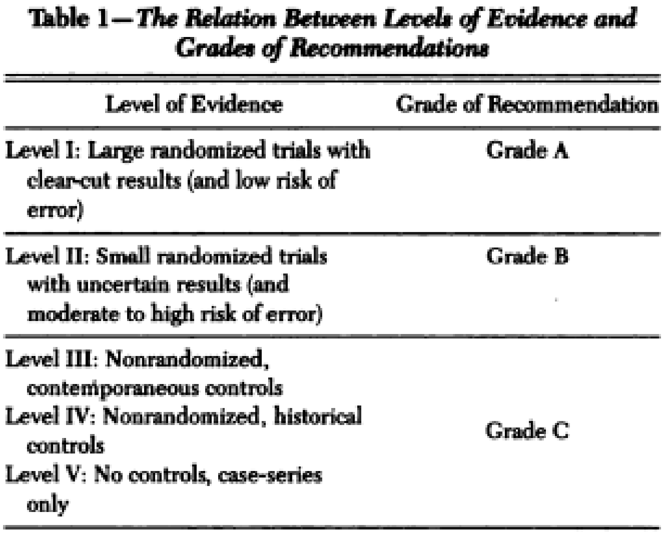


Table 1—*The Relation Between Levels of Evidence and Grades of Recommendations*

Level of Evidence	Grade of Recommendation
Level I: Large randomized trials with clear-cut results (and low risk of error)	Grade A
Level II: Small randomized trials with uncertain results (and moderate to high risk of error)	Grade B
Level III: Nonrandomized, contemporaneous controls	Grade C
Level IV: Nonrandomized, historical controls	
Level V: No controls, case-series only	

Figure 26: Clinical epidemiological hierarchy from Sackett, printed in *CHEST* in 1986 [298] and subsequently updated many times throughout the 1990s and 2000s [151,199,285,287].

Hierarchies are still defended by clinical epidemiologists independently of the EBM movement, generally following the same format—epidemiological study-designs in a relative ranking by level of evidence provided, and non-epidemiological evidence excluded (e.g. [34,142,327,328]). However, such hierarchies can be and have been influenced by the EBM movement's hierarchies.

The pre-EBM hierarchies, with the exception of CTF, all focused solely on the appraisal of epidemiological evidence bases for effectiveness claims and recommendations. CTF included expert opinion, but no hierarchy included mechanistic or laboratory evidence at that time. They have a narrow scope. Where superlative interpretations are endorsed, it is within this limited scope. None explicitly endorses absolute scales, though few explicitly exclude them. Interpretations are commonly made in terms of the 'level of evidence' studies provide, in-keeping with the motivation of assessing overall recommendations. Strength of recommendation is introduced as a secondary interpretation property. Broadly speaking, the pre-EBM hierarchies endorse a theory of *epidemiological* evidence, which claims that where there are positive RCTs, there is a higher level of epidemiological evidence for a treatment recommendation than when only observational studies are present. Modified interpretations are consistent with pre-EBM hierarchies, so one could equally adopt a more nuanced reading, e.g. evidence bases which included RCTs *usually* provide a higher level of evidence than solely non-randomised studies.

3.1.2: Early EBM—from Checklists to Hierarchies (c.1991-1998)

Evidence-Based Medicine grew from clinical epidemiology [66]. Many early proponents and innovators, such as David Sackett and Gordon Guyatt, were clinical epidemiologists at McMaster University, as well as practitioners and medical teachers. Similarly, the Cochrane Collaboration developed from clinical epidemiology—Chalmers and Enkin's intention was to facilitate the use of epidemiological evidence in clinical practice (for both patients and doctors) by creating summaries of the extensive and growing literature[38] [330]. Daly observes that EBM was in some respects a re-branding of clinical epidemiology ([73], pp.88-91).[39]

The term 'Evidence-Based Medicine' was coined by Guyatt in 1991[40] [64], and unveiled publically by the EBM Working Group in 1992 [39]. As the title of the Working Group's original publication indicates ('*Evidence-Based Medicine: A new approach to teaching the practice of medicine*'), the label 'Evidence-Based Medicine' was originally intended to apply to a new pedagogic approach focusing upon critical appraisal of the medical literature and independent problem-based learning, rather than absorption of inherited wisdom from experienced clinicians [39,332][41]. The movement quickly expanded its horizons to reforming clinical practice itself, not just medical teaching. The *Users' Guides* series, beginning in 1993 [334], attempted to train practising clinicians in critical appraisal as well as students [19], allowing them to find, appraise and apply epidemiological evidence in their practice.

A central purpose of the early EBM movement, then, was to try to change the behaviour of individual clinicians—evidence-based practitioners would find and appraise evidence to answer any clinical questions which arose in their practice (see [22,41,215,335,336]). In order to perform critical appraisal, clinicians needed guidance in appraising evidence from the various kinds of study available [334]. Appraising evidence is a complicated task—but the EBM movement needed appraisal techniques simple enough to be widely understood and applied quickly and easily in the limited time available to busy clinicians.

[38] The idea originates with Archie Cochrane [329], for whom the Collaboration is named.

[39] Clinical epidemiology suffered from resentment from both epidemiologists and clinicians—epidemiologists resented the implication that medical training was required to make epidemiological research 'useful', while some clinicians resented the idea that their clinical experience was subordinate to population-level studies, and saw the remit of epidemiology as limited to public health ([73], chs.2,5).

[40] Eddy [44] disputes priority for a 1990 paper [331] which discusses evidence-based inferences in medicine.

[41] EBM pedagogy is not to be confused with evidence-based pedagogy, a more recent movement to base teaching practice (across all disciplines) upon evidence from education studies—in recent years, an interest of the Campbell Collaboration, an evidence-based policy review group based on the model of the Cochrane Collaboration (e.g. [333]).

Their first approach to this problem was a checklist originating in Cook et al.'s 1992 update of the *CHEST* 'Rules of Evidence' [151]. This checklist delineates several features of a study which a clinician can check for—if these features are present, she can be confident that the evidence from that study is valid. The checklist took on a prominent role both in the early *Users' Guides* [19] and in Sackett et al.'s *Evidence-Based Medicine* [22], the movement's first textbook (see Figure 27).

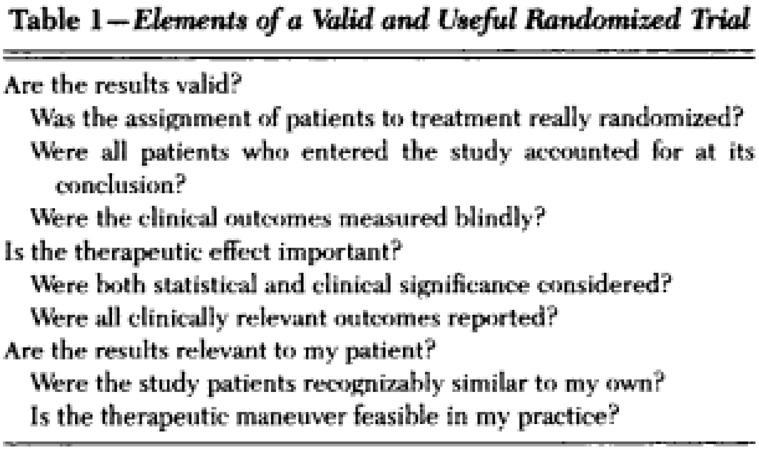


Table 1—*Elements of a Valid and Useful Randomized Trial*

Are the results valid?
  Was the assignment of patients to treatment really randomized?
  Were all patients who entered the study accounted for at its conclusion?
  Were the clinical outcomes measured blindly?
Is the therapeutic effect important?
  Were both statistical and clinical significance considered?
  Were all clinically relevant outcomes reported?
Are the results relevant to my patient?
  Were the study patients recognizably similar to my own?
  Is the therapeutic maneuver feasible in my practice?

Figure 27: EBM Checklist from Cook et al.'s ([151], p.306) update of the *CHEST* hierarchy, which was highly influential in the early *Users' Guides* [19] and the *Evidence-Based Medicine* [22] textbook.

The checklist makes three features of study-designs particularly significant—randomization, double-blinding and ITT-analysis[42]. Checklists, however, provide no information about other study-designs. The features highlighted by the checklist are supposed to guarantee the validity of a study—but it is left ambiguous whether studies which do not possess these features can be valid. In other words, the checklist features are supposedly jointly sufficient for validity, but it is unclear whether they are individually *necessary* for validity. Sackett et al.'s *Evidence-Based Medicine* seems to see randomization as necessary for validity—the textbook infamously states:

> *"If the study wasn't randomised, we suggest that you stop reading it and go on to the next article in your search."* ([22], p.94)

This advice was stated more dogmatically in later editions from 2000 onwards:

[42] ITT (Intention-To-Treat) analysis requires that all patients who entered the trial are included in the final analysis, despite any attrition (dropping out) during the study [246]. If we assume that patients who receive less benefit or greater side-effects from a treatment may be more likely to drop out, then analysis which only takes into account patients who completed the full course of the trial will tend to overstate the effectiveness of a treatment. Hence, ITT-analysis is promoted by the EBM movement to control for such 'attrition biases' [337,338].

> *"We can begin to rapidly critically appraise articles by scanning the abstract to determine if the study is randomised; if it isn't we can bin it."* ([10], p.118)

Although later editions, and indeed the later EBM movement, made and continue to make such absolute assertions of the necessity of randomization [339], the former quotation has sometimes been taken out of context by critics. In the first edition of *Evidence-Based Medicine*, the advice to "*stop reading it and go on*" comes in a section entitled "*Is the evidence from this randomized trial valid?*" ([22], p.94) The original advice occurs in the context of discovering that a purported RCT was not properly randomized: "*Was the assignment of patients to treatment really randomized?*" ([151], p.306). On the same page, Sackett et al. concede that evidence from nonrandomized studies can provide valid evidence (i.e. randomization is not necessary for validity)—when the treatment effect revealed is *"so huge that you can't imagine it could be a false-positive"* ([22], p.94)[43], or when the nonrandomized study concludes that the treatment was ineffective or harmful, as most biases in observational studies skew the results in favour of the efficacy of the experimental treatment. This is a precursor of some of the upgrading criteria of GRADE [202] (see Section 3.1.4, below).

The early EBM movement's checklists were ambiguous and provided only a very limited guide to the appraisal of evidence. The non-exhaustiveness of the checklist posed a serious issue if EBM was to provide a general method for clinical practice. The checklist approach left much to be desired, and several authors attempted to give more systematic and useful accounts of critical appraisal.

During this experimental mid-1990s period, various attempts were made to use hierarchies to guide critical appraisal. Early attempts were quite dissimilar to the precedents in clinical epidemiology. The first hierarchy within the EBM canon is Guyatt et al.'s *Users' Guides IX* in 1995 [190]. This hierarchy looks very different to those defended before or since (see Figure 28, below). The hierarchy includes only RCTs and observational studies, and evaluates the *grade of recommendation*, in part according to the consensus (or lack thereof) between the available studies. In this sense, it is most reminiscent of CTF's hierarchy, albeit with narrower scope. Guyatt et al.'s hierarchy is a clear hierarchy of evidence bases, using strength of recommendation as its interpretation property. Whether absolute interpretations are permissible is unclear—in this respect, it resembles Sackett's 1986 version [288].

[43] In this case, the authors note, it would be unethical to perform a randomized trial ([22], p.94)—for further discussion of the ethical issues where dramatic treatment effects are concerned, see Worrall's discussion of the ECMO case [187].

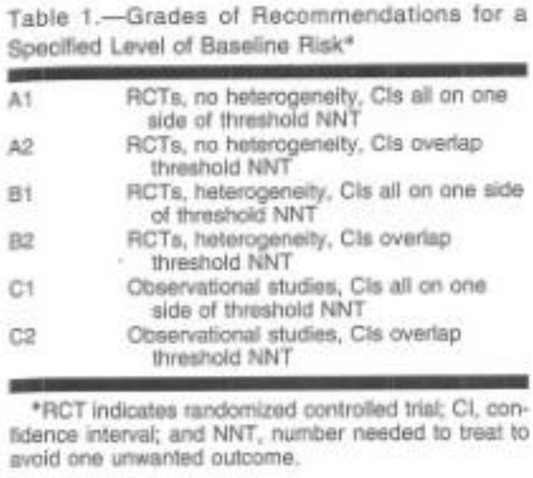


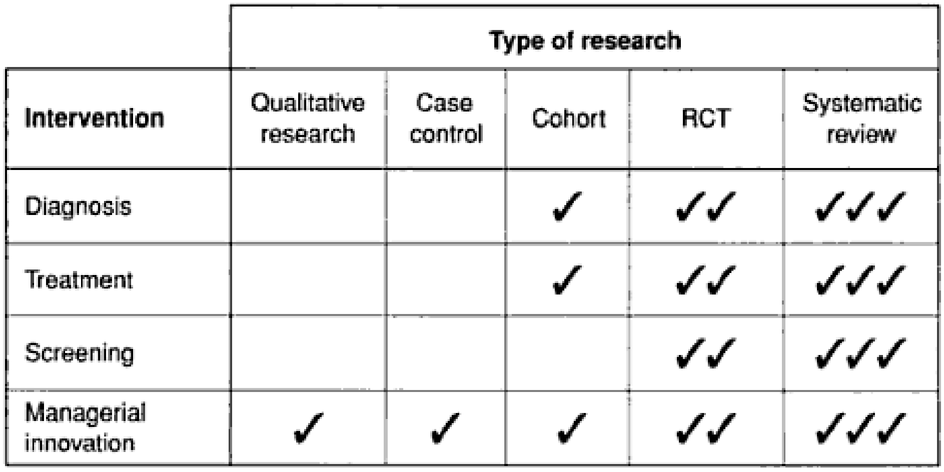
Table 1.—Grades of Recommendations for a Specified Level of Baseline Risk*

A1	RCTs, no heterogeneity, CIs all on one side of threshold NNT
A2	RCTs, no heterogeneity, CIs overlap threshold NNT
B1	RCTs, heterogeneity, CIs all on one side of threshold NNT
B2	RCTs, heterogeneity, CIs overlap threshold NNT
C1	Observational studies, CIs all on one side of threshold NNT
C2	Observational studies, CIs overlap threshold NNT

*RCT indicates randomized controlled trial; CI, confidence interval; and NNT, number needed to treat to avoid one unwanted outcome.

Figure 28: Guyatt et al.'s *Users' Guides IX* hierarchy [190], the first application of a hierarchy within the EBM movement.

In contrast, Muir Gray's 1997 textbook *Evidence-Based Healthcare* [340] introduced a variant which represents the quality of research for certain purposes by numbers of ticks, ranking systematic reviews at three-ticks, RCTs at two, and observational studies at one or none (see Figure 29, below). Muir Gray's hierarchy is perhaps most naturally interpreted as a hierarchy of methodology, but the interpretation property, and how to interpret the ticks in terms of that property, are left entirely up to the user.



	Type of research				
Intervention	Qualitative research	Case control	Cohort	RCT	Systematic review
Diagnosis			✓	✓✓	✓✓✓
Treatment			✓	✓✓	✓✓✓
Screening				✓✓	✓✓✓
Managerial innovation	✓	✓	✓	✓✓	✓✓✓

Figure 29: Muir Gray's variant upon evidence grading using tick-boxes ([340], p.122)

Perhaps the most influential introduction of a novel hierarchy into the canonical Evidence-Based Medicine literature is in Trish Greenhalgh's *How to Read a Paper* article series and book in 1997 (see Figure 19, above) [43,296]. Unlike most previous hierarchies, Greenhalgh's critical appraisal technique advocates appraisal of individual studies in isolation, according to the ranking in the hierarchy—Greenhalgh describes her hierarchy as *"Standard notation for the relative weight carried by the different types of primary study"* ([296], p.246). In fact, her hierarchy is unique in its combination of a restriction to clinical epidemiology study-designs with a simple list ranking and clear restriction to *relative* interpretations. Despite stating that the scope of the hierarchy is restricted to "*types of primary study*", systematic reviews and meta-analyses are included at Level 1. Greenhalgh introduces slight conditionality—RCT-evidence must be 'definitive' to access Level 2. Greenhalgh's hierarchy leaves the interpretation property open, but she suggests the:

> *"weight carried by the different types of primary study when making decisions about clinical interventions"* ([296], p.246).

Several similar hierarchies were subsequently adopted by EBM proponents, most notably in 1999's *Users Guides' XIX* [216].

In summary, the early period in the development of EBM hierarchies, lasting roughly from 1991 until 1998, was a period of experimentation. Hierarchies were introduced as one of a number of competing systems for appraising evidence. These systems needed to be simple and easy to use, allowing practising clinicians to evaluate evidence on the fly. As in clinical epidemiology, hierarchies were restricted to epidemiological evidence sources, ranking RCTs against observational studies and case reports. Almost no hierarchies in this period offered an absolute scale of evidence quality, suggesting only a relative ranking. No superlative language is used. Proponents experimented with a range of interpretation properties, and hierarchies of evidence were introduced in addition to evidence-bases. The use of conditions was common, especially to divide RCT evidence into sub-strata.

Overall, during this experimental period, the approach to evidence from the EBM movement is quite modest—under certain conditions, RCT evidence is stronger, or should be given greater weight, or provides a greater level of evidence for effectiveness claims, than does evidence from non-randomized epidemiological studies.

3.1.3: Later EBM—Hierarchies at the Centre (c.1998-2005)

In the late-1990s and early-2000s, some members of the EBM movement began advocating a different agenda. The movement's focus broadened and many proponents' aims became more radical. Some abandoned or sidelined (in some cases, reluctantly) targeting individual clinicians in favour of an attempt to reform medical practice entirely. Gordon Guyatt in particular became convinced that the attempt to give individual clinicians the tools to appraise evidence for themselves was inadequate [85,341].[44] There were two primary problems with the programme of clinician education: (1) the appraisal of clinical evidence is much more complicated and time-consuming than most clinicians have the time or training to perform; and, (2) many clinicians would fail to implement what they had learned systemically, or did not seek out the necessary training [341,344].[45] The model of 'passive diffusion' [345] of evidence and critical appraisal training—making the evidence and the training available and hoping clinicians would use these resources—was insufficient to make the changes to clinical practice envisioned by many within the movement.

Guyatt and colleagues responded by emphasising an intermediate stage between researchers and practitioners—professional critical appraisal. In an interview with Jeanne Daly in 2005, Guyatt said:

> *"I thought we were going to turn people into evidence-based practitioners (…) I no longer believe that. What I believe now is that there will be a minority of people who will be evidence-based practitioners, and that the other folk will be evidence users (…) they are not actually expected to read and understand the articles and really be able to dissect the methodology."* ([73], p.91)

Guyatt and others began strongly advocating various systems by which trained and dedicated critical appraisers could perform systematic appraisals of research evidence and disseminate their work to practitioners. The model for these evidence-appraisals was the Cochrane Collaboration. Moreover, a number of additional resources were developed and promoted, including summary journals such as the *ACP Journal Club* [100]. Students created 'CATs' (Critically Appraised Topics), systematic critical appraisals of individual studies (see [41,98,188]). Databases of peer-reviewed CATs provided another

[44] One contributory factor to this change in perspective among many in the EBM movement may have been the retirement of David Sackett in 2000 [111]. Sackett was amongst the most vocal and committed advocates of training clinicians—his publications and lecture-tours were aimed at teaching clinicians to appraise evidence and convincing them of the importance of applying the evidence themselves (see e.g. [20,41,78,336]). Although Sackett took a back-seat in EBM publications since 2000, he continued to target individual clinicians with his writings and editorials [180], especially in his "Clinician-trialist rounds" series (e.g. [342,343]).

[45] The problems are in fact mutually-reinforcing. The difficulties in practical application make clinicians less likely to apply what they have learnt or to try to learn to appraise evidence. Meanwhile, reluctance to make critical appraisal a regular part of practice makes it difficult to develop the experience and familiarity necessary to appraise evidence quickly and thoroughly.

potential resource [10,188]. Another proposal, BETs (Best Evidence Topics) [99,301], were similarly-produced overviews of all the available studies on a particular question (e.g. the effectiveness of a particular treatment for a particular presentation of a condition, or a particular diagnostic test). CAT-making could be both an educational tool, and a potential resource for others.

Importantly, if pre-appraised evidence was widely available, clinicians would no longer be justified in ignoring the research literature due to inability to appraise evidence or lack of time to perform detailed appraisals. This allowed the movement's rhetoric to intensify—clinicians have a duty to pay attention to the best evidence wherever they can, and with pre-appraised evidence, all clinicians can access and apply the best evidence [346]. The movement was expanding from an alternative movement focused on reforming the individuals involved to a broad reformative movement, focused upon changing all medical practice [347].

This change had a number of effects on the hierarchies defended by EBM proponents who subscribed to this new model of 'active dissemination' [345]. First, a new kind of hierarchy was introduced—hierarchies of pre-appraised evidence sources (often called the '4S' [145] or '5S' hierarchy [146]—see Figure 3, above)[46]. These hierarchies ranked types of pre-appraised evidence according to their quality, so that clinicians could locate the most reliable sources. The top-ranked evidence source is usually 'Systems' such as computerised decision support systems. The pyramidal ranking may suggest absolute interpretations (though no scale is given to help users construct these). There are no conditions and no suggestion of modification. The hierarchies are not, in fact, limited solely to pre-appraised evidence, as the lowest rung is always "studies" themselves. Presumably, the argument for ranking, say, a synopsis of a study (e.g. a CAT) above the study itself is that the synopsis-maker will identify any problems with the study which may be unclear from the study-report itself. This claim is problematic where untrained evidence-users create synopses, or when skilled appraisers draw on evidence from studies. Hierarchies like 4S rely upon the assumption that professional evidence appraisers can identify problems which ordinary evidence-users cannot.

The second major effect was the introduction of a new role for hierarchies. Producing systematic reviews, summaries and synopses requires a number of judgements about the quality of the studies at hand (e.g. which studies to include in a systematic review, and how to weight different studies). Clinicians must trust the judgements made by the producers of the pre-appraised evidence. One way to achieve reliability is to objectivise the appraisal process. Hierarchies offered a potential solution. If appraisers all use the same hierarchy, then everyone's appraisals should be similar and

[46] So-called because the 4 or 5 rungs of the hierarchy all begin with the letter S.

standardised. This should ensure that systematic reviews and synopses are reproducible and consistent.

Many institutions responsible for producing systematic reviews, guidelines and policies adopted or created a hierarchy to objectivise the selection and appraisal of evidence in their products during this period. Institutions introducing hierarchies between 1999 and 2005 included the National Institute of Health and Clinical Excellence (NICE) in the UK [2,284], the U.S. Preventive Services Task Force [23,24], the Scottish Intercollegiate Guidelines Network (SIGN) [133], the Institute for Clinical Systems Improvement (ICSI) [34], and the Australian National Health and Medical Research Council (NHMRC) [6,33]. These hierarchies use evidence *quality* or level of evidence as the interpretation property, and usually endorse *absolute* interpretations—such absolute interpretations are needed for the purposes of the institutions, which need to know how good each piece of evidence is, not just which sources are comparatively better than others.

The expansion in focus from training clinicians in critical appraisal to providing pre-appraised evidence for clinicians allowed many within the movement to make more radical claims—that *all clinicians* must use epidemiological evidence in their practice, not just those with the requisite training. Evidence-based guidelines and evidence summaries mean that all practitioners can access and use the relevant evidence. There are even suggestions that such guidelines may take on legal force in medical negligence cases [348].

A further development, accompanying the shift and expansion in focus, was an increase in the scope of the change advocated by many within the movement. While early texts had focused upon the *inclusion* of epidemiological evidence in clinical practice [19,20,22,39,78], increasingly proponents argued for the pre-eminence of epidemiological evidence (and especially RCT evidence) [8,10,82]. The rhetoric marginalised talk of inclusion and integration of different evidence-sources—clinical experience, biological studies and rationale, and epidemiology—and began arguing that epidemiological evidence outweighed non-epidemiological sources. Not only did RCTs provide the highest-quality *epidemiological* evidence, they provided the highest-quality evidence of any kind.

This development was clearly mirrored in the hierarchies produced in the period. For the first time in the EBM movement, hierarchies introduced non-epidemiological evidence in their rankings—invariably at the bottom levels. Hierarchies introduced clinical experience and expertise, “biologic rationale”, pathological theory and laboratory studies, placed below the lowest-ranked epidemiological studies. This may seem a minor alteration, especially in presentational terms, but actually represents a significant change in the intended interpretation of the hierarchies. Where superlative interpretations were defended, no longer was their claim that the highest-ranked study-design provided the highest-quality *epidemiological* evidence—rather, it provided the highest-quality *evidence*. Rather than using epidemiological evidence to supplement clinical experience and biological theory (or vice versa), the hierarchies now directed clinicians to use epidemiological evidence, while clinical experience and biologic rationale are only important when epidemiological evidence is absent. The first EBM hierarchy to include non-epidemiological evidence at the lowest level was Phillips et al.'s hierarchy in 1998 [130] (see Figure 30, below), which became the *CEBM* hierarchy [103] (see Figure 15, above), whose lowest level is reserved for "*Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles".*"

Grade of recommendation	Level of Evidence	Therapy: Whether a treatment is efficacious/ effective/harmful	Therapy: Whether a drug is superior to another drug in its same class
A	1a	SR (with homogeneity*) of RCTs	SR (with homogeneity**) of head-to-head RCTs
	1b	Individual RCT (with narrow Confidence Interval‡)	Within a head-to-head RCT with clinically important outcomes
	1c	All or none§	
B	2a	SR (with homogeneity*) of cohort studies	Within a head-to-head RCT with validated surrogate outcomes ‡‡‡
	2b	Individual cohort study (including low quality RCT; e.g., <80% follow-up)	Across RCTs of different drugs v. placebo in similar or different patients with clinically important or validated surrogate outcomes
	2c	"Outcomes" Research; Ecological studies	
	3a	SR (with homogeneity*) of case-control studies	Across subgroup analyses from RCTs of different drugs v. placebo in similar or different patients, with clinically important or validated surrogate outcome
	3b	Individual Case-Control Study	Across RCTs of different drugs v. placebo in similar or different patients but with unvalidated surrogate outcomes
C	4	Case-series (and poor quality cohort and case-control studies§§)	Between non-randomised studies (observational studies and administrative database research) with clinically important outcomes
D	5	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"; or non-randomised studies with unvalidated surrogate outcomes

Figure 30: Centre for Evidence-Based Medicine (CEBM) hierarchy from Ball & Phillips [131], a 2002 update of Phillips et al.'s 1998 version [130] with only cosmetic changes.

This move towards including non-epidemiological evidence at the lowest level quickly became standard in the EBM literature, with most hierarchies since 1998 following the practice. The change was solidified by prominent placement in the 2nd edition of *Evidence-Based Medicine* [135], *Users' Guides XXV* [82] and the *Users' Guides* book [83]. Most of the hierarchies which adopted this practice still only explicitly endorsed relative interpretations, though it was easy to read off an absolute interpretation. The property in question was evidence quality. Conditions are minor where present, and there is never any discussion of modification. The institutional hierarchies discussed above also took on these interpretative assumptions, but in combination with the need for clear-cut absolute scales of evidence quality (e.g. [2,23,24,27,134,305]).

A final development during this period was the increasing emphasis placed upon the hierarchies. The first edition of *Evidence-Based Medicine* [22] in 1997 made no mention of hierarchies, nor did the "*five linked ideas*" of Evidence-Based Medicine [21]. By contrast, Guyatt et al.'s ([83], p.10) *Users' Guides* book of 2002 made the hierarchy the first of two "*fundamental principles of EBM*", and the 2nd edition of *Evidence-Based Medicine* [135] placed its hierarchy at the heart of critical appraisal, the most significant section of the book.

In summary, during the late 1990s and early 2000s, many within the EBM movement changed their goals and ideas. Spearheaded by Guyatt, many began emphasizing provision of pre-appraised evidence to clinicians above training individual clinicians in critical appraisal. Hierarchies were no longer only pedagogic tools to train clinicians to understand epidemiological research, but became a ranking of *all* medical evidence, according to epidemiological precepts.

3.1.4: Contemporary EBM—GRADE and the New Hierarchies (c.2004—)

Since hierarchies of evidence took centre-stage in the EBM movement, and took a new role in the generation of pre-appraised evidence, interest in the details of hierarchies increased within the movement. Producing reviews and synopses according to rigid hierarchies introduces several problems. Experienced evidence-appraisers are constrained by the simple equation of evidence-quality with study-design. Many proponents and producers of pre-appraised evidence wanted to involve more factors than study-design in their appraisal—for instance, evidence of a dose-response gradient [202], and how representative the study-population is of the target-population [213].[47] While such considerations could be described in discussion sections of their reviews, there is no guarantee that such concerns would be read, understood and appreciated by their audience of busy practitioners, especially if EBM scaled back its attempts to educate clinicians in evidence-appraisal. Meanwhile, these considerations could not be explicitly included in the analysis without forgoing the standardisation and 'objectivity' provided by a shared hierarchical method. As such, there was a clear impetus to develop a more sophisticated hierarchy which would still standardise procedures, but allow for factors other than study-design to influence the appraisal of evidence.

From 2004 onwards, several attempts were made to defend more sophisticated forms of hierarchy. Most prominent amongst these attempts are the GRADE (Grading of Recommendations Assessment, Development and Evaluation) hierarchies produced by the GRADE Working Group (see Figure 9, above), first in 2004 [12,45], and then in major article series in 2009 [13,297] and 2011-2 (e.g. [14,195,350]). The GRADE system is explained in detail in Section 2.1.3, above, and examined in Chapter 6.

La Caze views GRADE as exemplifying a "*non-categorical*" interpretation [72]. As argued in Chapter 2, a more perspicuous way to understand GRADE is as a heavily conditional ranking. The interpretation property is explicitly "*quality*", which is translated into a strength of recommendation when combined with considerations of the risk/benefit ratio demonstrated by the evidence. The interpretations endorsed are unequivocally *absolute* (High, moderate, low or very low quality; strong or weak recommendation). Notably, though, GRADE hierarchies have removed non-epidemiological evidence from the ranking, reversing the step taken in the early-2000s. No GRADE hierarchy yet explicitly includes non-epidemiological evidence, though there is no reason why it could not be

[47] There are many alternative appraisal systems which take such factors into account, most using extensive checklists rather than hierarchical structures (see [152,349] for overviews). Perhaps the best-known is CONSORT (CONsolidated Standards Of Reporting Trials) [154], a 25-part checklist for reports of RCTs. Although this standard does not explicitly evaluate the quality of evidence (rather, the quality of reporting), it can be used as a systematic basis for synopses or summaries. Moreover, while CONSORT is designed only to apply to RCTs, there are extensions to other study-designs (e.g. [153]), and similar programmes tailored to those designs (e.g. STROBE—Strengthening The Reporting of OBservational studies in Epidemiology [313]).

included.[48] The omission of both non-epidemiological and pre-appraised evidence sources suggests that GRADE hierarchies are not meant to be given superlative interpretations. GRADE could simply be interpreted as having a narrower scope, more reminiscent of early hierarchies [195]. This is not necessarily a concession to proponents of non-epidemiological evidence—after all, GRADE is used to systematise which evidence is used in systematic reviews and meta-analyses; excluding non-epidemiological evidence may imply that it should be omitted entirely. In practice, GRADE might be read as including a further *inclusion condition*: only epidemiological evidence counts at all.

It has been claimed that GRADE is now the only relevant hierarchy, and is the one that has been adopted wholesale by EBM. For instance, Guyatt et al. talk about "*An emerging consensus on grading recommendations*" [48,351], and the World Health Organization (WHO) called GRADE:

> "*internationally agreed standards for making transparent recommendations*"
>
> ([4], p.37).

Goldet & Howick claim that GRADE is "*emerging as the dominant method for appraising controlled studies*" ([200], p.50). According to this picture, the history of hierarchies is one of evolution towards the GRADE standard, which was adopted in 2004. Since then, the process is one of refinement of that standard.

However, GRADE is not the consensus position it has been portrayed as. The range of groups who have adopted GRADE do not use it consistently. As a study by Alexander et al. [206] showed, organizations like the WHO [4] have adopted GRADE but rarely use it within their guidelines. Others have retained other hierarchies or developed new ones since the advent of GRADE, including NICE [2,27,28], the Australian NHMRC [7,352,353], SIGN [134], the US Preventive Services Task Force [5] and the Canadian Task Force on the Periodic Health Examination [32][49]. Prominent hierarchies such as Guyatt et al.'s *Users Guides* hierarchies [83], the *CHEST* hierarchies [35,199] and the *CEBM* hierarchies [15,103] are still updated, published and used. Moreover, while GRADE is influential in health policy

[48] A number of GRADE hierarchies include "Other evidence" catch-all categories [12,37,46,278], although whether these are meant to be read as including non-epidemiological sources or just epidemiological evidence not captured by the labels 'RCT' or 'observational study' is unclear.

[49] Several of these institutions were or are part of the GRADE Working Group [12], and some have technically adopted GRADE or used it for some publications, but continue to use alternative ranking systems. For instance, although NICE's guidelines manual formally adopted GRADE in 2009 [3], individual guidelines such as the 2011 guidelines on Tuberculosis [27] continue to use other systems, and some used mixed approaches such as the 2009 and 2015 guidance on borderline personality disorder [354] which uses SIGN to assess quality of evidence, then GRADE to assess recommendations. The same has been true of the WHO (see [281]). A 2014 study showed that only 37% of WHO guidelines use GRADE, and that over 55% of those made strong recommendations based on evidence which GRADE assessed as low or very low quality [206]. The Australian NHMRC has offered several alternative hierarchies (e.g. [6,33,352,353,355]), and their latest guidance allows guideline developers to choose either GRADE or an NHMRC approved hierarchy to appraise evidence and recommendations (see [7]). Several sources also make changes to GRADE beyond those endorsed by the GRADE Working Group [14], including NICE [3] and the American Academy of Neurology [36].

and guideline development, practitioners can still access and use other hierarchies—indeed the latest *CEBM* hierarchy specifically targets practitioners, while recommending GRADE for use by guideline developers [15,16].

Similar hierarchies inspired by or related to GRADE have also proliferated in recent years (e.g. [37,46,278]). In particular, several hierarchies use traditional list designs, but have multiple designs per level, and split the ranking of some study-designs according to the 'quality' of the study (e.g. [132,356]). For example, Eaves' 2011 hierarchy [132] (see Figure 31) ranks "*high quality*" RCTs "*with adequate power*" as Level 1, with "*Lesser quality*" RCTs *and* cohort studies both achieving Level 2. One might worry about the leeway for unsystematic judgements in terms such as "high" and "lesser quality"—however, such concerns could be ameliorated by offering GRADE-like or CONSORT-like criteria. Eaves also offers a novel presentation of the hierarchy on a scale from "validity" to "bias". Other developments include broadening the range of study-designs included in hierarchies (e.g. [208,299,357]) and greater detail in the rankings (e.g. [15,103]).

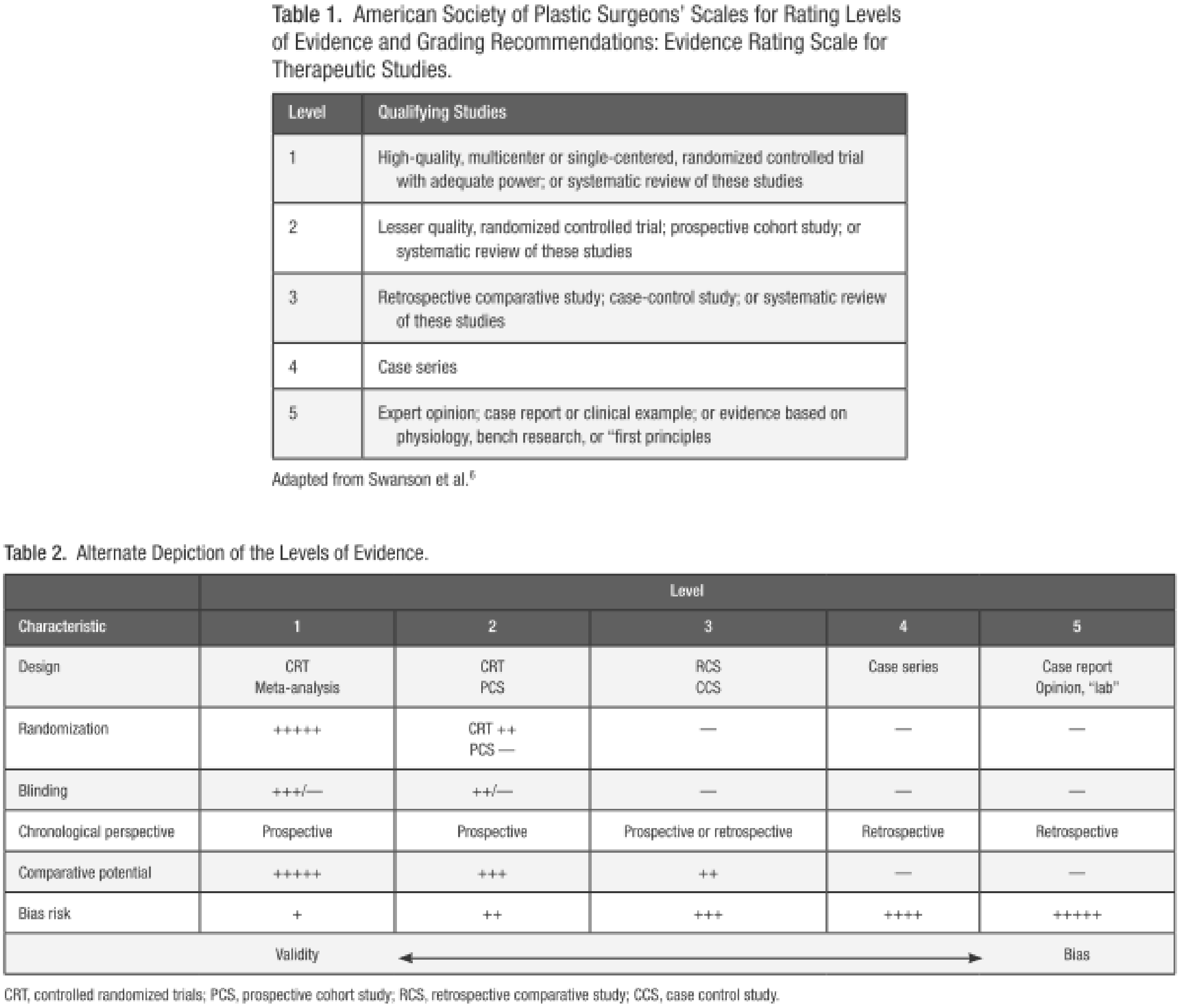


Table 1. American Society of Plastic Surgeons' Scales for Rating Levels of Evidence and Grading Recommendations: Evidence Rating Scale for Therapeutic Studies.

Level	Qualifying Studies
1	High-quality, multicenter or single-centered, randomized controlled trial with adequate power; or systematic review of these studies
2	Lesser quality, randomized controlled trial; prospective cohort study; or systematic review of these studies
3	Retrospective comparative study; case-control study; or systematic review of these studies
4	Case series
5	Expert opinion; case report or clinical example; or evidence based on physiology, bench research, or "first principles

Adapted from Swanson et al.[6]

Table 2. Alternate Depiction of the Levels of Evidence.

	Level				
Characteristic	1	2	3	4	5
Design	CRT Meta-analysis	CRT PCS	RCS CCS	Case series	Case report Opinion, "lab"
Randomization	+++++	CRT ++ PCS —	—	—	—
Blinding	+++/—	++/—	—	—	—
Chronological perspective	Prospective	Prospective	Prospective or retrospective	Retrospective	Retrospective
Comparative potential	+++++	+++	++	—	—
Bias risk	+	++	+++	++++	+++++
	Validity	←		→	Bias

CRT, controlled randomized trials; PCS, prospective cohort study; RCS, retrospective comparative study; CCS, case control study.

Figure 31: Eaves' 2011 hierarchy, in both a standard tabular format, and a scale from validity to bias ([132], pp.719-20).

While a general trend towards sophistication in hierarchies has been evident since 2004, not all developments have been sophisticated or increased the complexity of hierarchies. One notable trend is in presentation—many hierarchies in the 2010s take pyramidal form (see Figure 32, below), with RCTs at the top of the pyramid (e.g. [42,189,256,289,290]). Far from a reflective sophistication, the change to a pyramid form seems confused, and illustrates the tendency to fail to justify structural choices in hierarchy design. No guidance is offered to the differences in interpretation between a pyramid and a table. The natural assumption is that the pyramid implies that the evidence becomes scarcer as we move up towards the pyramidion. But no evidence is presented that, for instance, RCTs are rarer than cohort studies, or cohort studies rarer than case-control studies.[50]

[50] At least one instance of this pyramidal ranking is patently false—Niederman & Richards' [208] pyramid ranks RCTs above CATs ('Critically Appraised Topics'), which are summary reports of RCTs, and thus must be scarcer

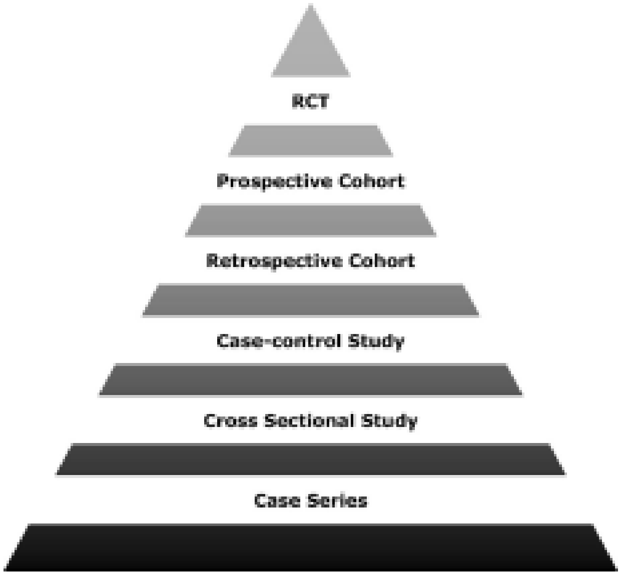


Figure 32: Example of an evidence pyramid from Ho, Peterson & Masoudi [256]

3.1.5: Discussion and Challenges

Several challenges might be made to the account developed here. First, it might be argued that the four stages delineated above are not clearly distinct periods in the history of hierarchies, and that there are exceptions to the trends highlighted during each period. There are a couple of hierarchies which include non-epidemiological evidence in the form of expert opinion at the lowest ranking during the 'pre-EBM' and 'early EBM' periods (e.g. [31,102,207]), and hierarchies which still rank only epidemiological designs during the 'later EBM' period (e.g. [6,23,216]).

Although this challenge is accurate, it does not undermine the account given here. First, only trends are presented here, and the dates given in this account begin from the onset of a trend. The trend within the EBM movement to include non-epidemiological sources in hierarchies began in 1998, but did not completely permeate the literature until at least 2000—the last hierarchy of the 'later' period to include only epidemiological evidence was Briss et al.'s in 2000 [23]. A trend in this sense

(at most, equal in number) than RCTs. This pyramid is prominently repeated in the 'EBM Page Generator' [289]. An alternative suggestion is that the pyramid represents a process of testing treatments: we begin with the lowest level test. Those which survive pass to the next level, and so on. Such an approach is reminiscent of the "Phase" system of therapeutic testing (those treatments which survive phase I pass to phase II, etc.). In this case, the number of treatments under scrutiny will fall as we climb the pyramid, as some treatments will fall at lower levels. But this is surely not the intention here: I take it that no EBM proponent believes that lower-level evidence must be provided before passing to the higher levels.

involves both the increase in frequency of the newer style and the decrease in frequency of the older style—sudden transitions are unlikely. Moreover, as discussed in Section 3.1.1, clinical epidemiology persists distinctly from the EBM movement [358]—hierarchies continue to be advocated which address solely the use of evidence *in clinical epidemiology*, rather than directly in medical practice (e.g. [35]). These hierarchies are not pertinent to the identification of trends within the EBM movement.

Similarly, those hierarchies which *did* include non-epidemiological evidence prior to the changes within the EBM movement in the late 1990s originate outside of the EBM movement. Of the three notable examples, one (CTF's 1979 hierarchy [31], see Section 3.1.1) considerably pre-dates the movement, while another (LaForce's 1987 hierarchy [207]) both pre-dates the movement and is explicitly limited to public health applications. The third, an anonymous editorial in *Bandolier* [102], makes clear reference to Evidence-Based Medicine, but seems primarily concerned with editorial policy for the journal not application in clinical practice[51]. The non-epidemiological source in all of these cases is expert opinion. Evidence from laboratory studies, clinical experience and mechanistic reasoning was not included in any hierarchy until 1998.

In summary, trends are identified, rather than sharply-delineated periods. Despite the overlap between the historical stages, there are major transitions in the interpretative assumptions made by many members of the EBM movement which are philosophically significant. The goal here is to identify coherent sets of interpretative assumptions, using an historical analysis to identify them.

A second challenge to this account comes from an alternative depiction of the EBM movement's development which has been offered by both critics and proponents of EBM—namely, that EBM began as a zealous revolutionary group (or idea), advocating a "*crude*" ([123], p.576) or "*hard-line*" [187] position (see also Section 3.2.1, below), and over time either mellowed and became more sophisticated (according to proponents [85,86,123]) or retreated and abandoned its untenable position in favour of vagueness and obfuscation (according to critics) (*cf.* [75,359,360]). By contrast, the account defended here holds that the EBM movement began with relatively modest aims (changing teaching and the practice of individual clinicians to *include* epidemiological evidence) and interpretations of hierarchies to match, and developed to a more radical position (changing the whole practice of medicine to be primarily dictated by epidemiological evidence).

This alternative account is generally supported with three claims: (1) that EBM originally presented itself as a paradigm shift in medicine [39], which are radical departures, and that the

[51] In particular, the *Bandolier* editorial hierarchy [102] seems to claim that the journal wishes to publish primarily results of RCTs, and that non-randomised studies will be subject to more stringent review.

movement later distanced itself from this radicalism (e.g. [83,86]); (2) that very radical claims about the importance of randomization were made in the early EBM texts—for instance, the injunction to go on to the next study if a study is found to be non-randomized in the 1997 edition of *Evidence-Based Medicine* [22]; and, (3) that the introduction of hierarchies into the process of critical appraisal represented a softening of their formerly 'hard-line' position in comparison to the early checklists (e.g. [19]), by admitting evidence other than RCTs into consideration. Each of these three claims is misguided and paints an unrepresentative picture of the development of the EBM movement.

First, the retreat from the 'paradigm' definition of EBM is actually primarily due to increased philosophical scrutiny, and the realisation that the term had been used without attention to its philosophical implications—such as the radical, incommensurable change a paradigm shift is supposed to involve (see Section 1.2.1). Moreover, the 'paradigm shift' originally proposed was a shift in *teaching* and *individual practices* [39]. It was not to be a new paradigm for the entire medical field, but more a new *option* for practice.[52] The 'new paradigm' in 1992 had a much narrower scope than the ethos for medicine defended in the early 2000s.

Second, the radical claims in the 1st edition of *Evidence-Based Medicine* have been overstated, as discussed above (see Section 3.1.2), and these claims were only embellished in later editions [10,339]. Moreover, contrary to the claim that the EBM movement softened its insistence upon randomization, most of the hierarchies which insist upon RCTs as the only worthwhile evidence in medicine were produced in the 'later' period—see for example van Tulder et al.'s [305] 2000 hierarchy (see Figure 20, above).

Third, and most importantly, the change from a checklist to a hierarchy was not primarily motivated by a new caution, nor was it a case of relenting on an earlier hard-line view by allowing non-randomized epidemiological studies into the fold. Rather, checklists suffered from ambiguity in interpretation, and failed to offer guidance to clinicians where RCTs were unavailable. The checklists instructed clinicians in the use of RCTs in addition to other more familiar sources. Introducing hierarchies served only to add *further epidemiological sources* to this repertoire. Note that, when they were first introduced, hierarchies included only epidemiological sources [190]—surely then, taking further epidemiological sources into consideration cannot be considered softening a hard-line position.

[52] This is as opposed to the new paradigm for the whole of medicine described more recently by Guyatt and colleagues ([86]; [73], pp.88-9).

This interpretation is understandable if EBM is understood as the promotion of *RCTs*, as opposed to the promotion of epidemiological evidence within clinical practice. However, conflating EBM with simple advocacy of RCTs is a mistake, certainly in the early EBM period; early defences of EBM all emphasise the importance of integrating *epidemiological* evidence [20,21,40,78]—RCTs are heralded as the epitome or pinnacle of these new sources [39,190]. Crucially, expanding hierarchies to include non-epidemiological evidence was not a weakening but a radical strengthening of the EBM line. Adding non-epidemiological evidence at the bottom level transformed non-epidemiological evidence from something separate to something subordinate.

The introduction of hierarchies into EBM publications was not the retreat from radicalism it has sometimes been deemed to be. Hierarchies pre-date the EBM movement, and had been used by numerous key figures in the movement, including Sackett [31,193,288], prior to their introduction into EBM. Furthermore, hierarchies were introduced during a period of experimentation in methods of critical appraisal, before any particular model had been widely accepted. Introducing a hierarchy which ranked epidemiological designs did not profoundly alter the views espoused within the movement. The crucial change came in the introduction of non-epidemiological evidence at the foot of the hierarchy, shifting the ideology from the integration of epidemiological evidence into clinical practice, to the pre-eminence of epidemiological evidence and RCTs over other kinds of evidence. This change took place from 1998.

3.2: Philosophical Interpretations

Section 3.1 showed that the interpretative assumptions made by authors of hierarchies have shifted over time. This is evident both from the surrounding literature and from the presentation of their hierarchies. Another source of potential clusters of interpretative assumptions is the philosophical literature which has emerged around EBM. Although these assumptions may not be made by *authors* of hierarchies, they may be made by *users*. This section identifies a number of philosophical interpretations of hierarchies, and reformulates them through the framework introduced in Chapter 2.

3.2.1: "Crude" and "Sophisticated" EBM

Brody, Miller & Bogdan-Lovis [123] distinguish between "*crude*" and "*sophisticated*" EBM. They argue that much of the damage done to the EBM movement originates from its "*friends*"—supporters taking a crude approach. These two positions have different consequences for interpretation of hierarchies.

The "crude EBM" position amounts to:

> *""decerebrate genuflection" at the altar of the RCT".* ([123], p.577, partially quoting [361]).

Crude EBM consists of the theses: (a) RCT evidence is the only high quality evidence for any kind of claim, and (b) RCT evidence is always high quality—there are no flaws to the methodology. This position is allied to the claim that RCTs are the "*gold standard*" of quantitative research [178,362].

In the interpretative framework, "*crude EBM*"'s interpretative assumptions are: absolute interpretations, with RCT evidence given a high-quality rank and everything else (with the potential exception of meta-analyses and systematic reviews of RCTs) a moderate or low quality rank; unconditional and unmodified interpretation; an 'evidence quality' interpretation property; and, extremely broad scope—the hierarchy applies to all questions, and includes all kinds of evidence (anything not explicitly mentioned is implicitly low-ranked—a form of exclusion criterion or catch-all category). Chapter 4 argues that most existing criticisms of hierarchies are devastating for the crude interpretation. But many existing criticisms are *only* problematic for crude interpretations. It is quite hard to find explicit examples of crude EBM amongst hierarchy authors—the exception is perhaps van

Tulder et al.'s 2000 hierarchy ([305], see Figure 20, above). While early EBM sources are sometimes cited as examples of crude EBM, this is often a misrepresentation, as argued in Section 3.1.5.

In contrast, "sophisticated" proponents accept that different kinds of evidence are "the best available" in different situations. RCTs are not always appropriate—sometimes non-epidemiological sources like clinical experience, pathophysiology and mechanistic reasoning are 'best'. Moreover, they accept the:

> *"degree and type of uncertainty that is inherent in any clinical application of knowledge and in any form of biomedical research."* ([123], p.577).

The implications of this for hierarchy interpretation are less clear. But "sophistication" imposes some constraints. First, there must be different hierarchies for different questions. Many hierarchy authors now offer distinct hierarchies for diagnosis, prognosis, harm and treatment (e.g. [15,90,103,104]). Daly et al. even present a hierarchy of qualitative evidence [189].

Second, awareness of uncertainty implies that unmodified and unconditional interpretations should not be defended—even the most reliable methods sometimes provide low-quality evidence (see the "Bad Implementation Problem", discussed in Chapter 4, below). Depending on the understanding of the "uncertainty" involved, conditions might be used to solve this problem; if the causes of lower-than-expected quality evidence are known (i.e. there is only uncertainty about *whether* a particular study will have these issues, not what the issues are), then downgrading conditions can be added accordingly. But if the uncertainty is at the level of the problems that can occur (i.e. there is uncertainty about what can go wrong, or about whether a known problem has or probably has occurred in a specific case), then it seems only modified interpretations will suffice in order to express the possibility that evidence from a particular method could be lower quality than expected.

Uncertainty might work in the other direction too—usually-weak methods might sometimes provide higher quality evidence. But that is not directly implied by the "sophisticated" view as conceived by Brody, Miller & Bogdan-Lovis. The central assumptions of the "sophisticated" view are, then: different rankings for different questions; modified (and possibly conditional) interpretation; only relative interpretations are needed[53]. The interpretation property is left open. The "sophisticated"

[53] To escape the RCT-worshipping trap of "crude EBM", the assumption that if there is no high-quality evidence, then there is no role for evidence, must be avoided. Crude EBMers are sometimes accused of assuming that whereof there are no RCTs, thereof we must remain silent. By contrast, for sophisticated EBMers, the absolute quality-level is not important—what is important is *using the best available evidence*. For that, only relative rankings are needed. Notably *superlative* interpretations are also necessary, which may cause difficulties if the hierarchies are not exhaustive, especially as Brody, Miller & Bogdan-Lovis' version of sophisticated EBM clearly includes all kinds of evidence, including pre-appraised and non-epidemiological sources.

interpretation, primarily through the use of modifiers, will escape many of the problems discussed in Chapter 4. However, the interpretations are of much more limited value—claims like 'The evidence from this RCT is probably higher quality than the evidence from this cohort study' have weaker implications for practice than do the claims of crude EBM, as is demonstrated in Chapter 5.

3.2.2: La Caze's Modest Interpretation

La Caze, in *"Evidence-Based Medicine Must Be…"* [11], offers his favoured interpretation of hierarchies. The interpretation has three features: relative ranking; internal validity as the interpretation property; and heavy conditionality and/or modified interpretation. La Caze also suggests that a hierarchy might be given a *ceteris paribus* reading. His view is a deliberately "modest" interpretation. He attempts to limit interpretative assumptions to the least philosophically controversial. One never makes claims about evidence *quality* or strength on the basis of hierarchies. The hierarchies only rank evidence by probable internal validity. One must make inferences about quality, strength, etc. on this basis, using context and further information. Moreover, hierarchies do not imply, say, that RCTs provide high-validity evidence—rather they only rank comparative internal validity (i.e. only relative interpretations are accepted). Some modification or heavy conditionality is also needed; no method *always* provides higher-validity evidence than another. The strongest interpretations under La Caze's modest schema resemble: 'The evidence from this RCT probably has greater internal validity than the evidence from this cohort study.' Such a modest interpretation would survive most criticisms levelled at hierarchies, as Chapter 4 will show. However, the modesty of the interpretation may be a considerable drawback in terms of practical usefulness, which will be the topic of Chapter 5.

3.2.3: The Heuristic Interpretation

In the supporting documentation for Howick et al.'s 2011 "OCEBM Levels of Evidence" [15], the authors claim that:

> *"in addition to traditional critical appraisal, [the hierarchy] can be used as a heuristic that clinicians and patients can use to answer clinical questions quickly and without resorting to pre-appraised sources."* [16][54]

Howick et al.'s hierarchy is supposed to be a *"heuristic search tool to find and use the likely best evidence"* [16]. This suggestion fits with a philosophical position recently defended by Djulbegovic, Guyatt & Ashcroft [86], and critically developed by Sturmberg [160]. A 'heuristic', for our purposes, is an unproven (and unprovable) rule—a way of attempting to reach a desired outcome which is not *sufficient* to ensure that the right outcome is reached (i.e. it is not proven to work, precisely because it does not always work) [86,317,363,364]. One might call a heuristic a "rule of thumb".

The better the heuristic, the more often users will achieve the desired outcome by using it. In practice, heuristics usually provide a high success rate under "normal" circumstances, but a considerably lesser success rate under other circumstances [317]. If a hierarchy is a heuristic, and the desired outcome is an accurate appraisal of evidence, then the measure of a hierarchy-qua-heuristic is how often accurate appraisals of evidence result from using a hierarchy. If EBM itself is a heuristic, and the desired outcome is improved patient-welfare, then the measure of EBM is how often using its process improves outcomes for patients.

Heuristics are usually based upon empirical evidence and implemented subject to experience and expertise. The experienced practitioner knows the rules of thumb, and knows when they are likely and unlikely to work. As such, one might ask whether a heuristic interpretation is compatible with placing clinical experience at the lowest rank in a hierarchy.[55] If EBM or the hierarchy is a heuristic grounded in clinical experience, then surely that clinical experience should not be low-ranked—or at least, ranking clinical experience as low quality will undermine the hierarchy itself (i.e. the hierarchy is not itself based on high-quality evidence). This challenge is mistaken. First, there are different roles for clinical experience and expertise. Howick outlines a number of roles for clinical experience and expertise in clinical practice [59]. Unsystematic experience can be poor *evidence for a claim*, yet be an important component of a process, such as the application of a hierarchy to evaluate evidence. To argue that clinical experience should not be used *as evidence* for the effectiveness of a treatment does not entail that clinical experience should not be used *in appraising* the evidence for the effectiveness of that treatment. Much as surgical expertise is necessary for effective surgery, expertise in evidence appraisal may be necessary for successful appraisal.

[54] When Howick et al. say *"without resorting to pre-appraised sources"*, they presumably mean that their hierarchy allows users to appraise evidence from scratch without needing to find summaries in summary-journals, like *ACP Journal Club* [100], or individual appraisals such as CATs [41]. They do not mean "without including any pre-appraised evidence", as each draft of the OCEBM hierarchy includes at least systematic reviews of RCTs, which are pre-appraised sources [15,103].

[55] Howick et al.'s *CEBM* hierarchy [15] does just this.

But perhaps clinical experience must have an *evidential* role in underpinning a hierarchy-as-heuristic. The heuristic is supported by experience—the argument for the heuristic is an induction over past successful use of the heuristic (under certain circumstances). As such, clinical experience *is* the evidence base for the heuristic hierarchy. To combat this response, one can invoke a second distinction: unsystematic experience can be poor evidence *for the effectiveness of a treatment*, yet good evidence *for a heuristic*. In other words, the ranking of evidence given by a hierarchy is not reflexive; one would not use the same hierarchy to appraise the evidence for the hierarchy.

The hierarchy as heuristic position is discussed in more detail in Chapter 7. For now, there are several interpretative assumptions embedded in this position. Primarily, a heuristic interpretation implies probabilistic modification (the *likely* best evidence, as Howick et al. put it). Heuristics sometimes misfire. There may be some element of conditionality placed upon the interpretations—a high success rate for the heuristic is only likely under certain conditions. These are different conditions to those usually imposed—the conditions focus on the circumstances of appraisal (are the conditions suitable for the hierarchical method to appraise the evidence?). Howick et al.'s heuristic view commits to only relative interpretations, but absolute interpretations are compatible with it, as long as modifiers are always included.

3.3: Conclusions

Section 3.1 showed that the interpretative assumptions made by hierarchy authors changed over time. There are four broad periods identifiable in the development of hierarchies, each associated with a cluster of interpretative assumptions. Pre-EBM hierarchies used level of evidence as the interpretation property, and are limited in scope to epidemiological evidence. They are compatible with, but do not explicitly endorse, absolute interpretations, superlative interpretations, and all forms of modification and conditionality. Early EBM was a period of experimentation in the use of hierarchies, involving hierarchies of evidence and of methodology. The interpretative assumptions were uniformly relative, superlative interpretations were not endorsed, and the scope was limited to epidemiological evidence. Interpretation properties such as strength and level of evidence were common. Conditionality was normal, and modification is sometimes suggested. In the Later EBM period, the scope broadened to include all evidence in medicine. Superlative interpretations were more common. Relative interpretations were generally used by EBM proponents, but institutional hierarchies used absolute interpretations. There was no modification and little conditionality. Finally, in contemporary EBM, the focus is largely on conditional or modified interpretations, though these interpretations are generally absolute. The scope of some hierarchies is limited again to epidemiology, though this might be read as the exclusion of non-epidemiological evidence altogether.

Section 3.2 presented several philosophical views of hierarchy interpretation. La Caze and Howick et al. both defend interpretations which they argue are the best way to understand hierarchies. La Caze's is a modest interpretation, which limits the philosophical controversy of the assumptions, at the cost of the usefulness of the information which results. Howick et al. understand hierarchies as heuristics, designed to give the right answer most of the time, without necessarily being grounded in rigorous argument. They are products of experience, not of theory. Brody, Bogdan-Lovis & Miller, by contrast, are concerned with discerning two types of interpretation they identify in the EBM literature: crude and sophisticated views. The crude view is diametrically opposed to La Caze's modest interpretation, licensing very strong interpretations. Its downside is philosophical untenability: Chapter 4 will demonstrate the range of serious problems which affect crude interpretations. The sophisticated view is primarily defined as the absence of crudeness, and so does not have much to add in terms of specific assumptions, but is perhaps the target for the contemporary period of EBM.

The philosophical views of La Caze and Howick and the Contemporary EBM hierarchies of the GRADE Working Group, Evans, and others, are primarily responses to perceived problems with the Later EBM period hierarchies. Unconditional, unmodified, absolute interpretations with broad scope are thought, by most of these authors, to be untenable. Howick et al. take a modifier-centred approach to avoiding these problems. GRADE uses conditionality to the same end. La Caze goes further, rejecting all of the assumptions which the crude view and the Later EBM period interpretations made. The next chapter turns to the problems with hierarchies raised in the EBM literature, and discusses the criticisms levelled at hierarchies, showing that these criticisms can almost universally be solved by either conditionality, modified interpretations, or a combination of the two. The downside to this tactic is a decrease in usefulness of hierarchies.

CHAPTER FOUR:

THE CHALLENGES TO HIERARCHIES

CONTENTS

Chapters 2 and 3 showed that there are many ways to interpret hierarchies, depending upon the interpretative assumptions made. The predominant interpretative assumptions associated with EBM have changed over time, and several authors have defended specific interpretations. This chapter reviews the major criticisms levelled at hierarchies and at EBM accounts of evidence, and argues that the criticisms so far explored in both the medical and philosophical literature are powerful with respect to *some* interpretations. However, no criticism yet undermines *all* interpretations of hierarchies. Moreover, the subset of well-criticised interpretations is far from covering the whole space of potential interpretations. In particular, most criticisms only apply to *absolute* interpretations, or where the interpretations are unconditional and unmodified. The choice of subject matter and interpretation property is also crucial to several criticisms. Several important criticisms only affect hierarchies with broad scope, which include non-epidemiological evidence. As such, there are sets of interpretative assumptions which avoid the major criticisms addressed at hierarchies to date.

The 'crude' interpretation, and the interpretations from the Later EBM period are largely demolished. Forms of Early EBM and Contemporary EBM interpretations survive some but not all criticisms intact. La Caze's modest interpretation and Howick et al.'s heuristic interpretation remain defensible despite the full set of criticisms, though the usefulness of hierarchies under these interpretations can be questioned. There are ways to avoid each criticism using by conditions and/or modifiers, and limiting the scope of hierarchies. This chapter is an argument for the theses that (1) suitably sophisticated GRADE-style conditional hierarchies, and La Caze and Howick's interpretations are the most tenable existing positions given existing criticism, and that (2) sets of interpretative assumptions more broadly are insulated from the existing critical literature where modification and/or suitable conditions are used, and where scope is limited to epidemiological evidence.

4.1: JUSTIFICATION

A number of authors have argued that EBM proponents cannot justify ranking RCTs above other forms of epidemiological evidence (observational studies, etc.) or epidemiological evidence above non-epidemiological sources such as mechanistic reasoning and clinical experience (e.g. [51,75,76,97,148,362]). EBM proponents rarely offer systematic arguments to justify their rankings. But some attempts have been made (e.g. [16,22,59,77,86]). This chapter first shows that objections to these attempted justifications only partially succeed (Section 4.1.1).

There are two main approaches to justifying a hierarchical ranking—empirical and theoretical justification. Empirical justifications appeal to evidence from observations and studies to justify a ranking. Empirical evidence could be used to support specific parts of the ranking (e.g. that RCT evidence is higher quality than observational study evidence) or to justify the claim that a hierarchy should be used in practice by showing that using hierarchies is beneficial. Evidence for a ranking will be considered in this section as an attempt at justifying *a hierarchy*, while evidence that *using a hierarchy* is beneficial will be considered in Chapter 7 in the context of a heuristic interpretation of hierarchies.

A theoretical justification presents a philosophical or methodological argument to show that, *in principle*, evidence from high-ranked designs possesses the interpretation property to a high degree (or to a higher degree than evidence from a lower-ranked design). For instance, the argument that randomization rules out a particular kind of bias which is not ruled out by non-randomized studies could form part of a theoretical justification for ranking RCT evidence above evidence from non-randomized studies.

4.1.1: EMPIRICAL JUSTIFICATION—EVIDENCE FOR A RANKING

This section explores the challenges critics have raised to empirical justification of EBM hierarchies. The two main challenges accuse EBM proponents of circular reasoning, and of confirmation bias in the evidence considered. First, this section reviews the evidence employed in empirical justifications of hierarchy-interpretations. Primarily, empirical justification is provided to support interpretations which assert the low or lower quality, strength or validity of evidence from (a) observational studies, (b) mechanistic reasoning, and (c) clinical experience, compared to RCT evidence.

(a) **Observational studies:** The evidence amassed to show that observational studies provide low-quality, weak or low-validity evidence generally consists of lists of treatments which were thought to be effective based on observational studies, but whose results were later contradicted by the findings of RCTs (e.g. [77,365]). In addition, many meta-epidemiological studies have been published which purport to show that observational studies exaggerate treatment effects, tending to give positive results in more cases than do RCTs (e.g. [366-370]).[56] Collins & MacMahon summarise this argument:

> *"it makes little sense to continue to base inference on observational studies when their results have been reliably refuted by large-scale randomised trials."* ([365], p.459)[57]

(b) **Mechanistic reasoning:** EBM proponents cite a range of treatments which were believed to be effective on the basis of theories about human physiology and purported causal mechanisms. In each case, the treatment was later either shown to be ineffective in epidemiological studies, or the mechanism or physiological theory was rejected in a favour of a new theory. Favoured cases include anti-arrhythmic drugs to decrease the risk of myocardial infarction in patients suffering from Ventricular Ectopic Beats (VEBs)[58] ([22], p.6; [20]; see [375,376])—a policy which was responsible for an estimated 10,000 deaths in the US [377], routine foetal heart monitoring [75], grommets for glue ear [187] and the use of drugs which increase exercise tolerance in heart failure patients such as prostacyclin and milrinone/Prostacor [217,378], which were subsequently shown to increase cardiovascular mortality [379,380].[59] Moreover, they cite more archaic examples such as blood-letting [381] which were accepted due to discredited physiological theories such as humorism [58], and uncontrolled clinical experience [381,382]. They further present lists of formerly accepted treatments subsequently for which RCTs have been negative (e.g. [77]).

[56] In response, several authors [51,58,69,121,192,371,372] cite studies which show observational studies to be more consistent (i.e. to more often agree with other observational studies about whether there is a treatment effect, and how great that effect is) than are RCTs, and that the results of suitably comparable RCTs and observational studies are generally in close agreement [371-373]. A recent Cochrane review of studies comparing randomised and non-randomised studies concluded that the evidence is too weak to draw substantive conclusions about whether non-randomised studies show different effects to randomised studies [374]. However, their interpretation was similarly circular—the authors interpreted this as further evidence of bias in non-randomised studies, claiming that the similar results were due to a mixture of observational studies over- and under-estimating treatment effects, cancelling out. They argued that this makes the bias of non-randomised studies even more pernicious, as it is not known whether the studies will be positively or negatively skewed.

[57] Note however that Collins & MacMahon [365] argue for an important role for observational studies in clinical practice: these studies are particularly adept at estimating absolute treatment effect *sizes* in actual practice, and in detecting side-effects and adverse effects.

[58] The link in the mechanistic chain at fault is the unsupported leap from a suppression of a non-lethal arrhythmia correlated with adverse outcomes, to a lower risk of myocardial infarction.

[59] Again, here the mechanistic link between the surrogate outcome (the increased exercise tolerance) and the patient outcomes (mortality, morbidity) is at fault [378].

(c) **Clinical experience:** There are two ways in which clinical experience is used as evidence—"general" and "individual" clinical judgement ([59], p.160; [58]). "General clinical judgement" takes the consensus of experienced practitioners as evidence of the effectiveness or ineffectiveness of a treatment. General clinical judgement was a prominent source of evidence prior to the influence of EBM—for instance, through consensus conferences. "Individual clinical judgement" refers to using one's own clinical experience and expertise to judge the likely effectiveness of a treatment for a patient. Evidence has been presented against both of these forms of clinical experience. Similar examples to those cited in (b) are used against the general case. Bloodletting is again a prominent case of clinical consensus being fooled by uncontrolled experiences and (possibly) placebo effects [58]. Howick cites the Halsted radical mastectomy for breast cancer, a treatment which persisted largely due to clinical consensus in face of mounting of epidemiological evidence that it was both dangerous to women and no more effective than minor lumpectomy combined with chemo- and radio-therapy ([59], pp.165-6). Another prominent case is the reluctance to give corticosteroids to women in premature labour, despite Victoria Crowley's meta-analysis which showed that the treatment was associated with major reductions in infant mortality [383].[60]

With respect to individual clinical judgement, Howick and colleagues [58,59] cite evidence that mechanical prediction rules outperform clinicians in most judgement tasks. For instance, the Ottawa Ankle Rules were developed to guide clinicians in determining whether patients with ankle pain should be offered an X-ray to check for bone fractures [384]. Studies showed that clinicians who followed the Ottawa rules missed fewer ankle fractures and performed fewer unnecessary X-rays than clinicians who did not follow the guidelines [384,385]. Although these judgement tasks do not generally include therapeutic decision-making—a task complicated by the need to take patients' preferences and values into account—the evidence does suggest that individuals who use their 'experience' to deviate from evidence-based guidelines commit more errors than those who stick to the guidelines (e.g. [386,387]).

4.1.2: Circularity in Empirical Justification

There are two major challenges to these attempted empirical justifications—circularity and confirmation bias. The circularity objection holds that attempts to justify a relative ranking placing RCTs above observational studies, mechanistic reasoning or clinical experience using this empirical evidence are unavoidably circular (see [51,55,388]). Judging observational studies, mechanisms or

[60] This case was significant in the call for more attention to be paid to systematic reviews and meta-analysis. The forest plot from Crowley's meta-analysis later became the Cochrane Collaboration's logo ([59], p.161-2).

clinical experience to have supported false effectiveness claims depends upon the assumption that the effectiveness claim is false. The effectiveness claim is believed to be false on the basis of RCT evidence. To show that RCT evidence is superior to non-RCT evidence, one must assume that, where RCTs and non-RCTs contradict, the RCT is right. This argument is circular.

In response, first note that not all of the evidence presented above required RCT evidence to contradict the non-RCT evidence. For instance, bloodletting was not rejected due to an RCT. Similarly, the evidence that mechanical prediction rules outperform clinical judgement does not depend on accepting the results of the epidemiological studies which underpin the prediction rule. The use of such prediction rules against individual judgement provides independent testability for the claim that the epidemiological study provides more reliable results.

Second, in many cases EBM proponents are not simply using a single RCT to 'trump' the non-RCT evidence. Although Lacchetti et al.'s list [77] does include cases where there is simple contradiction between a single RCT and a single observational study, most of the prominently-discussed cases involve a great deal more evidence. Basing the argument simply from Lacchetti et al.'s list would expose the EBM proponent to a charge of vicious circularity. However, where there are a large number of studies—and meta-analyses, systematic reviews, and studies other than RCTs—which contradict clinical experience, mechanistic reasoning, or a previous observational study, then rejecting the contradicted result on that basis seems justified. When a range of independent studies, using different methods, consistently agree, then a single unremarkable source of evidence which contradicts this consistent body of research is probably mistaken. This can allow the identification of invalid and unreliable results without necessarily having to make strong judgements about the methodology of studies which contradict that result.[61] This is clearly the case in at least the radical mastectomy, corticosteroids and VEBs cases.

Finally, suppose that, due to circularity, one cannot justify the relative ranking of RCTs above clinical experience, mechanistic reasoning or observational studies directly through this empirical argument. It does not follow that the empirical evidence cannot play a role in justifying this relative ranking, or that it cannot justify an absolute ranking. Suppose that the claim that RCTs provide high quality evidence can be justified through other means (i.e. theoretical justification—this is akin to setting up the RCT as the 'gold standard'). Then the circle is broken. The claim that the RCT-evidence is valid, plus the claim that the non-RCT evidence contradicts the RCT evidence, justifies inferences about the validity of the non-RCT evidence. So, while it may not be possible to establish the superiority of RCT evidence empirically, if one can independently establish the *absolute* ranking of the RCT evidence,

[61] Such a principle would clearly be a heuristic, as it cannot (by design) take account of the relative quality of the large body of evidence compared to the single isolated study.

then empirical evidence can support the inference that clinical experience, mechanistic reasoning, or observational studies tend to provide lower quality evidence.

It is not necessary to justify a strong absolute claim like “RCTs provide high-quality evidence” to set up a fixed reference point from which to use empirical evidence to justify further interpretations. Establishing a modified claim like “RCTs are highly likely to provide high-quality evidence”, plus the inconsistency between RCTs and other sources, licenses an inference that the other sources probably provides lower quality evidence. Thus a modified claim can be used to justify other modified claims.

A more challenging criticism of empirical justifications even for absolute interpretations comes from the confirmation bias of most attempts at empirical justification. Suppose a hierarchy is interpreted as implying: ‘Mechanistic reasoning provides low quality evidence for claims about treatment effectiveness’. A range of cases in which mechanistic reasoning provided low quality evidence for such claims is provided as evidence, as in (b), above. Is the claim justified?

No—the evidence exhibits confirmation bias: providing confirming cases for the claim, without considering whether there are disconfirming cases. It is a mistake to infer from a set of low quality mechanistic studies that mechanistic studies are low quality or even probably low quality. Exactly such reasoning from anecdotal evidence is what the EBM movement sought to discredit and avoid (see [39]).

Perhaps a probabilistically modified form could be justified, for instance: ‘In practice, mechanistic reasoning tends to provide low quality evidence for claims about treatment effectiveness’. But here the reasoning is threatened by a base-rate fallacy. From a number of cases of low-quality mechanistic reasoning, it does not follow that it is probable that mechanistic reasoning will be low-quality. The catalogue of cases might only amount to a few rare exceptions amongst a generally high-quality cohort of mechanistic studies. The evidence only justifies the claim that mechanistic reasoning tends to be low-quality if a large *proportion* of low-quality mechanistic studies was found amongst a proper sample of mechanistic studies. This objection is particularly strong given natural salience bias. People are more likely to notice and remember cases which were salient for some reason. In medicine, salient cases often involve serious harm and loss of life. Cases where catastrophe resulted from mistaken reasoning such as radical mastectomy, VEBs and bloodletting, which cost tens of thousands of lives and caused mutilation and suffering, are extremely salient. Cases in which mechanistic reasoning provides a sound understanding of the physiological effects of a drug are comparatively unremarkable. As such, any set of cases selected manually could be expected to be significantly skewed—a form of selection bias. Only a statistical sampling will justify the claim empirically, which is not typical of the current evidence base.

A more tenable argument would use the identified cases of mistakes made on the basis of mechanistic reasoning to demonstrate that *there is a risk* of mechanistic evidence misleading clinicians. Showing that reliance upon mechanisms has led clinicians to false beliefs about the effectiveness of treatments demonstrates that the quality-range of mechanistic evidence includes the low end. This is reminiscent of Howick and colleagues' common-sense position [58,59,389], that mechanistic evidence may often (or even usually) be low quality, but that does not mean that the few instances of high-quality mechanistic reasoning should be ignored. To integrate this insight into a hierarchical approach would require either upgrading conditions for mechanistic evidence setting out the criteria under which it can be considered high-quality[62], or at minimum a modified interpretation. No hierarchies yet support this approach for mechanistic reasoning. However, GRADE has gone some way to attempting this with respect to observational studies. The upgrading criteria in GRADE [14,202] are primarily meant to capture the loose requirements for considering observational studies as providing high quality evidence offered by Sackett et al. [22] (for instance that the effect shown is large, or that the biases affecting the study would skew against the result found).

4.1.3: Theoretical Justification

The main theoretical justification for hierarchies of evidence ranking RCT evidence above evidence from other kinds of study is that RCT evidence has higher internal validity. It has sometimes been claimed that randomization is the only way to achieve high internal validity (see [52,75,76,388]). For instance, Pocock & Elbourne claim:

> *"Only randomized treatment assignment can provide a reliably unbiased estimate of treatment effects"* ([390], p.1907)

Other authors defend the claim that RCT evidence has higher internal validity than non-randomized study evidence based on claims that the RCT methodology, and randomization in particular, prevents certain biases which could not be prevented otherwise, or is the most reliable way of preventing those biases (e.g. [20,22,390,391]). There are three main kinds of bias which proponents of EBM have claimed are ruled out by RCTs, or at least are better mitigated in RCTs than in non-randomized studies: confounding or allocation bias, selection bias and treatment bias.

[62] Howick [58,59,137,389] has offered such criteria, but they have not been integrated into hierarchies.

This section will defend a number of claims, the result of which is a justification in terms of exposure to selection biases for a ranking which places RCT *and* non-randomised experimental evidence, above observational study evidence, above case series evidence. Ultimately, a probabilistic argument is defended here for the claim that experimental studies are less likely to be biased by imbalances in unknown confounders than are observational studies. However, the justification only applies to a ranking given a modified and conditional interpretation. Moreover, no justification for ranking epidemiological evidence above non-epidemiological evidence is provided. The claims which constitute this argument are:

(1) Randomization does not guarantee high internal validity.
(2) Randomization does not prevent confounding due to unknown confounders, but baseline checks allow for prevention of confounding due to known confounders.
(3) Non-randomized experimental studies can prevent confounding due to known confounders to the same or better degree as RCTs.
(4) Observational studies are vulnerable to biases caused by self-selection, which can only be controlled for when they cause confounding on known confounders.
(5) Self-selection creates a difference in the exposure to unknown confounders between experimental and observational studies—while neither type of study can control for unknown confounders, observational studies are more likely to have imbalances in unknown confounders.
(6) Case series are vulnerable to another additional form of selection bias (i.e. inclusion bias).

4.1.3.1: Randomization does not guarantee high internal validity

It is not the case that RCT evidence always has high internal validity. It can be quickly demonstrated that some RCT results do not show true effects through a counterexample in which a well-conducted RCT result is patently false. One such case is due to Leibovici [392], discussed by Worrall [75,76]. Leibovici's RCT studied the effects of "*remote retroactive intercessory prayer*"—that is, a prayer said for patients' recovery by an unrelated agent, many years after the event. Leibovici found that the prayer had statistically significant results on several outcomes, such as duration of hospitalization. In this case, there is no need for an experimental result to contradict the RCT's finding—almost everyone will recognize that the claim that retroactive prayer is effective is absurd, and agree that the RCT evidence is invalid, despite being conducted in strict accordance with methodological standards. It presents a straightforward counterexample to interpretations of hierarchies like: 'RCT evidence has high validity'—though not, of course, to modified interpretations like: 'RCT evidence tends to have high validity'.[63]

[63] One might, like Tonelli [393], further take this case to show that mechanistic reasoning still has an important role as a *negative* argument—that is, if it is clear that no plausible mechanism could be provided for the

4.1.3.2: Known and Unknown Confounders

Worrall [52,75,76] argues that randomization is neither necessary nor sufficient for validity by eliminating the arguments for the special epistemic status of randomization.[64] In particular, some authors have claimed that randomization controls for all known and unknown confounders (e.g. [307,399]). For instance, Kunz, Vist & Oxman claim that:

> "*Randomisation is the only means of controlling for unknown and unmeasured differences as well as those that are known and measured*." ([226], p.2)

Worrall shows that this claim is false. As defined in Chapter 2, confounders bias study results and undermine causal interpretations of those results. Confounding occurs when factors which affect the outcomes being measured are not equally distributed between groups. Confounding which results from unequal initial allocation of patients to study groups is called 'allocation bias'.[65]

Randomization does not necessarily balance confounding factors. There is no *guarantee* that a random allocation will split a given factor evenly between the groups. If, say, 50 participants are male and 50 are female, it is possible (if unlikely) for a random allocation to distribute all 50 men to one group and all 50 women to another. Less extreme but still important imbalances are not so improbable. Moreover, if there are many confounding factors (and for complex medical interventions, it seems reasonable to expect that there will be a great many potential confounders), the chance that *all* confounding factors are near-evenly distributed in a given random allocation is low.[66] Worrall claims that the argument that all confounders are balanced by randomization results from a quantifier fallacy:

> "*Even if there is only a small probability that an individual factor is unbalanced, given that there are indefinitely many possible confounding factors, then it would seem to follow that the*

effectiveness of a treatment, then the treatment cannot be effective (and moreover any study to the contrary must be invalid). For further discussion of negative evidence in hierarchies, see Chapter 6.

[64] Note that Worrall [52] also discusses philosophical arguments for the special epistemic status of randomization from Papineau [394], Cartwright [362,395] and Pearl [396]. These arguments are not discussed here—the arguments made are not used by proponents of EBM, and have been comprehensively treated by Worrall [52] and Urbach [397,398].

[65] Not all confounding is allocation bias. Some confounding could be introduced after the initial allocation. For instance if patients drop out of the trial and are not accounted for in the analysis, then an initially balanced allocation could become unbalanced (aka. attrition bias) [246,337,338].

[66] This is exacerbated by the fact that validity is defined in terms of *confidence in* the accuracy of results and causal interpretation. So, even if all factors which *actually* affect outcomes are balanced, if there are factors which a reasonable person would *suspect* are confounders and which are unbalanced, validity is affected. Perhaps male and female patients being unevenly distributed in a trial did not affect the outcomes, but if it is reasonable to suspect it might, the validity of the trial is affected.

> *probability that there is some factor on which the two groups are unbalanced (when remember randomly constructed) might for all anyone knows be high*" ([51], p.S324).

Randomization does not guarantee that all confounding factors are balanced. Worrall cites the fact that standard practice and textbook advice is to check allocations for baseline imbalances. Worrall notes that:

> "*most guides to RCTs recommend checking the two groups (…) for 'baseline imbalances'.*" ([75], p.993)—(see e.g. [400-402]).

If possible, researchers will re-randomize where baseline imbalances are found, until a suitably balanced allocation is chanced upon. In some cases re-randomizing is not possible, for instance where patients are enrolled over time and begin the trial at different points. In these cases, the usual advice is to check for baseline imbalances when reporting the results. Statistical techniques to control for the effect of baseline imbalances have been developed [402].

It is this process of checking for baseline imbalance and re-randomizing, *not randomization itself*, which allows researchers to control for known confounders. Given that randomization does not guarantee balance, unless baseline checks are performed, one cannot be confident that known confounders are balanced and therefore not biasing the result. Therefore, it is the process of checking and adjusting for baseline imbalances which provides confidence that the results are not affected by allocation bias.

However, only known confounders can be checked for baseline imbalances. Therefore, it is always possible that an unknown confounder is unbalancing the groups, and thus biasing the trial result. RCT methodology cannot rule out unknown confounders, but can rule out imbalances in known confounders (assuming checks for imbalances are part of the methodology). To justify ranking RCT evidence above non-randomized study evidence on the basis of allocation biases, an argument that non-randomized studies cannot or do not exclude imbalances in known confounders would be needed.

4.1.3.3: Selection Bias in Randomized and Non-Randomized Experimental Studies

The difference between randomized and non-randomized studies is usually then cast in terms of *selection bias* (*cf.* [51,52,75,76])—imbalances in confounders due to selective allocation. In a non-random selective experimental study, the trialist may be responsible for dividing patients into groups. For example, if patients are allocated to the control or experimental group in alternation, a clinician who knows the schedule could easily manipulate the allocation. For example, they could exclude a patient who was less healthy than average if the next patient was to be included in the experimental group, or include a less healthy patient if the control group was next (see [1,226]). In observational studies, the patients self-select their group—for instance, a comparative observational study of treatment effectiveness compares outcomes in those patients who chose one treatment option to those who chose another. If patients with a worse prognosis tend to chose one treatment over another, then there will be bias against that treatment in the study. Randomization is supposed to eliminate selection because it pays no attention to patients' features. As Roberts & Torgerson put it:

> "*controlled trial randomisation ensures that allocation of patients to treatments is left purely to chance.*" ([401], p.185)

Worrall accepts that randomization does prevent selection bias [75,76] (where successfully performed—and, like in the case of blinding, there is good reason to be sceptical of the success-rates of some randomization procedures [314,315,400,403,404][67]). However, there is no reason to believe that random allocation is *the only* method which could exclude selection bias. One could, for instance, develop a computer program designed to perform a selective allocation which balances all known confounders. Clearly this method is selective. But it does rule out selection *bias*. One example of such a method is "minimization", in which an algorithm allocates each patient to the group which would minimize imbalances in confounders between the groups [405,406]. The Cochrane Collaboration has accepted that minimization is equally able to exclude selection bias as randomization:

> "*Minimization may be implemented without a random element, and this is considered to be equivalent to being random*" ([1], p.198).

Indeed, minimization may provide *more* balance in confounders than randomization, as one review found:

> *"minimization provides better balanced treatment groups when compared with restricted or unrestricted randomization"* ([407], p.662).

Randomization *can* produce baseline imbalances, and must be checked to ensure imbalances are not present. Even if baseline checks are carried out, some threshold level of acceptable baseline imbalance must be used. Computer allocation should be able to minimize imbalance between confounding factors. Randomization *might* stumble onto the optimal balance, but this seems unlikely (and less efficient). Moreover, a randomizer will not know when it has found the optimal distribution.

[67] In particular, randomization via an unconcealed string of random numbers and randomization by sealed envelopes is easily subverted (see [1,314,403]). If the order in which patients are to be allocated to treatment groups is determined by a random string, but this random sequence is known to the clinician in advance, then the randomization procedure is no better than simple alternation [400]. The randomness of allocation should not be equated with concealed allocation [315,400]. Concealed allocation is also not equivalent to blinding.

Therefore, it seems that a selective computer algorithm would balance known confounding factors *as well if not better* than randomization.[68]

Therefore, it is not necessary to use randomization in order to prevent selection biases. Non-random allocation methods can be just as good, if not better, at excluding confounding. Randomization is not sufficient to exclude confounders because, like any method, it cannot ensure balance in unknown factors. So, RCT methodology is neither necessary nor sufficient to ensure that allocation and selection biases are prevented.

4.1.3.4: Selection Bias and Observational Studies

However, while non-randomized studies can prevent selection bias, it does not follow that all methods are equally good at doing so. Borgerson [227] claims that non-comparative studies such as case series and case reports are not affected by selection bias because they do not divide patients into two groups. However, they are affected by selection in the form of inclusion bias—if a researcher reports only the cases in which a treatment was effective and not those in which it failed (or vice versa) then the case series will give a very biased impression of the effectiveness of the treatment. Without explicit selection criteria for inclusion in a case series, evidence from case series is very exposed to bias. To accommodate this, hierarchies could rank case series at a low level (as they uniformly do) with an upgrading criterion where explicit selection criteria are given which convince the appraiser that the case series represents a fair sample of patients.

Borgerson also shows that a similar argument could be mounted against observational studies. Non-random experimental studies can be insulated against selection bias by methods such as minimization. But in observational studies, patients self-select their group by their treatment choices. If treatment choice is correlated with a confounding feature, then self-selection biases the result.

In response, Borgerson [227] argues that checks for baseline imbalances and statistical techniques to adjust for these imbalances are equally applicable in such observational studies as in RCTs. Moreover, a 'matching design' can be used for the observational study, in which participants are chosen to be observed in such a way that confounding factors are balanced between the groups. Of course, neither baseline checks and adjustments nor matching can control for unknown confounders. But neither can randomization. Therefore, observational studies, like randomized studies, can

[68] It strikes me as strange that anyone would think an iterated random procedure plus checks for baseline imbalances would be better (or more efficient) at balancing groups than a computer programme designed solely to balance groups, which is designed to maximize fairness. A scattergun approach of randomization will always eventually find some sufficiently balanced allocation with respect to known confounders (assuming one exists), but users will have to perform checks to know when that has occurred. Minimization, by contrast, selects the best allocation without user input [405-407].

introduce controls for selection bias. There may be a case for giving an initial low rank to observational studies, and upgrading their evidence conditional on adequate matching or successful baseline checks. However, clearly the same reasoning should apply to RCT evidence. Borgerson also correctly notes that hierarchies do not distinguish non-randomized experimental studies from observational studies. One lesson from this discussion, then, is that hierarchy authors should make this distinction, and rank minimization experimental designs at the same (or higher) level as RCTs.

However, there remains a difference between the exposure to selection biases in observational studies as opposed to in experimental studies. Where self-selection takes place, selection could be causally related to an unknown confounder. It is not implausible that patients choose a particular treatment in part due to an underlying factor which makes them also more responsive to that treatment or at lower risk of side-effects. One relevant consideration is expectations and placebo effects. Patients who have high expectations of a treatment and believe that the treatment works tend to experience greater effects from that treatment. This may be because patients are more compliant with a treatment they believe is working, but may also be due to enhanced placebo effects associated with expectations of effect (see e.g. [229,230]). Patients may also experience more severe side-effects from treatments they believe are efficacious (aka. the nocebo effect [408]). When patients are allowed to choose their treatment, it seems plausible that they will choose the treatment which they most believe to be effective. Therefore, patients in observational studies may have exaggerated treatment effects due to their belief in the effectiveness of the treatment. This effect may be particularly strong where some patients had choice and others did not (for example, in natural experiments such as comparisons of areas in which a treatment was offered to areas in which it was not). However, patients' beliefs in the effectiveness of a treatment may be very difficult to check at baseline and adjust for, or to match in a matching design. This is only one example of potential associations between treatment choice and treatment effects. *Self-selection* may be the bias which truly distinguishes observational and experimental studies.

RCTs and properly-conducted non-randomised experimental studies cannot control for unknown confounders, but do not allocate patients to groups on the basis of factors potentially causally connected to unknown confounders. By contrast, treatment choice in observational studies may be causally linked to the effectiveness of the treatment. Therefore, there is reason to believe that observational studies will be more likely to be affected, and more severely affected, by bias due to unknown confounders. This constitutes a theoretical justification for ranking experimental study evidence above observational study evidence.

4.1.3.5: Treatment Bias

As defined in Chapter 2, treatment biases occur when different groups are treated differently other than in the treatment protocol they receive—for instance, when they are treated by clinicians with different levels of competence and experience, or with different levels of flexibility or monitoring. It has sometimes been argued that treatment biases are less likely in RCTs than in non-randomized studies (*cf.* [51,75,76,227]). Randomization, so the argument goes, permits us to control for this by facilitating double blinding. Where allocation is blinded, practitioners cannot deliberately treat the experimental group preferentially, as they are unaware which group is the experimental group. But it is not randomization, again, which prevents this bias—it is double blinding. Admittedly, it is difficult to blind observational studies. But experimental non-randomized studies can be blinded.[69] Additionally, blinding does not exclude treatment bias—blinding can be (and studies suggest often is [409,410]) broken, and groups could be treated differently irrespective of blinding (for instance, if one group is treated by different staff to the others). The fact that one group is experimental and the other control, which is concealed from the practitioners, is not the only motive which could create differences in treatment. What's more, not all RCTs can be blinded (the ECMO trials are a particularly well-discussed example of an un-blindable RCT—see [69,187,272]), not all RCTs which can be blinded are blinded, and not all RCTs where blinding was attempting are successfully blinded (see [409,410]). RCT methodology is, at least, the wrong property to prize in excluding treatment bias.

4.1.3.6: To which Interpretations does the Justification apply?

This section has shown that there is no justification for separating properly conducted non-randomized experimental study evidence from properly-conducted RCT evidence in terms of allocation, selection or treatment biases. Observational study evidence faces different challenges to experimental study evidence, in that self-selection could cause an increased likelihood of unknown confounding. Where observational studies are not affected by self-selection (e.g. where the study compares different times or areas where choice was not a factor in treatment allocation), or where selection is not conceivably connected to confounders, observational study evidence may be equally valid as experimental study evidence. But there is reason to believe that experimental studies are better able to prevent certain biases than non-experimental studies. Therefore, the justification provided here only applies to modified and conditional interpretations of hierarchies—it is *probably* the case that (subject to certain conditions regarding the ways the studies were conducted) evidence from an experimental study has higher internal validity than evidence from an observational study.

Note that there has been no attempt to justify the claim that epidemiological studies provide evidence with higher validity than mechanistic evidence or clinical experience. It seems difficult to

[69] For instance, if patients are allocated to treatment and control groups by a computerised matching program (see below), the practitioners and patients can be equally unaware of the allocation.

imagine such a theoretical justification, primarily because the kinds of claims which are supported by mechanistic evidence are often different to the kinds of claims supported by epidemiological evidence (see Chapter 6 for further discussion). So, the justification described here applies only to hierarchies with scopes limited to epidemiological evidence.

A justification can be provided which is compatible with the arguments made by Worrall. RCT methodology cannot rule out bias caused by unknown confounders. But self-selection introduces an additional source of confounding, in addition to chance, which does not affect experimental studies. Unless self-selection does not occur or is not plausibly connected to any unchecked confounder, there is some reason to have higher confidence, ceteris paribus, in the evidence from experimental studies than from observational studies. This provides a benchmark to start using empirical evidence to assess studies' findings, as described above.

Chapter 6 will show that this justification ultimately fails to support a hierarchy of evidence. The point for now, though, is that the existing critical literature does not comprehensively show that no nuanced interpretation of a hierarchy can be justified.

4.2: Overvaluing RCTs

This section addresses a range of criticisms of RCT methodology in particular, which attempt to show that ranking RCT evidence as strong or high quality is an error. The arguments demonstrate once again that absolute unmodified unconditional interpretations of hierarchies are untenable. In each case, though, a more nuanced interpretation involving conditions and modifiers can still be defended.

4.2.1: The "Bad Implementation" Problem

Grossman & MacKenzie [388] and Bluhm [55,56] rightly observe that some RCTs are performed badly, and some observational studies are performed well. RCT design is no guarantee that the advantages discussed above will actually obtain. They argue that: (1) it is absurd to claim that all RCTs provide evidence which is of the same quality or strength; and, (2) at least some RCT evidence is low-quality and/or low strength. A whole range of methodological problems might be encountered in RCTs: the trial may be unblindable, randomization may be subverted, baseline imbalances may be unchecked or unaddressed, or treatment bias may be rampant (see Section 4.1.3, above).

Even if the trial methodology is well implemented, there is variation in the quality and strength of evidence: a trial might be very small—a trial of only a handful of patients is an 'RCT' just like a trial of 10,000; it may be underpowered to detect effects; it may have enrolled only a very specific subset of the population; a large number of patients may have been lost to follow-up, dropping out during the trial, etc. Clearly, not every RCT provides the same quality evidence as every other. Therefore, it is absurd to rank evidence from all RCTs at the same absolute quality level. Moreover, an RCT which has many of the defects identified here would provide low-quality, low-strength evidence. In both respects, Grossman & MacKenzie and Bluhm are wholly correct. Unmodified and unconditional absolute interpretations ascribing high quality to evidence from all RCTs are undermined. However relative and/or modified interpretations go unscathed, as they do not require that all RCTs be of equal quality, or that all RCTs provide high quality evidence. Conditional interpretations which downgrade RCT evidence where these problems are evident are also still defensible.

The criticism goes further, though: if some observational studies are well-performed, very large, well-powered, and representative of the target population, while some RCTs are badly-performed, very small, underpowered and unrepresentative of the target population, then surely some observational studies provide *higher* quality evidence than some RCTs (see also [148]). Even unconditional unmodified relative interpretations do not escape unscathed, then.

The solution to the bad implementation problem is to move to a more nuanced interpretation. Modified interpretations avoid the challenge. Some RCTs providing low quality evidence is consistent with the claim that RCT evidence is probably high-quality. Conditional interpretations may also circumvent the challenge, if they include sufficient conditions to downgrade RCTs which suffer the methodological defects discussed above, and to upgrade observational studies which avoid methodological problems. GRADE certainly survives this objection. The bad implementation problem limits which hierarchy interpretations are defensible, rather than showing hierarchies themselves to be untenable.

4.2.2: Commercial Bias in RCTs

Section 4.1 discussed ways in which RCTs might exclude biases which are not easily controlled for by other methods. Brody, Miller & Bogdan-Lovis [123] have argued that RCTs may in practice suffer from an additional bias which other designs tend to avoid—commercial bias.

A substantial proportion of RCTs are funded or conducted by pharmaceutical companies to try to demonstrate the efficacy of their products. Various governmental licensing agencies require demonstrations of efficacy in RCTs prior to approval, which heavily motivates commercial enterprises to produce positive results. One might expect commercially-funded RCTs to tend to be positively biased, due both to intentional and subconscious influences upon the findings. Many studies of funding bias indicate that this is indeed the case (e.g. [411-417]).

In addition, commercial bias may manifest as publication bias [415]. Publication bias occurs when studies are less likely to be submitted for publication, published, published prominently, and republished because they did not have a positive result [198,239,240,418]. While some of the blame for publication bias lies at with editors and consumers [247], there is evidence that commercial funding correlates with publication and re-publication of positive results and non-publication or delayed publication of negative results [239,241,245,415,419]. Some pharmaceutical companies have litigated against researchers or editors to delay or prevent publication—Rennie [420] reports a case in which litigation by Boots Pharmaceuticals Inc. delayed publication of a negative finding ([421]) by 7 years. As such, commercial bias could justify the claims that: (a) positive trials are more likely to be biased than negative trials, and (b) negative trials are less likely to appear in searches. Reliance on searches of RCT evidence, then, may give a misleading picture. Furthermore, the problem may be exacerbated, not solved, by drawing upon meta-analyses and systematic reviews unless analysts and reviewers are meticulous in tracking down suppressed data and reviewing trial protocols and adherence [241,245].

However, the commercial bias objection has a limited impact upon hierarchies of evidence. First, there is no reason why other forms of evidence could not be subject to commercial bias and publication bias. While RCTs may more often be funded by pharmaceutical companies due to licensing requirements, one could argue that they are actually methodologically less susceptible to manipulation than other methods—more safeguards against manipulation such as double blinding are feasible. Moreover, the argument only affects absolute interpretations—RCT evidence could still be *higher* quality than non-randomized studies, even if commercial bias undermines claims that the evidence is *high* quality. Hierarchies can be adapted to take the threat of commercial bias into account—one could, for instance, add a condition for the highest-level evidence that the researchers performing a study must be independent from commercial interests (or at least that the findings have been independently verified). Such a condition would salvage even absolute high-quality interpretations, as well as the justification given in Section 4.1.3. Modified interpretations are again unaffected unless commercial bias is the norm.

Commercial bias is perhaps a stronger challenge to hierarchies which rank systematic reviews and meta-analyses at the top. If reviews are susceptible to publication bias then they may be systematically positively biased, and so should be lower ranked. Modification may not provide a sufficient solution for this issue for hierarchies—if systematic reviews and meta-analyses are *systematically* positively biased, then it would not suffice to say that, for instance, meta-analyses 'tend to be high quality', unless evidence could be provided that most meta-analysts successfully test and control for publication bias. There are methods available to test for publication bias, such as funnel plots (see e.g. [198,250,251]). Commercial bias could be assessed in meta-analysis by comparing the subset of commercially-funded trials to the subset of non-commercial trials within the meta-analysis, and by sensitivity analysis, performing the analysis with and without commercially-funded trials included. The most plausible suggestion to adapt hierarchies to deal with publication bias is to impose conditions which downgrade reviews and meta-analyses which fail to impose such checks.[70]

[70] To date, conditional hierarchies like GRADE do not include meta-analyses or systematic reviews, so this option has not been pursued. But GRADE does downgrade for evidence of publication bias [198].

4.2.3: The Problem of Small Effects

As Penston [178] and Worrall [228] observe, there is a strong tendency towards larger clinical trials, oriented towards discovering small differences in effect-size between similar treatments. There are clear commercial incentives at work: pharmaceutical companies can enter a crowded but lucrative market very profitably if they can demonstrate that their version of a popular drug is marginally more effective than competitors. But when searching for small effects, the problems of biases are amplified. Even a small bias could produce the appearance of a small difference in effect. In other words, high-quality evidence of a small effect must rule out even small biases. Penston and Worrall both provide plenty of reasons to suspect that these biases cannot be adequately excluded. As such, the evidence from large trials for small effects should not generally be regarded as high quality evidence.

This is an important point—one should indeed be very sceptical of small differences in effects, even when demonstrated in large trials. Statistical power will not exclude small biases. However, this argument will not undermine many hierarchy interpretations. The more small effects are investigated, the more false positives should be expected. However, this is not a phenomenon unique to RCTs. Seeking small effects will be difficult and will increase the false-positive rate no matter how evidence is gathered. Hence, while the problem of small effects might make one doubt the *absolute* quality of RCT-evidence for small-effect claims, it does not demonstrate that RCTs do not provide *higher* quality evidence than other sources—relative rankings are unaffected. Moreover, to preserve some absolute interpretations, one need only add conditions to require that only RCTs which demonstrate large (or no) effects are classed as 'high' quality—GRADE has already gone some of the way to adding this condition (large effects allow upgrading, but marginal effect-sizes do not yet allow downgrading) [14,202]. More should be done to link quality of evidence with the size of the effect demonstrated in hierarchies. But this could be achieved within the framework of current hierarchies.

4.3: Undervaluing Mechanisms and Clinical Experience

Existing arguments do not establish that relative, modified and conditional interpretations of hierarchies which rank RCTs highly cannot be justified. Interpretations such as La Caze's [72] and Howick's [16] survive, as does a suitably updated GRADE. However, relative interpretations could be undermined from the opposite direction—by demonstrating that lower-ranked evidence is higher quality than these hierarchies admit. This section considers three arguments, which Howick calls "anomalies" [422], which might show that mechanistic reasoning, clinical experience and expertise are seriously undervalued by hierarchies. It then reviews the arguments from Clarke, Russo, Williamson and Illari [62,63] that hierarchies undervalue mechanistic evidence.

4.3.1: The "Paradox" of Effectiveness

Howick calls the most general form of this anomaly argument[71] the "Paradox of Effectiveness" [59,393,422]. The paradox can be formulated in one of two ways. The strongest version of this "paradox" claims:

> There are a number of treatments which are known to be effective *(i.e. the level of evidence that the treatment is effective is high)* but which are based on evidence which hierarchies would appraise as *weak* or a low level of evidence.

This is the most recognisable form of the 'paradox', and has been discussed previously by Worrall[72] [51,75,187]. This argument dates back at least to Rubin's work in the 1970s, in which he pointed out that most scientific inquiry does not employ RCT methodology [423]. Worrall and Howick provide a range of treatments which are very strongly believed to be effective, yet which are not supported by

[71] Howick has offered a number of forms of this argument, both in his book and articles [58,59,137,249]. He gives a more detailed breakdown in a presentation available online [422]. Howick does not defend these arguments himself, but should be credited with formulating and cataloguing the challenges. Howick's own defense of EBM against these criticisms is not explored here—a precise and thorough defense can be articulated through the framework of interpretation.

[72] Note that Worrall also discusses these cases from an ethical perspective: it is unethical to claim that RCTs must be performed where there is clear consensus that a treatment is effective, justified by non-epidemiological evidence. In this regard, the argument remains very powerful, unconnected to Howick's version as a challenge to hierarchies of evidence.

evidence given high ranking in hierarchies of evidence. The list includes: aspirin for headache, penicillin for pneumonia [51,187], appendectomy for appendicitis, and the Heimlich manoeuvre for clearing blocked airways [58,59,187,422].[73] An infamous example comes from Smith & Pell [426], who satirised the EBM movement by claiming that a double-blind RCT was needed to test the effectiveness of *"Parachute use to prevent death and major trauma related to gravitational challenge"*. Clearly there is well-justified strong belief in the effectiveness of parachutes. Yet there is no strong evidence for this belief, or the level of evidence is low, according to EBM hierarchies.[74]

What interpretation of an EBM hierarchy would these 'anomalies' contradict? Initially, it seems the required interpretation is: 'The evidence from *X* is weak/low level evidence for the effectiveness of *Y*', where *Y* is the treatment, and *X* is the existing evidence for the effectiveness of *Y*. So, for example, 'The evidence from physical theory and unsystematic experience is weak evidence for the effectiveness of parachute use'. Normally, the assumption is that there is no epidemiological evidence at all for the effectiveness of the treatments concerned—aspirin, appendectomy and parachute use are all known to be effective based on clinical experience and mechanistic reasoning alone.

This interpretation is absolute—it claims that the evidence is *weak*. A relative interpretation may be acceptable in some of these cases. For instance, one might accept that *if there was* an RCT supporting the effectiveness of aspirin for headaches or appendectomy for appendicitis, this would be even stronger evidence still. But this response will not work in every case—certainly, it seems there is the strongest evidence, the highest level of belief, that could possibly be justified in claims like 'Parachute use is effective in preventing gravitational trauma'. The only way this could be accommodated is if the relative ranking was also non-strict—that is, lower-ranked evidence is *as strong as or weaker than* higher-ranked evidence. So, the evidence from mechanistic reasoning that parachute use is as strong as RCT evidence would be.

Moreover, the paradox also only affects unmodified, unconditional interpretations. If the interpretation has probabilistic modification, then the problem is resolved—a modified interpretation is consistent with evidence which is given a low ranking actually being strong evidence in some cases. It can equally be overcome using upgrading conditions—for instance, by upgrading the assessed strength of clinical experience and mechanistic reasoning if it demonstrates a very large effect

[73] These anomalies may be much less clear-cut examples than previously claimed. For instance, a meta-analysis [424] and a systematic review [425] both found that medical management is inferior to appendectomy for acute appendicitis (NB: these studies were published after the original claim by Worrall [51]). If this criticism were more substantial, I would advocate dividing the potential anomalies according to the reason for the absence of epidemiological studies—some are not performed because the effectiveness of the treatment is universally accepted (e.g. parachute use), but others due to the cost, ethics or circumstantial difficulties (e.g. in extreme emergency cases such as the Heimlich manoeuvre).

[74] This intuition is even more strongly illustrated when considering obviously harmful or needless procedures. For example, a randomized trial is clearly not needed to check whether leaving surgical implements inside patients is truly harmful.

(*cf.*[202]). This latter response seems to be the one generally endorsed by Howick, who argues that confounding and other methodological issues can be outweighed if a study demonstrates a large enough effect. Bird calls this:

> *"Howick's rule of evidence: studies provide good evidence when the effect size outweighs the combined effects of plausible confounders."* ([116], p.4)

Hierarchies can accommodate this rule. GRADE has done so by including upgrading criteria which include all potential confounders reducing the effect size demonstrated, and the study demonstrating a large effect size [202]. But GRADE does not currently include clinical experience or mechanistic reasoning in its ranking. The scope of GRADE would have to be expanded to implement this response in practice.

The paradox is also only problematic for hierarchies which include non-epidemiological evidence in their ranking. Hierarchies which do not include non-epidemiological evidence—the majority—are not susceptible to the criticism because they do not entail any claim about the strength or quality of that evidence. Unless, however, they rank evidence bases by level of evidence, in which case omitting evidence from clinical experience and mechanistic reasoning *could* remain a problem. It is possible to have a low level of evidence from epidemiological studies, but a high level of evidence from omitted evidence such as clinical experience and mechanistic reasoning. So the assessment that there is a low level of evidence for the hypothesis made by such a hierarchy would be false because it omits the evidence which justifies belief in the hypothesis.

However, this challenge to hierarchies of evidence bases with narrow scope can be sidestepped by clearly explicating what such hierarchies actually entail. A hierarchy of evidence bases which ascribes a low level rating to an evidence base cannot entail that *there is a low level of evidence for hypothesis X* unless the evidence base assessed includes all evidence. Where the scope is narrower, the only claim justified is: '*The evidence base assessed* provides a low level of evidence for hypothesis X'. The same applies to hierarchies interpreted in terms of strength of recommendation. Instead of reading such as a hierarchy as entailing that a strong recommendation cannot be justified, it should be interpreted as entailing that *the evidence base assessed* cannot justify a strong recommendation.

The "Paradox of Effectiveness" problem is a devastating challenge to hierarchy interpretations which are absolute, unmodified and without adequate conditions, which include non-epidemiological evidence from clinical experience and mechanistic reasoning within their scope—precisely the kinds of interpretations popular during the 'Later EBM' period, and associated with 'Crude EBM' [123]. Any one of a narrower scope, conditions for upgrading mechanistic evidence and evidence from experience where the effect is clear and obvious, modifiers, or a non-strict relative ranking can resolve the challenge.

4.3.2: The "Large Effect" Paradox

Howick considers a further version of the 'paradox'—the "Large Effect" paradox [422]. This version uses examples in which an effect is so large and obvious that no systematic studies are needed to for observers to know that the treatment works. One such case might be the Heimlich manoeuvre. The patient is clearly experiencing a blocked airway. Performing the manoeuvre causes the blockage to fly out—the effect is so clear and obvious that no further studies are needed. Other examples include defibrillation to restart a stopped heart, mother's kiss to remove a foreign blockage from a nostril, and adrenaline injections for anaphylactic shock [422].

The large effects version of the "paradox" is not significantly different from the original challenge. The same kind of response suffices—this is only a problem where the hierarchy includes non-epidemiological methods, and the interpretation is absolute, unmodified and unconditional. Authors of conditional hierarchies which include non-epidemiological evidence must be sure to include a very large upgrading condition for observations which immediately demonstrate the effectiveness of the treatment.

In fact, a number of hierarchies have made specific provisions for this kind of case. An example is the Phillips & Ball [130,131] *CEBM* hierarchy, which included "*All-or-none*" evidence at the highest level—this refers to:

> *"when all patients died before the Rx became available, but some now survive on it; or when some patients died before the Rx became available, but none now die on it."* ([103], n.§)—where 'Rx' refers to treatment regimen.

The 'all-or-none' category was included in Sackett et al.'s 2000 hierarchy, and retained in the 2009 *CEBM* Levels of Evidence [103], although removed without explanation in the 2011 version [15,16]. This might include cases like appendectomy, defibrillation and adrenaline injections for anaphylactic shock, and could be tweaked to include cases like mother's kiss by substituting death for complete recovery.

A final response to this kind of argument might consider exactly what the role of hierarchies within appraisal should be. The role of hierarchies could be tweaked so that they are used to appraise evidence where it is not already obvious that a treatment works. Perhaps a hierarchical appraisal should be restricted to cases where systematic evidence appraisal is needed. Where the evidence is so obviously compelling that no one would dispute it, appraisal is not only superfluous but inappropriate, and hierarchies are inapplicable.

4.3.3: Phillip's Paradox and the Paradox of Efficacy

Finally, Howick considers a version of this argument which he calls "Phillip's Paradox" [249,422], originally from Ney, Collins & Spensor [427]. Phillip's Paradox is the problem that, in double-blind trials, very large effects undermine blinding. If the effects of the experimental treatment are extremely large or visible, then patients experiencing these effects will be aware that they are in the experimental group, not the control. Similarly, their physicians will also become aware of their allocation. As such, Howick argues, the most effective treatments cannot be tested in a gold standard double-blind RCT. Nevertheless, there is the highest level of evidence for the effectiveness of such treatments.

The simplest response to this case is merely to discard it—very few EBM hierarchies make any claims about blinding. Ranking 'RCTs' at the top level does not entail that only double-blind RCTs count. Only a hierarchy in which ranking is conditional upon blinding will be affected by Phillip's Paradox, and then only if a more nuanced condition (i.e. 'unless the effect is very large') is omitted. GRADE, for instance, allows for additional biases that might be introduced by unblinding to be offset by a large effect size. As yet (and perhaps surprisingly), no hierarchy explicitly includes double blinding as a condition for top-ranked evidence.

A variant upon this argument which is not considered explicitly by Howick (though he has discussed a similar argument in other contexts [428-430]) could be called the "paradox of efficacy". Several EBM proponents have argued that, in order to demonstrate the efficacy of a treatment, a *placebo-controlled* RCT is necessary (see [431]). They argue that active-controlled studies are not *assay sensitive* (that is, they are not able to detect whether the experimental treatment has an absolute beneficial effect—studies comparing two 'active' treatments are only able to tell which has the greater effect and by how much, not the actual size of the treatment effect) [432]. If this argument is right, then there is only good evidence for the efficacy of treatments which have been tested in placebo-controlled trials. A large number of treatments (including but by no means limited to those listed above) have never been tested against placebo. Yet there is, justifiably, a high level of belief that these treatments are efficacious.

This argument (if the assay sensitivity concern is correct, which is controversial [428-431]) would only affect EBM hierarchies if they enshrined the view that only placebo-controlled trials can establish efficacy. However, no hierarchy has ever made mention of placebo controls as a methodological condition. As such this apparent 'paradox of efficacy' dissolves in practice.

Phillip's paradox and the paradox of efficacy serve as cautions against overzealous application of conditions. They are species of a larger problem which does affect some hierarchies, in which conditions are imposed because they help to reduce certain biases or increase the size or precision of the observed effect, but no provision is made for evidence in which these biases do not pose a threat or the effect estimates are already precise. For instance, some hierarchies stipulate that top-ranked evidence must come from large multi-centre trials [31,105,132,207]. But these hierarchies do not make provisions for evidence which comes from smaller, single-centre trials but which shows a very large effect. It is important to consider the various ways in which high-quality, strong evidence can be provided, and ensure that these are included, when producing conditional hierarchies.

4.3.4: The Russo-Williamson Thesis

Recent work by Clarke, Russo, Williamson and Illari has presented an argument against hierarchies based on an account of the epistemology of causal inference. Russo & Williamson's "*epistemic theory of causality*" [60] states that in order to confirm the claim that A is a cause of B, evidence that A and B are correlated *and* evidence that there is a mechanism linking A to B which accounts for A's effect upon B are both required. According to Clarke et al., the 'Russo-Williamson Thesis' (RWT) states:

> *"In order to establish that A is a cause of B in medicine one normally needs to establish two things. First, that A and B are suitably correlated—typically, that A and B are probabilistically dependent, conditional on B's other known causes. Second, that there is some underlying mechanism linking A and B that can account for the difference that A makes to B."*
>
> ([63], §2.2)

Note that it is only *normally* necessary to establish correlation and mechanistic connection in order to establish causation in this version of RWT. Clarke et al. do not suggest circumstances in which the two are not individually necessary, nor whether it is correlation, mechanism, or both which can sometimes be unnecessary. Perhaps they are considering the kinds of cases Howick called anomalies, above, in which unsystematic experiences or mechanisms alone suffice to establish causation.

Second, as Illari [61] has emphasised, RWT does not require two *kinds of evidence* or even that two different sources of evidence are needed to establish a causal claim. It is consistent with RWT for a single study to provide both evidence of correlation and evidence of mechanistic connection, and therefore provide evidence for causation.

Finally, RWT provides necessary conditions for *confirming causal claims*, not necessary conditions for the existence of a causal connection. It may often be the case that a causal claim cannot be confirmed given the available evidence. This does not entail that no such causal relation exists. Therefore, the inability to provide evidence of correlation between A and B or the inability to provide a suitable mechanism does not entail that A does not cause B, only that a causal connection from A to B cannot currently be confirmed. Therefore, even a close adherent of Russo & Williamson's epistemic theory of causality may provide different criteria for establishing *non-causation*[75]—that is, there is an open question as to how one might confirm the claim that A *does not* cause B. The question of confirming non-causation is discussed further in Chapter 6.

Clarke and colleagues offer a sustained defence of RWT in the health sciences [60,62,63,433]. Drawing upon the criticisms of RCTs and epidemiological evidence described in this chapter, they argue that evidence of correlation alone is insufficient to confirm a causal claim in medicine. They accept some of the criticisms of causal inference on the basis of mechanistic evidence alone raised by authors such as Sackett [20,22] and Howick [58,59]. Suppose that RWT is true—it will almost always be necessary to have evidence of correlation and mechanistic connection in order to confirm causal claims in medicine. Does it follow that hierarchical approaches are untenable?

Clarke et al. state:

> *"It is this balance, we would argue, that has been lost in present-day evidence hierarchies."*
>
> ([63], §2.2)

There is a sense in which hierarchies of evidence lose the balance between evidence of correlation and evidence of mechanism. 'Mechanistic reasoning' has certainly been allocated a low level in many hierarchies. For many hierarchies, the challenge can be resolved. In hierarchies of evidence which do not include mechanistic reasoning, and are not interpreted as entailing that excluded evidence is low-quality or non-evidential, the hierarchy entails no judgment about the quality or strength of mechanistic evidence. It remains problematic that evidence from an RCT alone

[75] I will use the term 'non-causation' to refer to 'A does not cause B'. I intend this terminology to help in avoiding ambiguity in phrases such as 'negative evidence for causation'—which can refer to *evidence that A has a negative causal effect on B* (e.g. A prevents B) or *evidence that A does not cause B*.

could be judged high quality or strong evidence for a causal claim. A condition could be added to these hierarchies (either for each level in the hierarchy, or for the hierarchy to apply at all) which requires a plausible causal mechanism for the treatment effect being considered. Surely, where evidence for a causal mechanism already exists, further evidence of correlation can be strong and high-quality evidence for the causal claim. Evidence for a causal mechanism could be appraised in parallel, either through a novel hierarchy of evidence for mechanisms (a similar approach was taken to evidence from qualitative studies by Daly et al. [189]), or through some other criteria (e.g. Howick's criteria for high-quality mechanistic evidence [58] or a modified version of Bradford Hill's guidelines [137,434]). Alternatively, a heuristic hierarchy could resolve the problem if a plausible causal mechanism is normally in place for treatments being tested. Finally, a non-strict relative hierarchy avoids the problem, as the claim that RCT evidence is as strong or stronger than observational study evidence for a causal claim does not entail that the RCT evidence is at all strong.

A stronger challenge to hierarchies from RWT affects hierarchies of evidence bases, interpreted in terms of level of evidence or strength of recommendation. In GRADE, an evidence base of RCT evidence justifies a high degree of belief in the hypothesis that a treatment effect is produced, and (if the cost-benefit ratio is sufficiently clear-cut) a strong recommendation that the treatment should be used. But if RWT is correct, then without mechanistic evidence, the RCT evidence base cannot justify a high level of belief in the hypothesis that the treatment caused the treatment effect. Here, unlike in hierarchies which appraise quality or strength, the restriction of hierarchies to epidemiological evidence does not solve the problem. Parallel hierarchies or other appraisal procedures for assessing mechanistic evidence are superfluous, because an RCT evidence base alone suffices to justify a high level of evidence.

There are several alternatives available to attempt to integrate the insights from the work of Clarke et al. into the framework of a hierarchy of evidence bases. GRADE, in particular, has some of the tools to consider an important role for mechanistic evidence. However, at present, no hierarchies of evidence bases succeed in doing so. Chapter 6 will return to this problem, arguing that it undermines the majority of hierarchies of evidence bases, but that a modified version of GRADE could deal with at least some of the force of mechanistic evidence—particularly where biological implausibility tells against a causal claim.

This challenge undermines less sophisticated level of evidence hierarchies, and provides a serious problem for GRADE which requires augmentation of the appraisal procedures. If RWT is correct, then GRADE needs at minimum to require that mechanistic evidence be included in any evidence base which is assessed as high level. Perhaps the most natural way to do this would be to stipulate that evidence of a plausible mechanism is a minimum threshold condition for the hierarchy to apply. So, in order for appraisal even to be appropriate, mechanistic evidence must first establish that the treatment effect hypothesis is plausible. It is only *given the existence of a plausible mechanism* that a GRADE hierarchy can assign levels of evidence. This may again threaten the balance between mechanistic evidence and evidence of correlation, as mechanisms become a basic requirement, and then evidence of correlation is the arbiter of level of evidence. However, it is consistent with RWT. A different account of RWT would be needed to establish that not only is evidence of both mechanism and correlation necessary to establish causation, but that both play an equal role—RWT as stated in Clarke et al.'s paper does not entail that *balance* is required.

A similar approach would be to include downgrading criteria where the mechanism by which a treatment has its effects is unclear, and/or upgrading criteria where a clear, well-confirmed mechanism is offered. This latter approach requires the weaker version of RWT as stated in Clarke et al.'s paper—that evidence of mechanism and correlation are both *normally* required—as opposed to the stronger version in which evidence of both mechanism and correlation are always necessary to confirm causal claims. This approach allows for evidence of correlation alone to establish causation where the evidence satisfies enough of the upgrading criteria to counteract the downgrading effect of a lack of mechanistic evidence.

In summary, there are ways of interpreting hierarchies of evidence which are consistent with RWT. Hierarchies of evidence in which the interpretation property is quality or strength may be consistent with RWT given suitable conditions, modification or a relative reading. Hierarchies of evidence-bases are more clearly problematic given RWT. If a high level of evidence for a causal claim cannot be justified without mechanistic evidence, then in effect mechanistic evidence should be a condition required to access the higher levels in such hierarchies. This is not implemented in any current hierarchies. Restricting scope to epidemiological evidence will not overcome this challenge. Changes are necessary to accommodate RWT, either in making higher level rankings conditional on a plausible mechanism, or in making the applicability of hierarchies themselves conditional on mechanistic evidence.

4.4: External Validity

One of the most repeated challenges to EBM concerns *external validity* (e.g. [11,55,69,72,97,148,178,213,214,256-259,261,362,435-437]). In chapter 2, external validity was defined as justifiable confidence that the result and effect found in a study in study-population P_S generalises to target population P_T. Meanwhile, internal validity is the confidence in the accuracy of a result and effect in the study population, P_S. Where P_T=P_S, external and internal validity are the same.[76] However, in medicine studies are not often undertaken in the target population. The classic problem of external validity is that of warranting confidence in the accuracy of a projection of a result in P_S into P_T, where $P_S \neq P_T$.

Cartwright [257,258,362], Rawlins [97] and Ho et al. [256] argue that RCT evidence is not usually useful for clinicians because RCTs generally provide evidence with low external validity. 'External validity' alone is not well-defined: one must specify external validity *in some population* P_T. According to Ho et al. [256], P_T is the population of patients a doctor sees in her routine practice. This varies by specialism and location. They argue that RCT evidence rarely has sufficiently high external validity in P_T for it to be useful for clinicians (see also [435]).

There are many features of clinical trials which threaten external validity in P_T (see [213,222]). Most of these features are due to steps taken to enhance internal validity within the trial population P_S. Steps taken to enhance internal validity include stringent exclusion criteria for participants [192].[77] Many clinical trials exclude patients with co-morbidities, non-standard presentations of the target condition, patients outside the 18-65 age bracket, and even women [439-441]. The choice to enrol in a trial correlates with specific psychological, physiological and economic features of patients [69,442]. Study conditions may be highly regulated. Patients may be checked for compliance or treatments may be directly observed [254]. Treatment may take place in a hospital setting, often in highly-specialized units, given by experienced and specialized clinicians with a strong understanding of the intended treatment regimen [148]. Dosage may be clearly and narrowly prescribed, and variation from guideline dose may be prohibited, even where side-effects are apparent [254,443,444]. Placebo effects may be exaggerated due to the expectations that an experimental treatment is radically effective [229]—data suggests that patients treated with the same therapy respond better if told that the drug

[76] This is sometimes the case in medical research—for instance in an *n*-of-1 trial, and in some forms of outcomes research.

[77] The idea is to select a fairly homogeneous cohort with relatively little variation in the major confounding factors. However, this might also be done to increase the likelihood of positive results or serve commercial interests [438] (see Section 4.2.2), as the homogeneous cohort usually selected is the group most likely to respond well to treatment with the fewest side-effects.

is experimental or that they are in a trial [232-234,445]. Self-reporting of symptoms and effectiveness may be affected by both expectations and the desire to cooperate with the perceived 'desired result' [229,230,444]. Side-effects may be enhanced or even generated by nocebo effects due to the expectation that experimental drugs are risky and potent [230,408,446].

By contrast, 'in the wild' these conditions are unusual [76,254]. Clinicians and support workers generally are not as well-versed in the specifics of the treatment regimen. They may make more errors in directing patients and administering interventions. Compliance is likely to be much lower, especially in treatments which the patient is solely responsible for administering, and for conditions which affect memory or attentiveness (see e.g. [447,448]). Especially for certain conditions (e.g. arthritis [260] and Alzheimer's [449]), many or even *most* sufferers may be outside the age range studied in the majority of trials [439,440]. Many patients will suffer co-morbidities, and be undergoing or have undergone other treatments [450]. Clinicians will want to modify dosage to enhance responsiveness in low-responding patients, or to ameliorate suffering where patients experience serious side-effects [443]—going "off-label" is a common practice (see [451,452]). Even if clinicians do not experiment with lower doses, patients may self-regulate dosage, increasing dose if they experience high or low benefit, and decreasing dose if they experience side-effects. Patients may also self-medicate for side-effects, which can cause unanticipated drug interactions, or take other medications which affect the treatment's effectiveness.[78]

Some authors claim that internal and external validity are 'in tension' or even opposed (e.g. [256,456,457]). This is a misunderstanding which springs from the problem that some *ways of enhancing* internal validity threaten external validity, and vice versa. But internal validity is clearly an important consideration in determining external validity—if one is not confident that the effect has been accurately estimated in P_S, this surely affects confidence in a generalisation to P_T [213].

Nonetheless, the tension between ways of enhancing internal validity and external validity is important. Ho et al. may be right that standard heavily-controlled RCT designs fail to *balance* internal and external validity concerns. The emphasis upon ensuring internal validity may come at the expense of generalisability. The components of high-quality evidence identified in Chapter 2 included generalisability to the target population and clinical significance as part of a recommendation or decision-making process. If RCT evidence normally lacks these latter two properties, then RCT evidence is not normally high-quality evidence, contrary to the claims implied by most interpretations of most hierarchies.

[78] A notable example of the latter is St. John's Wort, a commonly used alternative medicine which has been shown to diminish the effects of almost 50% of commonly prescribed drugs, including the contraceptive pill [453-455].

At first glance, this challenge only applies where a hierarchy is given an absolute interpretation. If relative interpretations are used, it could be that RCT evidence is fairly low quality due to inapplicability to practice, yet is still better than lower-ranked evidence (perhaps because lower-ranked evidence has low internal validity, which entails low external validity and applicability). To make this challenge pertinent to relative interpretations, it would be necessary to argue that lower-ranked evidence such as that from observational studies *has* external validity and applicability, in addition to arguing that RCT evidence lacks it. It would be a mistake to assume that observational studies usually have high external validity just because they usually have design features conducive to external validity—broad study populations, inclusive criteria, real-world conditions and 'normal' practitioners, etc. [313] As noted above, all of these features will not guarantee external validity even in a very similar target population unless confidence in the study result in the study population is justified—in other words, unless the study has sufficiently high internal validity. Contrary to what is sometimes assumed in the literature (e.g. [69,255,266,456,458,459]), an argument for external validity of observational studies will take more than demonstrating that they have pragmatic features (*cf.*[148]).

There are a number of ways to approach this problem while retaining even absolute interpretations of hierarchies.[79] One method is to retreat to a modest position like La Caze's [72]. La Caze's approach uses internal validity as the interpretation property for hierarchies. Given his interpretative assumptions, hierarchies do entail any claim about evidence quality—only about the likely internal validity of the evidence. Further inferences to the quality of that evidence must be made separately (for La Caze, by using biological theory and mechanistic reasoning [124]). While this tactic insulates the hierarchy from this objection, it leaves the clinician without guidance in crucial steps of evidence appraisal. La Caze's position makes a *hierarchy* itself no longer particularly clinically useful.

A more user-friendly approach might offer an additional hierarchy using the interpretation property of external validity, allowing clinicians to use both hierarchies together to estimate both internal and external validity, and make inferences about quality accordingly. But this system is relatively complex, and no hierarchy has yet been proposed to rank evidence according to external validity. The task is substantially complicated by the observation that the external validity of evidence varies according to the target population. The relationship between study-population and target-population will surely be more significant than the methodology used in determining the external validity of the evidence.

[79] La Caze ([72], pp.12-3) argues that 'categorical' hierarchies cannot overcome this problem, as generalizing beyond study populations requires understanding of biological mechanisms (see [124]), which are low-ranked by EBM hierarchies. Even if this is so, however, La Caze's argument is incorrect: not all hierarchies include mechanistic reasoning/biological theory at all, and moreover that mechanistic reasoning is low-ranked *as evidence* for a treatment effectiveness claim does not entail that it *cannot* be used either as evidence or as part of an inference using a different evidence base (see below for further discussion).

Modification could be employed to address the problem, but is less helpful here than in combating the objections above. The thrust of Ho et al.'s [256] argument is that *most RCTs* have low external validity. As such, a probabilistic modifier might not solve the issue. Capacity modifiers could solve the problem—'RCTs *can* provide high quality evidence'—but leaves the situation vague. Again, hierarchies interpreted this way are not useful for clinicians, who want to know *under what conditions* RCT evidence is high quality.

Conditions seem to be the most fruitful approach to tackling this challenge. A hierarchy like GRADE could downgrade evidence with low external validity, and upgrade evidence that is readily generalisable to P_T. GRADE hierarchies currently downgrade evidence where there are problems of 'indirectness' of evidence (see especially [129]). This term may capture inferences from surrogate outcomes to patient-oriented outcomes, or inferences across different populations and settings. A more explicit consideration of generalisability and applicability in the GRADE criteria would be welcome to clarify that the problem of external validity is taken into account, along with potential upgrading criteria for direct applicability.

In combination with a conditional approach, hierarchy authors could take a more fine-grained approach to research design. Despite Ho et al.'s objections, there are RCTs which prioritise external validity. Pragmatic RCTs have been performed for decades (see [262,263])—these attempt to replicate real-world conditions closely, using 'normal' clinicians, in 'normal' settings, on broad study-populations [153,255,264]. No hierarchies as yet differentiate between pragmatic and explanatory RCTs in their ranking.[80] Of course, the range from pragmatic to explanatory RCTs is a sliding scale [264]—but the degree of pragmatism could be taken into account in a suitable conditional hierarchy. Where hierarchies of evidence bases are used, the hierarchy could take account of whether the evidence base of RCT evidence includes evidence from both explanatory and pragmatic RCTs, and upgrade accordingly. Moreover, *n*-of-1 trials—also called 'therapeutic trialling' [460], in which a clinician tries out different treatments on a single patient to find the one which works best for them [218]—have clear external validity in the target patient ($P_S = P_T$) to the extent that the trial is internally-valid. Guyatt et al.'s *Users' Guides* [8,83] hierarchy ranks *n*-of-1 trials at the top level, above standard RCTs, although this has not been widely adopted. Therefore, some hierarchies clearly do consider external validity.

In summary, RCT evidence does not necessarily have low external validity, but critics may be correct that it often does in some target populations. It is not clear, however, that non-RCT evidence has higher external validity given that confidence in a study result is a prerequisite for confidence in a

[80] This is particularly surprising where, a la La Caze [72], internal validity is the intended interpretation property. In the reverse of Ho et al.'s concern, pragmatic RCTs will generally have lower internal validity *ceteris paribus* than an explanatory trial, so should be lower-ranked in a La Caze style hierarchy.

generalisation of that result. External validity challenges can be accounted for by introducing conditions which downgrade RCT evidence for inapplicability and upgrade non-randomised evidence where it directly applies to the target population. Hierarchies could take better account of external validity as a component of evidence quality by including such criteria, as well as taking the lead of Guyatt et al. in including RCTs which have high external validity (e.g. *n*-of-1 trials, pragmatic trials) at the top end of the ranking. Level of evidence interpretation properties and hierarchies of evidence bases more generally are less affected by this problem as they could allow for the assessment of evidence from multiple trials in different populations and settings, including pragmatic and explanatory RCTs together. Although this has not yet been implemented in any hierarchy, a ranking could consider the breadth of patients and settings including in trials within the evidence base, and whether both pragmatic and explanatory RCTs have been conducted with consistent results. On the basis of this objection, expanding GRADE's up- and down-grading criteria to take account of these factors is strongly recommended.

4.5: Ethical Challenges

There are serious and legitimate concerns about the ethics of clinical trials. Unethical behaviour can take place in the trial context—patients can be dehumanised and their wishes can be manipulated or overlooked (*cf.* [187,272,427,431,443,444,461,462]). It may be unethical to perform a clinical trial investigating a specific question. Often, the ethical defence for a trial is that of *clinical equipoise* [463]. There is clinical equipoise about X where the medical community at large does not agree about whether or not X is the case. Medical uncertainty is used to justify a trial, despite the risk of harming patients through substandard or ineffective treatment. But trials have been conducted where the condition of equipoise is barely satisfied or not at all. Truog [272], Worrall [52,76,187] and Bluhm [69] have discussed the ECMO (Extra-Corporeal Membraneous Oxygenation) case in detail as an example. One study found that most trials in their sample from rheumatology violate equipoise [438]. Placebo-controlled trials are particularly susceptible to this criticism (*cf.* [427,428,430-432]).

There are two forms of ethical objection here, a strong and a weak objection. The strong objection holds that RCTs are never ethically justifiable. This is not the view held by Truog, Worrall and Bluhm. The weaker objection holds that there are situations in which RCTs are unjustifiable. This should be obvious—for instance, if there are already a number of convincing RCTs providing the required evidence. But Worrall and Bluhm go further—there are situations in which no randomized trials have been performed, yet it is unethical to perform any. There are various such situations. They include Howick's [59,422] 'anomalous' situations, in which there is already such strong evidence from non-RCT sources that an RCT is superfluous to requirements, as well as further cases like ECMO, in which evidence from observational, historically-controlled data is suitably compelling to make RCTs redundant and harmful [187].

The ethical argument must be taken very seriously by EBM proponents. But unlike the hierarchies of the pre-EBM period, EBM hierarchies are not intended to guide research design. The fact that RCTs top many hierarchies does not entail that (a) only RCTs should be performed, nor that (b) RCTs are always appropriate.[81] To derive such claims from even the strongest later-EBM-style hierarchies requires further assumptions: for (a), that only the highest quality evidence is worthwhile, whatever the costs; for (b), that obtaining more evidence is always worthwhile, whatever the costs. Assumption (b) is clearly unfounded—cost-benefit analysis should be performed, at the very least. Most medical ethicists would go much further than that basic requirement. Assumption (a) might be

[81] A possible exception to (b) is Gray's 1997 tickbox hierarchy [340] which does explicitly refer to the appropriateness of research designs.

argued for by some hard-line proponents of RCTs (e.g. [305]) but is certainly an unusual position amongst EBM proponents.

If anything, hierarchies are a positive tool against over-zealous insistence on RCTs. Hierarchies reflect the fact that, even if the lower-ranked evidence is lower-quality or weaker, it is still evidence which can be taken into account in decision-making. It is not necessary to have RCTs in order for medical practice to be evidence-based. For example, Guyatt et al. in the *Users' Guides* are explicit that hierarchies imply that RCTs are not the only valuable source of evidence:

> *"The hierarchy makes it clear that any statement to the effect that there is no evidence addressing the effect of a particular treatment is a non sequitur. The evidence may be extremely weak—the unsystematic observation of a single clinician, or generalization from only indirectly related physiologic studies—but there is always evidence."* ([82], p.1293)

A hierarchy which ranks RCTs highly does not entail that RCTs should always be performed. Inferring that RCTs should be performed from the claim that RCTs provide high quality evidence commits a naturalistic fallacy. Thus, while ethical issues are important and must be tackled by EBM proponents, hierarchies themselves are not at fault, nor are they undermined.

There are more subtle ethical issues to consider. These include whether patients fully understand the nature of clinical trials. The 'therapeutic misconception' is the tendency of patients to believe that trialists have the same obligations to ensure they receive the highest standard of care as do their doctors [443,444]. As such, they may enrol in trials which are against their best interests, misled by the belief that the trialist would not encourage them to enrol otherwise [462].[82]

Finally, Worrall's discussion of the ECMO case brings concerns about stopping rules to the fore [187]. Stopping rules are protocols which terminate a trial if it becomes clear that either treatment arm is endangering patients—i.e. if one treatment arm vastly outperforms the other. Pocock criticises stopping rules for weakening the quality and strength of evidence provided by trials (e.g. [391,464]). Worrall objects that clearly the evidence obtained through such aborted trials is nonetheless very strong—strong enough to convince researchers that the trial must be stopped. In a sense, both sides are correct: a trial which terminates early may provide lower-quality, weaker evidence than had it continued; however, it may still provide overwhelmingly strong evidence. GRADE-style hierarchies offer a potential solution here: evidence might be downgraded where trials were stopped early, but this effect could be offset by upgrading for the large effect observed. The question again is whether,

[82] On the contrary, there is some evidence of a 'trial effect' [232-234]: patients in clinical trials tend to get better outcomes than patients receiving the same treatments outside a trial. This may be due to a higher standard of care, higher level of specialisation amongst caregivers, or due to enhanced placebo effects due to the raised expectations associated with trials.

and at what stage, ethical concerns outweigh epistemic ones. A proponent of the ‘strong ethical objection’ might argue that it is always unjustifiable for epistemic concerns to outweigh ethical concerns (and in particular, ethical obligations to patients). No matter how strongly this argument is made, hierarchies will be unaffected, unless they are given a very crude interpretation which demands that RCT evidence be obtained in spite of the consequences.

4.6: CONCLUSION

This chapter considered 5 broad classes of challenge to EBM hierarchies. In each case, the challenges were either surmountable, or apply only when certain interpretative assumptions are made. Relative interpretations survive largely intact, as do modified interpretations. When conditionality is used, considerable effort must be expended to avoid these challenges—certainly, the minor levels of conditionality at play in most Contemporary EBM hierarchies (e.g. [15,118,132,357]) and even GRADE hierarchies (e.g. [12,13,195]) is insufficient to overcome all of these challenges alone. But conditionality could be a very adaptive response to criticism.

Several sets of interpretative assumptions survive the array of criticisms directed at hierarchies of evidence. Pre-EBM and Early EBM interpretative assumptions tended to restrict scope only to epidemiological evidence, and then with conditions and relative interpretations. As such, they are largely unscathed. Philosophical approaches to hierarchies such as Howick's [16] and La Caze's [72] are defensible. Versions of GRADE with serious updates to the up- and down-grading conditions can meet the criticisms. By contrast, the absolute, unmodified and unconditional interpretations commonplace in the Later EBM period, and exemplified in Brody, Miller & Bogdan-Lovis' [123] characterisation of the "crude interpretation" are thoroughly dismantled, as is the normal interpretation from the Late EBM period.

However, while hierarchies with relative interpretations, modification and suitable conditions survive these direct challenges, adopting these interpretative assumptions may pare away the useful functions of hierarchies. Doctors want to know how good the evidence is for a range of effectiveness claims. When restricted to modified and relative interpretations, hierarchies offer them only information about the comparative quality of evidence, with considerable reservations. As such, an approach which uses sophisticated conditions (along with a heuristic reading like Howick's [16]) might be favourable to a modest approach like La Caze's [72]: although the process of formulating and applying the hierarchy will be more complex, it will at least yield more practically-useful interpretations. This will be the topic of Chapters 5 and 6.

CHAPTER FIVE:

RELATIVE RANKINGS

CONTENTS

This chapter will analyse relative ranking interpretations of hierarchies. There are several ways of understanding relative rankings presented in the EBM literature. It is argued here that these approaches make hierarchies either far too strong or far too weak to be useful. Therefore, limiting the interpretation of hierarchies to relative rankings is not a viable interpretative approach, or solution to the problems posed in chapter 4.

To illustrate the argument, a practical case study is presented in section 5.1, considering the position of a practitioner trying to use a relative ranking while evaluating the effectiveness of a treatment for rheumatoid arthritis amongst patients who have not responded well to standard therapies.

5.1: Tocilizumab

Tocilizumab, also known as Actemra, is an IL-6 receptor inhibitor which has been suggested as a treatment for rheumatoid arthritis (RA) [125,126,465,466] and juvenile idiopathic arthritis [467] in cases in which standard therapies have been ineffective. There are a range of standard treatments for RA, collectively known as DMARDs (disease-modifying anti-rheumatic drugs) including methotrexate (MTX) and tumour necrosis factor inhibitors (TNFis) [468,469]. MTX is the most common treatment, and usually the first option tried [469]. Some patients do not respond to DMARD therapy or cannot tolerate MTX. Many studies have investigated the potential of tocilizumab as a treatment for these patients, either alone or in combination with MTX (e.g. [125,126,465,470]).

The effects of anti-rheumatic drugs are measured in various ways. Amongst the most common outcome measures are ACR20, ACR50 and ACR70 (whether the patient experienced ≥20%, ≥50% and ≥70% improvement respectively on the American College of Rheumatology scale [267]), swollen and tender joint counts, the DAS-28 score (the Disease Activity Score at 28 joints [471]), and the EULAR response criterion (categorising responsiveness to treatment as good, moderate or none based on the DAS-28 score and the change in scores [472]). Laboratory tests such as C-reactive protein levels (CRP) and erythrocyte sedimentation rate (ESR) are also used as measures of inflammation (e.g. [125,126]). 'Holistic' measures of quality of life such as Health Assessment Questionnaires (HAQs) and impact-of-disease assessments on a Visual Analogue Scale (VAS) are also used (e.g. [125,126]). Several standards measure whether the disease is in remission. DAS-28 Remission is a threshold criterion, defining remission as DAS-28 < 2.6, with a threshold of DAS-28 ≤ 3.2 used to define low disease activity [473]. The ACR-EULAR Boolean criterion is a stricter approach (see [474]) which defines remission as simultaneously satisfying *all* of: at most 1 tender joint, at most 1 swollen joint, CRP level < 1mg/dL, and patient global assessment of disease activity ≤ 1 on a 10-point scale [475].

Suppose a patient presents with chronic RA and a history of inadequate response to standard DMARDs. Her clinician wants to know whether to recommend tocilizumab—alone, or combined with MTX—for this patient. She consults the literature, and finds a number of studies, including randomised controlled trials such as RADIATE [465] and Maini et al.'s trial [126], and observational studies by Forsblad-d'Elia et al. [125] and Yoshida et al. [470]. Assume for the sake of this case-study that these four studies will form the evidence base for her decision.

RADIATE [465] is a multi-centre three-arm double-blind RCT, assigning 499 patients to receive monthly intravenous 8mg/kg or 4mg/kg of tocilizumab or a placebo, in combination with MTX. Participants were RA patients who had been unsuccessfully treated with TNFi therapy, and did not suffer from any related co-morbidities. The primary outcome measure was ACR20 after 24 weeks, with secondary measures including ACR20 after 4 weeks, ACR50, ACR70, EULAR response, DAS-28, CRP and ESR. On each outcome measure, the study found statistically significant differences between the treatment and placebo groups.

Maini et al. [126] conducted a 7-arm RCT, enrolling 359 patients, randomised into 2mg/kg, 4mg/kg and 8mg/kg tocilizumab groups with and without MTX, and a placebo group. Participants had active RA and had experienced inadequate responses to MTX. The primary outcome measure was ACR20 after 16 weeks, with secondary measures of ACR50, ACR70, swollen and tender joint counts, physician and patient assessments of disease activity, pain scores, ESR, CRP, HAQ, DAS-28 and duration of early-morning stiffness. Measurements were taken at 4, 8, 12, 16 and 20 weeks. The primary endpoint found significant differences between both 4mg/kg and 8mg/kg groups and the control, but the 2mg/kg group was only significantly different from the placebo when combined with MTX. The study does not report comparisons between the tocilizumab and combination therapy arms. In terms of the secondary outcomes, only the 8mg/kg combination group had a significant difference from the control on the ACR50 and ACR70 measures, while both 8mg/kg groups showed significant reduction in DAS-28. For most of the other outcomes, all groups showed a reduction in symptoms, but the placebo group outperformed the groups treated only with tocilizumab.

Forsblad-d'Elia et al.'s observational study [125] used data from the Swedish biologics register to analyse the drug adherence and response of 530 chronic RA patients. The study was designed to test for predictors of drug adherence and responsiveness to tocilizumab. Using a number of analyses, they identify low initial CRP level, high exposure to previous treatments and high initial HAQ assessments as predictors of discontinuation of treatment, and high disease activity and low initial HAQ assessments as predictors of responsiveness. They record that 52% of patients reached low disease activity levels, and 37% reached remission. 80% and 34.9% of patients displayed a 'moderate' or 'good' EULAR response respectively under intention-to-treat analysis. The study population was broad in terms of age, gender, presentation and co-morbidities.

Finally, Yoshida et al.'s cohort study [470] analysed 247 patients in one registry in Japan. They compared tocilizumab with TNFis in terms of effectiveness, safety and drug survival. The primary outcome measures were remission rates measured according to DAS-28 and the ACR-EULAR Boolean remission criterion, after 6 months. They found that remission rates amongst tocilizumab patients were more than double that of TNFis on the DAS-28 measure, but nil on the Boolean criterion, as the tocilizumab patients all had higher swollen joint counts than the threshold for the Boolean criterion. Safety and drug survival profiles were very similar.

The clinician reads the abstracts of these studies and turns to a hierarchy of evidence to decide which to read further in her limited available time, and how to appraise the evidence each provides. The relative rankings almost all agree that the RCTs provide higher quality or stronger evidence than the observational studies. It is unclear whether the hierarchies are meant to imply that the two RCTs and the two observational studies provide the *same* quality or strength evidence as one another. In the next section, these examples will be used to discuss these rankings and argue that they provide information which is either too strong or unhelpful. The important information about the studies for the subsequent discussion is summarised in the table below (Fig. 33).

Study	Methodology	Comparison	Population	Primary outcome measure	Secondary outcome measures
RADIATE [465]	3-arm RCT	8mg/kg Tocilizumab vs. 4mg/kg vs. Placebo—all in combination with MTX	499 North American and European patients with inadequate response to TNFis and no serious co-morbidities	ACR20 at 24 weeks	ACR20 at 4 weeks; ACR50; ACR70; EULAR response; DAS-28; CRP; ESR
Maini et al. [126]	7-arm RCT	8, 4 & 2mg/kg Tocilizumab, with and without MTX, vs. Placebo	359 European patients with inadequate response to MTX	ACR20 at 16 weeks	ACR50; ACR70; swollen and tender joint counts; physician and patient assessments of disease activity; pain scores; ESR; CRP; HAQ; DAS-28; all at 4, 8, 12, 16 & 20 weeks.
Forsblad-D'Elia et al. [125]	Observational study (outcomes research)	Patients who continue vs. discontinue treatment; Patients showing high vs. low EULAR responsiveness	All 530 tocilizumab patients with data recorded in the Swedish biologics register	Drug adherence and EULAR responsiveness over longitudinal timeframe	Low disease activity (LDR) rate; remission rate; HAQ; CRP; ESR; DAS-28; ACR20, ACR50, ACR70.
Yoshida et al. [470]	Cohort study	Tocilizumab vs. TNFis	All 255 patients starting TNFi or Tocilizumab treatment at the Kameda Institute in Japan in 2003-2010.	Remission rates according to DAS-28 and ACR-EULAR Boolean criteria after 6 months	Safety (serious adverse event rate); drug survival.

Fig.33: Table reporting study methodology, comparison, patient population, and primary and secondary outcome measures for four studies of tocilizumab as a treatment for rheumatoid arthritis where DMARDs have been ineffective.

5.2: Interpreting Relative Rankings

Individual studies provide evidence for a range of claims, not just the central claim the study is designed to test. This creates an issue in the interpretation of relative rankings. Suppose the clinician applies a relative ranking like Guyatt et al.'s 2002 hierarchy (see Figure 1, above) to the four studies she has identified.

RADIATE and Maini et al.'s studies are randomized trials, so outrank Forsblad-d'Elia et al. and Yoshida et al.'s observational studies. According to this hierarchy, then, RADIATE and Maini et al. provide stronger evidence for prevention and treatment decisions than do Forsblad-d'Elia et al. and Yoshida et al.'s studies.[83] If the clinician used Greenhalgh's influential 1997 hierarchy [43], she would find very similar results. Her hierarchy ranks RCTs above cohort studies (though omits outcomes research), so the clinician can at least infer: 'Evidence from RADIATE and Maini et al.'s trials should be given greater weight than evidence from Yoshida et al.'s study when making decisions about clinical interventions.'[84]

The clinician infers that the RCT evidence has more weight in her decision about which treatment to recommend to her patient than the observational study evidence. But none of these studies bear directly on the decision of whether to recommend tocilizumab to her patient. Each study provides a range of information relevant to her decision.

For example, RADIATE [465] is designed to test whether MTX plus 8mg/kg or 4mg/kg of tocilizumab monthly is more effective than MTX plus placebo at treating RA. RADIATE was designed to test whether there is a statistically significant difference in ACR20 rate after 24 weeks between the 8mg/kg tocilizumab group and the control group. The study establishes that there is a significant difference. It thereby provides evidence for the causal interpretation of that information given by the authors—namely that tocilizumab *causes* an increase in ACR20 rate in comparison to placebo.

But RADIATE provides evidence for other claims, for example about the effects of tocilizumab on the secondary outcome measures like ACR50 and ACR70, and about tocilizumab's effectiveness in 4

[83] This is subject to a caveat which downgrades poorly conducted RCTs and upgrades observational studies showing dramatic effects which does not apply in this case—"*If treatment effects are sufficiently large and consistent, carefully conducted observational studies may provide more compelling evidence than poorly conducted RCTs*." ([8], p.12)

[84] Again, this is subject to a caveat about evidence quality: "*not even the most hard line protagonist of evidence based medicine would place a sloppy meta-analysis or a randomised controlled trial that was seriously methodologically flawed above a large, well designed cohort study*." ([43], p.54)

weeks as well as 24 weeks.[85] Moreover, these outcome measures were used as a proxy for the outcomes in which patients and practitioners are really interested—pain, quality of life and disease burden. RADIATE provides some evidence that tocilizumab decreased the average pain levels and increased quality of life compared to placebo in this group of RA patients.

There are also generalisations beyond the study population for which RADIATE provides evidence. It provides evidence for the claim that the average effect will be similar in the population of all chronic RA patients who respond inadequately to TNFis. The evidence for this claim is weaker than the evidence for the internal claims, but RADIATE still provides some confirmation for this generalisation. One can generalise to other populations too—there is some confirmation here that *all* RA patients would benefit from tocilizumab more than placebo, for instance. The study also provides some evidence concerning dose-response, giving some confirmation to the hypothesis that an 8mg/kg dose has a larger effect on ACR20 rates than a 4mg/kg dose.

Some studies also provide evidence for claims about subgroups of the study population. For instance, Forsblad-d'Elia's observational study [125] provides evidence for the claim that patients already treated with DMARDs are no less likely to continue using Tocilizumab, and that patients with high HAQ scores are less likely to experience clinically significant benefits from tocilizumab. These subgroup analyses can again be generalised. While many authors have concerns about the power, rigour and potential for exploitation in subgroup analyses (e.g. [300,477,478]), these studies nevertheless provide *at least some* evidence for claims about specific subgroups.

In addition, studies confirm specific claims about the effects upon individual patients in the study and predictions about the effects on individuals in the future, and about the frequency and incidence of desirable and adverse events in the study. For instance, Forsblad-d'Elia's study shows 37% of the patients studied achieved remission, information which provides evidence for a prediction of similar remission rates in target populations. The importance of such specific claims to clinicians and patients should not be overlooked—there is great value in decision-making to the ability to inform a patient that on average the treatment leads to remission in 1/3 of patients, for instance. This information may be far more important to patients in decision-making than information about the treatment's ability to outperform placebo on ACR20 measures.

There are many claims for which each study provides evidence. It follows that if a hierarchy implies, say, that 'RADIATE provides stronger evidence than Yoshida et al.'s study', then one should ask *for which claims* RADIATE provides stronger evidence. Both Guyatt et al. and Greenhalgh's

[85] Zarin et al.'s [476] analysis of the ClinicalTrials.gov database shows that large numbers of secondary outcomes are often measured in studies. They found a median of 3 secondary outcome measured in RCTs, with up to 122 secondary outcomes in some trials.

hierarchies delineate the scope of the quantifier as ‘when making treatment decisions’. But it is not clear whether they mean this to apply to *all claims* relevant to the treatment decision, or to some *specific claim*—e.g. ‘Tocilizumab is best treatment option for this patient’, or ‘Tocilizumab is an effective treatment for patients similar to this patient.’

5.3: "All Claims" Interpretations are Too Strong

One way of reading relative rankings like those of Guyatt et al. [8] and Greenhalgh [43] is as implying that *for all claims* for which the studies provide evidence relevant to the clinical decision, RCTs provide stronger evidence than observational studies. But this claim is far too strong. The strength and quality of evidence for a claim depends both on the methodology *and* the relevance of the study to the claim. For example, finding that tocilizumab is more effective than placebo on ACR20 provides some weak, indirect confirmation for the hypothesis that tocilizumab is more effective than TNFis—it shows that tocilizumab is efficacious, which surely makes the notion of outperforming TNFis more plausible.[86] But Yoshida et al.'s study which directly compares tocilizumab to TNFis clearly provides stronger evidence for this claim. Similarly, RADIATE shows that patients who had inadequate responses to TNFis benefitted from tocilizumab which provides some support for the hypothesis that previous exposure to TNFis is not a predictor of low responsiveness to tocilizumab. But Forsblad-d'Elia et al.'s study provides much stronger confirmation for that hypothesis.

In general, the justification for ranking RCT evidence above observational study evidence breaks down when evidence the RCT provides for claims other than the main claim under test is included (see [11,97]). EBM proponents accept that the advantages of an RCT designed to test a specific hypothesis does not carry over to analyses of secondary outcomes, subgroup analyses, or inferences from the result such as generalizations (see [11,299,477,478]). The advantages of RCT design which even the staunchest critics concede also do not necessarily apply here (see [11,52]). Therefore, there is no philosophical justification for an "all claims" reading of relative rankings, and these readings will in practice provide false information.

Nevertheless, some authors of hierarchies have defended such a reading. In particular, Roman, Silberzweig & Siu [299] criticised the tendency to lump all of the evidence from a study into the same category in a hierarchy, and provided a hierarchy which differentiates between the main result of an RCT, an inference from that result about one component of the intervention[87], and generalizations to other populations or treatment regimens. Their hierarchy does commit to the claim that in each of these cases, RCTs provide stronger evidence than observational studies (see Figure 34, below).

[86] Moreover, if there is another study which compares TNFis to placebo, and Maini et al.'s study shows that tocilizumab outperforms placebo by more than TNFis did in the other study, then this provides some limited evidence that tocilizumab is more effective than TNFis. However there are serious limitations to inter-study comparisons (see [255,479,480]). Populations and treatment regimens, including placebo regimens, differ, as do outcome measures, follow-up periods, and sample size.

[87] For instance, in Maini et al.'s trial, there are various groups combining different doses of tocilizumab with or without MTX. The study provides some evidence that MTX augments the effects of tocilizumab as a number of the combination groups fared better than the groups taking solely tocilizumab. On Roman, Silberzweig & Siu's hierarchy, this evidence ranks lower than the evidence for the main claim under test.

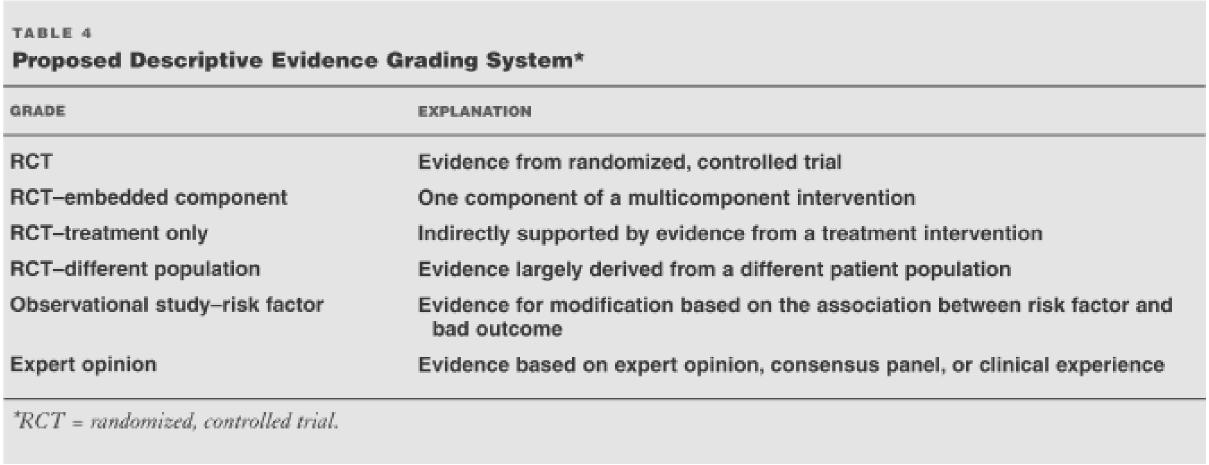


TABLE 4
Proposed Descriptive Evidence Grading System*

GRADE	EXPLANATION
RCT	Evidence from randomized, controlled trial
RCT–embedded component	One component of a multicomponent intervention
RCT–treatment only	Indirectly supported by evidence from a treatment intervention
RCT–different population	Evidence largely derived from a different patient population
Observational study–risk factor	Evidence for modification based on the association between risk factor and bad outcome
Expert opinion	Evidence based on expert opinion, consensus panel, or clinical experience

**RCT = randomized, controlled trial.*

Figure 34: Roman, Silberzweig & Siu's 2000 ([299], p.88) hierarchy ranking evidence other than the main result of an RCT (e.g. RADIATE, Maini et al.) above that from an observational study of risk factors (e.g. Forsblad-d'Elia et al.)

Similarly, McAlister et al.’s hierarchy [216] from the *Users’ guides* incorporates inferences across placebo-controlled RCTs (see n.4, above), inferences across subgroups of RCTs, and inferences from surrogate outcomes to patient-relevant outcomes within their more detailed hierarchy (see Figure 35, below). Again, their hierarchy places RCTs used as evidence for claims in these ways above observational studies. In this case, inferences from subgroups within RADIATE or Maini et al.’s trial would be classed as providing stronger evidence about the efficacy of tocilizumab than Forsblad-d’Elia et al.’s study, despite the latter being specifically designed to test hypotheses about certain subgroups, having larger sample sizes within these groups, and using sophisticated techniques to control for confounding factors, and despite the justification for ranking RCTs highly failing to apply to inferences from subgroup analysis.

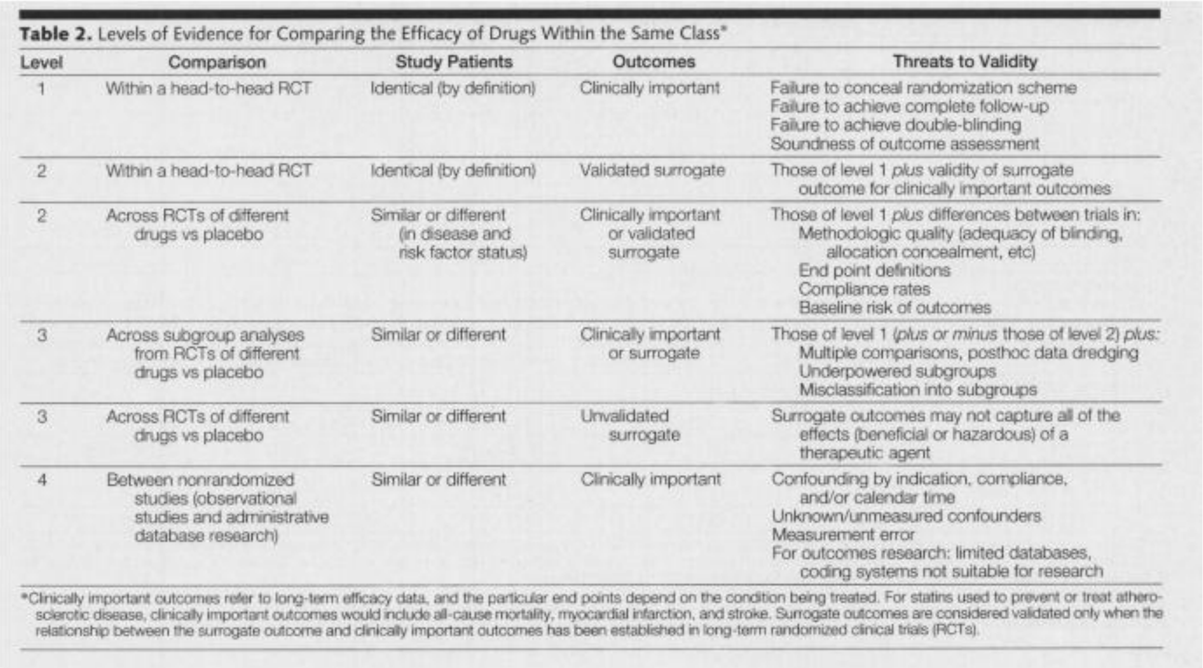


Table 2. Levels of Evidence for Comparing the Efficacy of Drugs Within the Same Class*

Level	Comparison	Study Patients	Outcomes	Threats to Validity
1	Within a head-to-head RCT	Identical (by definition)	Clinically important	Failure to conceal randomization scheme Failure to achieve complete follow-up Failure to achieve double-blinding Soundness of outcome assessment
2	Within a head-to-head RCT	Identical (by definition)	Validated surrogate	Those of level 1 *plus* validity of surrogate outcome for clinically important outcomes
2	Across RCTs of different drugs vs placebo	Similar or different (in disease and risk factor status)	Clinically important or validated surrogate	Those of level 1 *plus* differences between trials in: Methodologic quality (adequacy of blinding, allocation concealment, etc) End point definitions Compliance rates Baseline risk of outcomes
3	Across subgroup analyses from RCTs of different drugs vs placebo	Similar or different	Clinically important or surrogate	Those of level 1 (*plus or minus* those of level 2) *plus:* Multiple comparisons, posthoc data dredging Underpowered subgroups Misclassification into subgroups
3	Across RCTs of different drugs vs placebo	Similar or different	Unvalidated surrogate	Surrogate outcomes may not capture all of the effects (beneficial or hazardous) of a therapeutic agent
4	Between nonrandomized studies (observational studies and administrative database research)	Similar or different	Clinically important	Confounding by indication, compliance, and/or calendar time Unknown/unmeasured confounders Measurement error For outcomes research: limited databases, coding systems not suitable for research

*Clinically important outcomes refer to long-term efficacy data, and the particular end points depend on the condition being treated. For statins used to prevent or treat atherosclerotic disease, clinically important outcomes would include all-cause mortality, myocardial infarction, and stroke. Surrogate outcomes are considered validated only when the relationship between the surrogate outcome and clinically important outcomes has been established in long-term randomized clinical trials (RCTs).

Figure 35: McAlister et al.'s 1999 hierarchy from the *Users guides XIXb* ([216], p.1373).

5.4: "Main Claims" Interpretations are Not Useful

Adam La Caze argues that interpretations of relative rankings must be strictly limited to comparisons of the evidence the studies provide for the main claim[88] being tested. He states:

> "*the statistical techniques employed in analyzing randomized trials provide optimal warrant only to the primary hypothesis under test*" ([11], p.522).

He concludes that hierarchies only inform the reader that RCT evidence for the primary hypothesis the RCT tests has higher internal validity[89] than does evidence from an observational study for that claim.[90]

However, this approach is extremely limiting. Unless the observational studies in question are also designed to test the same primary hypothesis as the RCT, this information seems obvious. A study designed to test a specific hypothesis is more able to safeguard against biases in the testing of that hypothesis. Every well-designed study is designed to eliminate biases in its testing of the main hypothesis. Assuming a well-designed study, it is hard to justify denying the reverse claim—that an observational study designed to test hypothesis A probably has higher internal validity than an RCT designed to test hypothesis B but which provides some indirect evidence for A. In the case of the tocilizumab studies, it seems clear that RADIATE provides stronger evidence that tocilizumab is more effective than placebo in patients for whom TNFis have been ineffective than do Forsblad-d'Elia et al. and Yoshida et al.'s studies, even though both do provide some evidence to that end. But it equally seems obvious that for instance Yoshida et al.'s study provides stronger, more valid evidence that tocilizumab is more effective at increasing remission rates than TNFis are than either RADIATE or Maini et al.'s trial.[91]

"Main claims" readings only offer any useful information where the main claim of two studies is the same (or contradictory). Otherwise, if the study is well-designed, it should be expected that a

[88] Note that there is not always a unique primary comparison being tested. Analysis of the ClinicalTrials.gov database by Zarin et al. [476] showed that the median primary outcome measures was 1, but that some trials have as many as 71 outcomes pre-specified as primary.

[89] Unlike Guyatt et al. [8] and Greenhalgh [43], La Caze specifies internal validity, not strength of evidence, as the interpretation property for hierarchies. He claims that this is the only property for which a hierarchical ranking can be justified. The points made here apply to either interpretation property.

[90] Again, this is subject to conditions—that the RCT must be properly performed and all the necessary techniques used to prevent biases, including properly randomized allocation, double-blinding, tests to ensure blind is not broken, intention-to-treat analysis and checking for baseline imbalances between the comparison groups ([11], p.522).

[91] Additionally, it seems strange to omit these rankings where studies have the same design. Clearly, RADIATE provides stronger, higher validity evidence that tocilizumab is more effective than placebo in patients non-responsive to TNFis than Maini et al.'s trial does, which focuses on patients not responsive to MTX—and vice versa. But hierarchies would place these on the same level of evidence for each claim.

study which tests the claim of interest will probably provide stronger evidence than one which provides indirect evidence.[92] But two studies will very rarely test the same main hypothesis, especially where methodology is different. It is rare to have two studies testing the same main claim. Even two very similar studies like RADIATE and Maini et al.'s trial test considerably different claims. Unless a broad reading of the 'main hypothesis' which a study is designed to test is used, studies would have to be performed on the same cohort of patients to count for this kind of comparison.

With a broader reading of 'main hypothesis', studies on populations with similar features—the same age range and distribution of ages, gender, etc.—might count. But there are a great many ways in which studies differ which affect the hypotheses tested. In many trials, the precise treatment regimen used is extremely carefully delineated (*cf.* [255,443,444]). If these regimens differ between studies, despite both testing the same core experimental treatment, the main claims are different as they test different regimens.[93] The control treatment can similarly vary. There is even great variability in the use of placebo treatments, and how closely they match the administration of the experimental treatment (see [430]). The same goes for the length of follow-up of patients. Unless a study is designed as a replication of an existing trial, it seems unlikely that main claims will match even on a loose set of similarity requirements. Although replication is a sound scientific approach, attempted replications will use the same study methodology, and so too will not be amenable to analysis via hierarchies.

La Caze offers another reading of relative rankings in terms of counterfactuals—he claims that the hierarchy reveals what kind of study *would provide* the most valid evidence about a specific claim. This is a form of relative ranking, where the comparator is all other studies of the question. He claims:

> "*EBM's hierarchy provides clear advice. The most accurate method to measure the efficacy of Drug X, all other things being equal, is via a well-conducted randomized controlled trial*" ([11], p.524).

However, again this information is not particularly helpful. First, this account flattens the landscape of clinical trials dramatically. In reality, trials do not simply *"measure the efficacy"* of a drug—rather, they measure its effects on some endpoint in some population relative to some control intervention(s). Simply locating an RCT performed on a drug will not provide information about all the effects of the drug, nor in all populations or against all comparators. Except in the rare circumstance that an RCT addresses precisely the same main claim as an observational study, the only information provided by

[92] There are obvious exceptions to this rule, hence the probabilistic modifier. For example, sample size may be a factor: a study designed to compare tocilizumab to TNFis in 30 patients will probably provide weaker, lower quality evidence of the effectiveness of tocilizumab than a less directly applicable study enrolling thousands of patients.

[93]These small differences should not be overlooked—many trials are devoted to questions of which of a number of similar regimens is more effective. Much of the value of Maini et al.'s trial is in attempting to ascertain the best dosage for tocilizumab. There is also variation in how carefully controlled the regimens are—for instance, while both RCTs control the dosage and frequency of both tocilizumab and MTX, both observational studies allow variation in dosage and frequency.

this account is the counterfactual: '*If* this observational study had been an RCT, *then* it would have provided more accurate evidence about its main claim'.

This information is not useful for practitioners. It provides no way of estimating the reliability of the evidence the study actually provides. Furthermore, the counterfactual claims sometimes seem patently false. While Yoshida et al.'s study could plausibly have been designed as an RCT, in the case of Forsblad-d'Elia's study, it seems impossible to imagine redesigning the study as a workable RCT. One cannot randomize some patients to have higher or lower initial HAQs or disease activity. It is not clear that research into the causes of variation in responsiveness and continuation of treatment like that performed by Forsblad-d'Elia et al. *can* be performed as RCTs (see Chapter 6 for further discussion). It seems strange to suggest that the *highest* standard of evidence for claims about causes and predictors of responsiveness and treatment continuation cannot be reached. If RCT methodology cannot be applied to test certain claims, then not every non-randomized study could be improved by redesigning it as an RCT.

The only useful consequence of this account seems to be advice to researchers—if you want your trial to provide the most reliable evidence, then you should design it as an RCT. However, in the light of the challenge of whether RCTs are always available to test interesting medical claims, this advice is undermined. Furthermore, EBM and the hierarchies of evidence associated with EBM are not intended to guide researchers but to guide the application of existing evidence to clinical practice, so this advice is not pertinent to EBM's goals.

5.5: Other Interpretations are also Unhelpful

One can envision alternative readings being offered. In particular, some authors focus the EBM approach specifically upon the strength and quality of evidence for ‘effectiveness’ claims—generalizations of estimates of treatment effects to the target population (*cf.* [339]). Casting the clinician’s problem in these terms translates it as ‘Which studies provide the strongest evidence for the claim that tocilizumab is (or is not) an effective treatment for RA in patients similar to this patient?’ Alternatively, one could jump directly to the clinical decision problem as Guyatt et al. [8,82] and Greenhalgh [43] do, and cast the question as: ‘Which studies provide the strongest evidence for the claim that the clinician should (or should not) recommend tocilizumab for this patient?’

When represented in this way, the relative rankings are localized to the effectiveness claim or the ‘best recommendation’ claim. In other words, ranking RADIATE higher than Yoshida et al.’s study entails that RADIATE provides stronger evidence for the claim that tocilizumab is an effective treatment, or that the clinician should recommend tocilizumab, than does Yoshida et al.’s study.

However, this information is again not helpful for practitioners. What practitioners need is an estimate of the strength or quality of the evidence each study provides for such a claim, so that they can determine which treatment to recommend and how strongly to recommend it. Knowing that the evidence from RADIATE is *stronger* than that from Yoshida et al.’s study for these specific claims does not provide this kind of information. This ranking is consistent with both studies providing very strong evidence, with RADIATE’s marginally stronger, and with both providing very weak evidence, with Yoshida et al.’s marginally weaker. In practice, a clinician would want to know how strong and how reliable the evidence from each study is, and then combine the evidence from each.

These relative rankings are also dubious. Inferring from even very strong, high-quality evidence directly to a recommendation is very problematic, as Chapter 6 will argue. Even the inference from the results of a study to an effectiveness claim varies in strength between studies. RADIATE and Maini et al.’s trials provide stronger or weaker evidence depending upon the population in which the effectiveness claim is based. If the population is composed of patients for whom TNFis have been ineffective, RADIATE provides the stronger evidence, whereas if it is composed of low responders to MTX, Maini et al.’s study provides stronger evidence. In a mixed population, it is unclear how well each study supports the effectiveness claim. Moreover, Chapter 6 will emphasise that effectiveness claims and recommendations really depend on the *balance* of harms and benefits. Even if RCTs provide the strongest evidence about the benefits of tocilizumab, they may not be good approaches to the study of harms and risks, as La Caze [11,72] and Rawlins [97] argue. The claim that RCTs provide stronger or higher quality evidence for an effectiveness claim than do other sources of evidence is at best unhelpful, at worst false.

5.6: Conclusion

In summary, the EBM literature has been unclear as to how to understand relative rankings. This lack of clarity stems from the underappreciated fact that studies provide evidence for a variety of claims, and that any comparison of the strength, quality or validity of evidence provided by studies must specify *for which claim(s)* the comparison holds. Broad readings will be far too strong. Narrow readings make hierarchies so rarely applicable as to be clinically unhelpful, or yield false or unjustified claims. No attempt to clarify the interpretation of relative rankings has provided a way of generating usable information for practitioners. Therefore, restricting the interpretation of hierarchies to relative rankings of any kind is no solution to the problems raised in chapter 4, nor is it a viable approach the use of hierarchies in EBM. To be clinically useful, hierarchies will have to offer guidance in inferring the absolute strength or quality of evidence, or the absolute level of evidence in an evidence base.

CHAPTER SIX:

EVIDENCE AND EVIDENCE BASES:

CHALLENGES TO GRADE AND LEVELS OF EVIDENCE

CONTENTS

This chapter argues that although GRADE offers the most sophisticated current hierarchical approach to evidence appraisal, the GRADE approach omits important information. Evidence about the variation in treatment effects and the causes of different effects is not considered. This information is vital to a full appraisal of the evidence base for clinical claims.

First, it is demonstrated that simpler approaches than GRADE are inadequate. A hierarchical model must appraise the evidence-base as a whole rather than study-by-study to provide an accurate appraisal. Furthermore, approaches which (unlike GRADE) equate strength of recommendation with level of evidence are flawed. The ways in which the GRADE approach avoids these problems are then outlined.

However, this chapter will then argue that GRADE focuses primarily on evidence for estimates of average treatment effects. But information about variation and the causes of variation in treatment effects ("heterogeneity of treatment effects" [457,481,482]) is vital to clinical recommendations and decision-making. GRADE offers no solution to appraising and integrating evidence of variation. Four case-studies from a range of medical fields establish this claim. Solutions to this problem which employ hierarchical tools are discussed, but are inadequate. Finally, this problem is not solved by limiting the scope of application of hierarchies to health policy and public health, excluding questions of patient care. The same problem also affects policy and population-level decisions. Therefore, although GRADE avoids or could be reformulated to meet the challenges posed to hierarchies to date, the system is not an ideal method for appraising evidence bases in clinical medicine. Systems which take account of the distribution of effects and the causes of variation in treatment effects are preferable, and such systems are not primarily hierarchical.

6.1: Atomistic Approaches to Appraisal

Some hierarchies take an atomistic approach to evidence appraisal. The evidence from a study for a claim is appraised in isolation from other studies. This section argues that atomistic appraisal is inadequate in the medical context. Evidence from different studies interacts in a number of ways. Approaches which appraise the evidence from a study in isolation cannot draw upon information about that evidence from other studies—meta-evidence which affects judgements of the strength and quality of that evidence.

One can distinguish several versions of atomism. The simplest version—call it strong atomism—assumes that the quality and/or strength of the evidence a study provides for a hypothesis can be determined by looking at that study alone. Many basic hierarchies seem to enshrine this principle. Particularly, absolute unconditional hierarchies imply that knowledge of the methodology of a study alone can be used to determine the quality or strength of the evidence provided by that study. Assuming that the methodology can be determined from the study alone, this unconditional approach permits strong atomism in the appraisal of medical evidence.

This section argues that strong atomism is indefensible. Recall that the definitions of strength and quality of evidence referred to justifiable confidence in evidence, in terms of internal validity, external validity and clinical relevance (see Section 2.1). Information from one study can affect an agent's confidence in each of these features of the evidence from another study. Therefore, evidence from one study can affect the strength or quality of the evidence from another study. As such, the quality and strength of evidence cannot be determined from the study alone.

6.1.1: Internal validity

Consider two studies, A and B. There are many ways in which the evidence from B can affect the quality of the evidence from A for some hypothesis. First, B could provide information about the internal validity of A which would cause a reasonable agent to increase or decrease their confidence that the results of A are accurate, and in a causal interpretation of those results.

If A and B test for the same correlation, then their results will affect confidence in each others' findings. If B's estimate of the size and direction of the effect is similar to A's, then this successful replication increases confidence in the accuracy of the effect measured by A. Bias or spurious correlation is a less plausible explanation of the finding in A, given that B replicated the result. So the evidence from B increases the quality of the evidence provided by A. This is especially true if A and B were undertaken by independent investigators. Independent replication could also be helpful in ruling out the effects of conflicts of interest. In particular, meta-epidemiological evidence suggests that studies performed or sponsored by the pharmaceutical industry are more likely to show statistically significant benefits (see [244,415,416,483], and Section 4.2.2). An independent replication by a party without conflicts of interest is evidence that the study-result was not affected by commercial bias or experimenter bias, and thus increases justifiable confidence in the validity of the result.

By contrast, if B's findings disagree with A's, then confidence in each result may be diminished. Inconsistent results provide a reason to suspect that at least one of the studies is affected by bias. Disagreement amongst study-results is usually referred to as "heterogeneous results" (see e.g. [1,190,197]). Systematic reviewers and meta-analysts use techniques such as meta-regression to analyse heterogeneous results, testing for correlations between the properties of studies and their results to attempt to explain the variation in results ([1,484] and see below for further discussion). In the absence of such analysis, variation undermines confidence in the results due to the possibility of bias.

B may act as a defeater for A, demonstrating that A is invalid. For example, suppose that A compares treatments for rheumatoid arthritis. A shows that the experimental treatment outperformed the control. However, B presents outcomes research showing that rheumatoid arthritis patients with gene X fare much worse given either the control or the experimental treatment from A than do patients without X. If re-analysis shows that Trial A was imbalanced with respect to gene X (i.e. that the experimental group had a smaller proportion of X-positive patients than the control group), then the trial is confounded on the previously unknown confounder variable X, and the results are invalid. The rival hypothesis that patients in the experimental group did better than the control as a result of differences in X explains the findings in A without recourse to the claim that the experimental treatment has a greater effect than the control. The evidence from B (coupled with the re-analysis of A) provides evidence against the claim that A provides high-quality evidence.[94]

Finally, evidence that a particular bias was not successfully eliminated would be evidence against the validity of the evidence from a study. There are tests to check whether blinding was successful in a clinical trial.[95] If these tests show that blinding was broken, then confidence that the study has eliminated biases due to differences in expectations amongst the patients will be lower. Similarly, if the researchers' blind was broken, researchers may have treated patients in different groups differently, causing a treatment bias. More generally, evidence that blind is often broken [249,409,410], that randomisation is often unsuccessful or subverted [403], or that experimenters tend to (consciously or unconsciously) manipulate results [236] may provide reason to decrease confidence in the validity of study results, especially where studies do not provide details of how

[94] See the discussion of Temozolomide and the MGMT gene in Section 6.4.3.3 for a somewhat similar case.

[95] David Sackett [485-487] argued against testing whether blinding was successful in RCTs—his influence led to the removal of testing for successful blinding from the CONSORT quality criteria for reporting of RCTs [154,316].

blinding and randomisation was ensured (*cf.* [153,154]). Evidence that blinding was successful and that randomisation was not manipulated would provide evidence for the claim that the trial provides high-quality evidence.

6.1.2: EXTERNAL VALIDITY

There are many concerns which may decrease an agent's confidence in the generalisability of a result to some target population. These include concerns about whether the treatment will be as effective in patients with different features, different presentations of the disease, and in different settings. Similarly, the treatment may have more severe or more prevalent side-effects in a different population. The same could be true of control treatments.

B could provide evidence which ameliorates or enhances these concerns about the evidence from A. If B shows a similar effect to A in a different population, this may increase confidence in the generalisability of A (both to that population, and to wider populations as B provides some evidence that the effects of the treatments are not limited solely to the population tested in A, and some (albeit weak) evidence that the treatment effects are fairly consistent). Researchers often distinguish between 'explanatory' and 'pragmatic' studies [255,262,263,459]. Explanatory studies are performed in narrow, well-defined populations, under controlled conditions. Pragmatic studies are performed in broader populations under normal clinical conditions. If a pragmatic study finds a similar result to an explanatory study, then this supports the claim that the explanatory study's finding generalises to normal clinical populations and circumstances.

B could also undermine a generalisation from A's results. Clearly, if B's results in Population B differ from what would be expected when generalising A's results in Population A to Population B, then this is evidence that A's result does not generalise to Population B (and by extension, that it is not more broadly generalisable). This counter-evidence could be more remote—for instance, B could show that the effects of a *similar drug* to that tested by A failed to generalise to Population B, which would give clinicians reason to suspect that the effect found in A may also not generalise well to Population B. This evidence could also be mechanistic—focusing on the causal process of the treatment tested by A. If the causal process underlying the effect in A is known, or believed to be known, then if B provides evidence that some feature of patients in Population B will interfere with that causal process, inhibiting it, then B is evidence that the effect found in A may not generalise to Population B.

6.1.3: *Clinical relevance*

The evidence from B could support the clinical relevance of the evidence from A in several ways. As discussed in Chapter 2, in order to be clinically relevant, a study must provide information about a treatment's effects on outcomes that matter to the patient ("patient-oriented outcomes", in the EBM literature [127,128]). If A measures the effect of a treatment on a surrogate outcome (for instance, on a physiological measure of joint swelling in arthritis like C-reactive protein levels, rather than on patient's actual mobility and quality of life), then a study B which validates the surrogate outcome as a reliable measure of the patient-oriented outcome is evidence in favour of the clinical relevance of the evidence from A. If B shows that the surrogate outcome does not reliably track the patient-oriented outcome, then B is evidence against the clinical relevance of A. So, B could increase or decrease the quality of A.

6.1.4: *Exercise on Prescription—an example of evidence interacting*

In 2008, Lawton et al. conducted a study into the effect of exercise on prescription on the amount of physical activity reported by physically-inactive women aged 40-74 [488]. The evidence from this study provides an interesting case in which other evidence from within the trial and from other studies affects the quality of the evidence from the trial. In particular, the clinical significance of the evidence is undermined.

Lawton et al. recruited a cohort of 1,089 women aged 40-74 who reported low levels of physical activity, and randomized them to receive either a physical activity intervention, or usual care. The physical activity intervention consisted of a so-called "green prescription"—a short (7-15 minutes) counseling session led by a nurse according to a 'green script' which focuses on motivation, choice of activity and overcoming barriers to physical activity. The script is oriented to motivate patients to undertake a minimum of 30 minutes moderate physical activity, five days per week. In addition to the counseling session, intervention patients received phone calls over the following nine months (lasting 15 minutes with an average of 5 calls per patient), and a follow-up visit after six months consisting of a 30-minute meeting with the nurse in which the nurse would ask whether the patient had increased her physical activity, as well as providing further motivation and measuring secondary outcome measures such as blood-pressure, weight and waist-circumference. 'Control' patients received no additional intervention, and their outcomes were simply monitored.

The primary outcome measure for the study was self-reported level of physical activity, as ascertained by long-form physical activity questionnaire. To give a binary measure of physical activity, the threshold of 150 minutes moderate physical activity in the last week was used. The claim which the study is designed to test is:

(1) 'Exercise prescription increases the reported physical activity of low-activity women aged 40-74.'

The study found a significant increase in the reported physical activity of *both* the control and intervention group compared to the baseline level measured at the start of the study, and a significant differential increase—the patients in the intervention group were considerably more likely to achieve the desired 150-minutes level than the patients in the control group after 12 and 24 months. Lawton et al. claim that this result provides strong evidence for (1).

Claim (1) alone is of almost no interest to practicing clinicians—it has very low clinical significance. Clinicians are not interested in increasing how much physical activity female patients *report* doing. They may be interested, however, in increasing patients' actual level of physical activity—but primarily as a proxy for other health benefits. Reported physical activity, the investigators hope, will provide a proxy measure for actual physical activity. In turn, actual physical activity, they hope, will be a contributing cause in improved health outcomes such as quality of life, mobility and probability of adverse events such as infarction.

They measured a number of secondary outcomes, including quality of life, weight, waist-circumference, blood pressure, concentrations of fasting serum lipids, glycated haemoglobin, glucose, and insulin in blood samples, and physical performance in a three-minute step test. They found no statistically significant differences between the control and intervention groups on these secondary outcome measures.

There are two claims that might be interesting to clinicians for which this study *might* provide some evidential support:

(2) Exercise on prescription increases the physical activity levels of previously-inactive women aged 40-74.

(3) Exercise on prescription is beneficial to the health of previously-inactive women aged 40-74.

The study does not provide good evidence for either claim. Crucially, showing this relies on other evidence from within the study and from other studies. To make an evidence-based inference to (2) or (3) on the basis of these study-results would require additional premise

(4) Women aged 40-74 reporting higher levels of physical activity in this trial is good evidence that those women engaged in higher levels of physical activity as a result of the intervention

Claim (4) is very difficult to defend. This trial involves a very clear example of potential for *observer-expectancy bias*. Observer-expectancy bias occurs when the observer (in this case, the investigators in the form of the nurse) communicates the expected outcome to the patient. When patients know what the investigators expect to see, they tend to conform to that expectation [36,231-233]. This is an especially potent effect where the expected outcome is something that the patient can control [36]. In this case, the patient has direct and complete control over the primary outcome measure—their self-reported physical activity. The nurse has emphatically communicated the expectation of the study to the intervention group: the nurse wants them to engage in more physical activity. Due to the observer-expectancy bias, one expects that patients in the intervention will report higher levels of physical activity because they believe that they should.

Moreover, the trial is strongly exposed to the Hawthorne effect [231-233]—the phenomenon that people behave the way they think they are expected to behave more often and more explicitly when explicitly monitored. Just being told that they are in a trial and that their reported level of physical activity (along with weight, waist-circumference, etc.) will be monitored may lead patients to over-report their physical activity because they believe they are being assessed on it, and even to actually increase physical activity for the period of the trial, but crucially *not because of the intervention*, but because of the monitoring required to collect the trial data.[96] It may be the case that *even if* the patients are actually engaging in more activity and not simply reporting that they are, it is not exercise prescription that has this effect, but the knowledge that the nurse will be assessing their level of activity and measuring their weight and waist-circumference. This is particularly salient due to trial inclusion criteria which identifies the women as abnormally low-activity individuals, and has a particularly strong psychological impulse in this case because of the negative connotations attached to weight and waist-circumference, especially amongst the target population demographic.

Moreover, this interpretation of Lawton et al.'s data is supported by their own secondary outcome measures. There were no significant improvements in secondary outcome measures which are believed to be correlated with levels of physical activity. If physical activity genuinely increased in the trial population, then changes in weight, waist-circumference, laboratory measures and physical fitness would be expected. That these changes were not observed provides support for the explanation of the findings through the combination of observer-expectancy and Hawthorne effects.

Note that what makes the evidence from Lawton et al.'s study weak and low-quality is a combination of internal and external evidence. The evidence for an observer-expectancy effect and Hawthorne effect provides an alternative explanation of the data which does not require any

[96] This is a component of the "trial effect" [234]—the phenomenon that patients generally do better in trials than in normal practice. For further discussion, see Section 6.2.2.

treatment effect for exercise on prescription. The evidence from secondary outcome measures in the study corroborates this interpretation, as does the observation that the control group's reported physical activity also increased.

Lawton et al. defend their choice of self-reported outcome measure by citing evidence that the long-form physical activity questionnaire used has been independently validated against both heart-rate monitoring and the IPAQ-Long (another self-reported questionnaire). This external evidence pushes against the defeaters. But ultimately it is unsuccessful as counter-evidence because these validations were not performed under relevantly similar conditions. In tests of the reliability of a measure, there is relatively little or no pressure on the patients to over-report their physical activity (and indeed if the patients are aware that the study is designed to test the accuracy of their self-reporting, they may report more reliably than usual due again to observer-expectancy effects). What is needed to support (a) is evidence that patients do not over-report physical activity when the investigators apply strong pressure to increase their reported physical activity. The psychological studies demonstrating observer-expectancy effects, coupled with background knowledge that low-physical activity women aged 40-74 are likely to be concerned about having their weight and waist-circumference regularly measured and recorded, and reinforced by the absence of change in secondary measures, undermines the link between reported activity and actual activity.

The internal validity of the evidence from Lawton et al.'s study is undermined by the evidence that the effect was not due to the treatment. External validity is undermined by this, plus the knowledge that the effect of observer-expectancy would be unlikely to continue outside the trial setting, where data was not being recorded. Clinical significance is further undermined by the weak links between reported activity and actual activity, and then actual activity and health. Appraising the evidence from Lawton et al.'s study in isolation would be an error. Internal and external evidence needs to be included in the appraisal, and itself assessed for quality, in order to provide a good estimate of the quality and strength of the evidence.

6.1.5: Summary

Evidence from B can affect the strength and quality of evidence from A. A full critical appraisal of the quality of evidence for an individual source requires the user to consider the effect of meta-evidence on confidence in the internal and external validity and clinical significance of the evidence. A simple atomistic hierarchical model cannot accommodate this. Not only do simple hierarchical models omit meta-evidence, they *cannot* include meta-evidence because to do so requires the user to first appraise the quality of that evidence. For instance, if B's result contradicts A's, then the extent to which this affects the estimate of the quality of A's evidence will depend on the quality of B's evidence. A large, well-conducted trial which contradicts A will have a stronger effect on the quality of A's evidence than does a small, poorly-conducted one. If B has itself been repeatedly contradicted by studies C, D and E, then its effects on A (and C, D and E) will likely be lower. Because of the inter-relationship between judgments of quality of evidence amongst medical studies, atomistic approaches to evidence cannot properly take account of meta-evidence. Therefore, atomistic approaches to evidence will overlook important information. A fully evidence-based approach cannot be implemented by appraising the evidence from individual studies in isolation.

6.2: Levels of Evidence

Proponents of hierarchies can respond to the problems of strong atomistic appraisal by moving to a *hierarchy of evidence bases* rather than a simple hierarchy of evidence. As examined in Chapter 2, evidence base hierarchies assess the strength or quality of an entire evidence base for a hypothesis. Users assess the evidence from multiple studies together. Hierarchies of evidence bases include GRADE, but there a great many others. A sufficiently sophisticated hierarchy of evidence bases should be able to take account of the interaction between multiple studies.

This approach has a number of important advantages. It could take account of the importance of replication as a method of ruling out certain biases, and evidence from studies in different populations as support for generalising from results. It can allow for an evidence base containing both pragmatic and explanatory trials to be higher quality than one solely composed of explanatory trials. Heterogeneity of study results could potentially be considered. However, the implementation of hierarchy of evidence bases is challenging. There are several ways that they could be designed and used.

6.2.1: Cumulative and Threshold Approaches to Evidence-Bases

There are two straightforward approaches to hierarchies of evidence bases—cumulative and threshold-based approaches. The cumulative approach is in effect a more complex atomism. The simplest cumulative approaches assume strong atomism—individual studies are appraised in isolation, and the evidence each provides (weighted for quality) is summed. This is one way that GRADE has sometimes been used (*cf.* [206]) in practice—at first glance, GRADE *appears to* ascribe a quality judgement to the evidence from individual studies. So, because GRADE assigns an initial high quality ranking to evidence from RCTs, some users read this as GRADE ascribing a high-quality level to the evidence from an individual RCT. In fact, the GRADE authors clarify that the intention is to assess the whole evidence base for a hypothesis about a treatment's effect on a particular outcome. They state: *"GRADE judgments refer not to individual studies but to a body of evidence"* ([14], p.403). But a cumulative approach to GRADE has occasionally been used, and such an approach simply sums together the evidence from a number of studies. Another cumulative approach, albeit with weighting for quality, is used in simple meta-analysis—studies are weighted for quality, usually according to size, and a weighted average of their estimates of effect size is taken (*cf.* [1]).

This cumulative approach inherits the problems of atomistic approaches to evidence and will not solve the problems raised in Section 6.1. Weighting studies for quality and then summing them does not take account of the inter-relationship between different sources of evidence and their quality and strength. Like strong atomistic appraisal, these simple cumulative approaches do not take account of effects on the strength and quality of evidence of replication, heterogeneous study results, defeaters, and other meta-evidence.

A more sophisticated cumulative approach could allow for atomistic appraisal of studies, but accept that the individually appraised evidence may not be simply summed in a linear fashion. Some interaction between study results could be accounted for in this way. For instance, multiple independent studies with results which concur may be mutually reinforcing, such that the strength of this evidence base is greater than the sum of the strengths of the individual studies. But again this allows only for a portion of the interaction within evidence-bases—as far as atomistic appraisal of individual studies is maintained, the effects of one study on the quality and strength of another, not just on the overall strength or quality of the evidence base, is omitted. A more common approach uses threshold levels of evidence. A threshold-based approach specifies a threshold condition which the evidence-base must meet in order for the level of evidence for a hypothesis to be high (and moderate, and low, etc.). Examples of threshold-based hierarchies of evidence-bases include Cook et al.'s influential 1992 *CHEST* hierarchy [151] (see Fig.13, above), the SIGN hierarchy [133,134] previously used by NICE [28] (Fig.6, above), and the Australian NHMRC hierarchy [6] (Fig.7, above). The most common threshold for a high level of evidence is that at least one high-quality RCT, systematic review or meta-analysis provides evidence for the hypothesis. Other thresholds which require more studies could easily be formulated, although this has not occurred in the EBM literature. Examples include the FDA's former criterion for pharmaceutical treatment approval, which required two large RCTs with p-value < 0.05 [489].

There are a number of problems with a threshold-based approach. First, unless these approaches are given a modified interpretation or many additional conditions, they will be susceptible to a range of challenges from Chapter 4, including the "Bad Implementation Problem". An RCT may be small and badly conducted, while an observational study is large and well-controlled. But hierarchies like those shown above allow RCT evidence to justify a higher level of belief in a hypothesis than an observational study, even where the RCT is poorly implemented and the observational study is well implemented. Modifiers or a great deal more conditionality would be needed to make the thresholds track with common sense appraisals of evidence.

The strongest challenge to level of evidence rankings identified in Chapter 4 came from Clarke et al.'s application of the Russo-Williamson Thesis (RWT). RWT claims that in most cases, mechanistic evidence is needed in addition to evidence of correlation in order to warrant belief in a causal hypothesis. The levels of evidence tables either omit mechanistic evidence, as in the case of those pictured above, or rank it at the lowest levels (e.g. [15,103]). But if mechanistic evidence is truly *necessary* to justify a high level of evidence for a causal hypothesis, then these threshold conditions in levels of evidence rankings are incorrect. Evidence from at least one RCT is not sufficient for a high level of evidence for a hypothesis.

A second challenge compounds this problem. Thresholds such as those offered in the tables above do not generally take evidence of non-causation or non-association into account. For instance, there are two ways of interpreting NICE's hierarchy's Level 1b requirement for *"Evidence from at least one randomised controlled trial"* [28]—that there is at least one RCT which supports the claim that the treatment is effective, or that there is at least one such positive trial *and* that other trials do not (generally) contradict that result. Either reading is very problematic. The next section presents a case-study which illustrates both objections—the importance of mechanistic reasoning in an evidence base for a causal claim, and the importance of negative evidence. These are compounding problems because mechanistic reasoning and laboratory studies may often be important evidence against causal claims.

6.2.2: The "Paradox" of Evidence-Based Alternative Medicine

There is an intriguing pseudo-paradox in the literature surrounding evidence-based Complementary and Alternative Medicine (CAM) (see e.g. [490-493]). According to some researchers, there is an evidence-base which supports claims for the efficacy of some alternative therapies for certain conditions (see e.g. [494-497]), including RCTs and systematic reviews which provide evidence for the efficacy of acupuncture for arthritis [498] and nausea due to chemotherapy [499], chiropractic for back pain [500], and intercessory prayer for bloodstream infections [392]. Yet most evidence-based practitioners and philosophers are unwilling to accept that these treatments are efficacious and introduce them into the medical canon (*cf.* [490,491,493,501]).[97]

[97] A related argument, also referred to as a paradox of alternative medicine [490,492,493,496], is that alternative medicines cannot work by definition—if they worked, they would be scientific medicines. As Fontanarosa & Lundberg put it: *"There is no alternative medicine. There is only scientifically proven, evidence-based medicine supported by solid data or unproven medicine, for which scientific evidence is lacking."* ([496], p.1618) But this pseudo-paradox dissolves. A treatment could 'work' yet be excluded from scientific medicine for other reasons—for instance that it treats a condition which is not recognised as a disease by orthodox medicine. For example, homosexuality was removed from the Diagnostic and Statistical Manual of Mental Disorders (DSM-II, 7th printing) in 1974 [502]. The American Psychological Association condemned the use of conversion therapies in a statement in 2000 [503]. Such therapies are also rejected for the pseudo-disorder of ego-dystonic homosexuality, which was similarly removed from the DSM in 1987 (DSM-III-R) [504] (see [505]). Whether conversion

This ‘paradox’ can be resolved in two ways: rejecting the practices of these practitioners as ill-informed, or rejecting the account of evidence bases which judges these to justify a high level of belief in the hypotheses. This section argues for the latter resolution. A handful of RCTs with positive results is an inadequate evidence base for the claim that the treatment is efficacious, for three reasons: (a) evidence against efficacy claims is under-represented in the literature and under-appreciated in levels of evidence rankings; (b) some of the studies which comprise this evidence base are undermined by other evidence; and, (c) there is a lack of independent replications in the evidence base. To provide a truly compelling case for effectiveness, the evidence base would need to contain more replications and more variation in studies. The lessons learned from this case apply equally to research in conventional therapeutics, but the case of alternative medicine is particularly instructive as a case in which the discrepancy between the results of applying threshold levels of evidence and clinical consensus is apparent.

Many clinical trials are designed to study whether two treatments are at least equivalent in average effect on some outcome (“equivalency trials” or “non-inferiority trials”), or whether the experimental treatment outperforms the control (“trials of benefit”) (*cf.* [506,507]). Due to the statistical framework usually accepted in clinical epidemiology, negative results in these trials are usually interpreted as providing no evidence of benefit (or equivalency), as opposed to evidence that the treatment is on average less effective or ineffective on the outcome (see [242,252,508]).[98] There is an asymmetry in the methods needed to provide strong, high-quality evidence of effectiveness vs. ineffectiveness.

In addition to the focus upon trials of equivalency or benefit, there are three features of medical research and the existing literature which make it unlikely for evidence that a treatment does not have an effect on an outcome of interest to predominate even where treatments are ineffective: publication bias, selective reporting, and experimenter bias.

Publication bias refers to the well-documented tendency for trials with positive results to be disproportionately represented in the medical literature (see [198,239-241,248,418,419,509,510]). Positive results are more likely to get published than negative ones, in more prestigious journals, and reach publication more rapidly. There is growing evidence that positive results are often repeatedly published, sometimes by different authors, under different titles and in different journals, which may

therapies succeed in changing sexual orientation or behaviour is irrelevant—they are not part of scientific medicine because they do not treat a recognised condition.

[98] In particular, the Fisherian statistical significance framework has dominated much of medical statistics [179]. Confusion over the term “significant effect” may lead some to view negative results as “insignificant” in the sense of unimportant (see [179,252,489]), which is reinforced in the language of Fisherian hypothesis testing: “failing to reject the null hypothesis”. Meta-epidemiological studies have demonstrated that p-value correlates strongly with publication, time to publication, and prestige [242,508].

lead to double-counting in systematic reviews and by individual practitioners searching the literature (see [239,244,245,511]). It can be difficult to determine whether two publications are reporting the same data. Industrial sponsors may suppress or delay publication of negative findings (see [247,411,415]). Journal editors and peer reviewers may be less impressed by and less willing to publish negative findings (see [238,512]). Rejection from journals may slow the publication process for negative results. There is a (changing) perception that negative results are less valuable than positive results [239,419]. Editors may also be deterred by the possibility of litigation where pharmaceutical companies are involved, especially as sponsors (*cf.* [420]).

A potentially pernicious variant of publication bias is *selective reporting bias*. Most trials test for more than one association between variables. For instance, a trial of acupuncture for arthritis might measure the comparative effects of the treatment against a placebo on many measures, such as pain, quality-of-life, ACR-20, EULAR-response, Boolean remission, etc. (see Chapter 5 for further discussion of outcome measures). In cases of selective reporting, only the positive results from the trial are reported. Phrasing may make it appear as if these were the only relevant relationships tested, and omit the evidence of a lack of correlation between other variables. Selective reporting is commonplace [241,300,483,509,513,514], and can have serious effects on meta-analyses and systematic reviews [300,515,516]. The trial might only report results from some subgroup analyses, or only the positive subgroup analyses, performed on the trial population [300]. Yet negative tests of secondary outcomes can be important evidence both as part of the evidence-base for a claim, and as evidence relating the quality and strength of evidence, as was seen in the case of Lawton et al.'s exercise on prescription study.

Experimenter bias occurs when an experimenter prejudges the intended or expected outcome of her study, and consciously or unconsciously manipulates the study to provide these results [237]. Where stigma attaches to negative results, experimenters may subtly manipulate studies to provide positive outcomes (see [236]). There are many ways to interfere with experiments. Blinding will not prevent all forms of experimenter bias. First, blinding is not easy to ensure or enforce, especially if experimenters do not want it to be [249,409,427,486,487]. Checks are rarely carried out for successful blinding at the end of studies [316,410,485-487][99], and the small amounts of evidence so far collected suggests that *if they were*, most studies would turn out to have been unblinded at some point [409,410].

The result of these mutually-reinforcing problems is that negative findings are likely to be under-represented in the medical literature, and interpreted as non-findings rather than evidence that

[99] Hróbjartsson et al. [410] report that only 2% (31/1599) of blinded RCTs from the Cochrane Central Register of Controlled Trials reported any testing for the success of blinding.

the treatment was not efficacious. Simply amassing the results of RCTs will favour treatments which are ineffective.

In the light of the evidence concerning publication bias, selective reporting and experimenter bias, three features of the evidence base become particularly important: independence, replication, and tests for publication bias. Independence is important to decrease the threat of experimenter bias permeating the evidence base. If different studies are performed by different research groups, and these research groups are funded by different sources and have different interests, then one expects a decreased risk of experimenter bias. Thus, consistently similar results from a range of independent sources provide stronger and higher quality evidence together than each can in isolation. Where results from different sources with different motivations conflict, one's confidence in each study is weakened due to the threat of experimenter bias.

Replication is also important to guard against selective reporting. Where one result from a multitude of tests is positive, one is justified in suspecting a spurious correlation or in treating the result only as exploratory (*cf.* [477,489]). But where the negative tests are not reported, one cannot tell whether to suspect spurious correlation. An independent replication specifically designed to test for the correlation found in the original study would corroborate or undermine the original study. Again, here, the study plus replication can be stronger and higher-quality than either alone, as the threat of spurious results due to data-mining is weakened.

Tests can also be performed on the evidence base itself to detect publication bias. Meta-analysts have developed techniques such as funnel plot analysis to try to determine whether negative results have been suppressed (see [1,250,251]). Meta-evidence from funnel plot analyses can support the evidence provided by the evidence-base where publication bias is not detected, or undermine the evidence where publication bias is found. Reviews which systematically search the "grey literature" (unpublished academic literature, including working papers, reports, conference papers and preprints [517]) for unpublished results can also discredit previous reviews where it is found that negative results were suppressed [1,32,43,517]. One study showed that reviews which omit grey literature searching were considerably more likely to report a positive effect and generally reported larger effects [518].

In addition to these challenges, methodologies exist which systematically privilege the experimental treatment, despite technically using a form of randomised controlled design. One such example is the "cohort multiple RCT" (cmRCT) developed by Relton et al. [519]. The method was described in a *BMJ* article in 2010 [520]. Relton, herself a homeopathic practitioner, performed a pilot study using cmRCT methodology on homeopathic treatment for menopausal hot flushes [521]. Other cmRCTs are underway testing homeopathic treatments for depression [522], collaborative care for sub-threshold depression in over-75s [523], for scleroderma [524], and in a large-scale (n≈20,000) study of obesity and chronic disease in South Yorkshire [525]. Relton et al. claim that cmRCTs combine the best of both RCTs and observational studies [520]—they combine the longitudinal follow-up and large cohort of an observational study with the rigour and control of an RCT. They claim the same epistemic privilege for cmRCT evidence as for RCT evidence, and presumably the same position in hierarchies.

cmRCTs use a complex methodology which conceals systematic privileging of the experimental group. The design involves recruiting a large observational cohort, and then selecting a subgroup within this cohort for a trial. Multiple trials can be performed on the observational cohort, but not necessarily involving any of the same patients. The trial cohort is randomised into control and experimental groups. Only then are the experimental group patients contacted and asked to consent to the trial. The control group are never aware that they are participants in a trial. This approach is termed “Zelen consent” or “post-randomisation consent” (see [526-528])[100]. The major issue here concerns ‘trial effects’. Trial effects occur when patients perform better when they know they are part of a clinical trial [232-234]. This can be due to Hawthorne effects and enhanced placebo effects due to the expectation of benefit from the experimental treatment, or due to treatment bias—patients in trials are generally more closely monitored and treated by healthcare professionals with greater specialisation and expertise in their condition (see [229,230,234]). Because only the experimental group are aware they are part of a trial and are treated as such, they are likely to benefit from these trial effects, while the control group will not. Therefore, it should be expected that even in cases where the treatment has no greater average effect than the control, the cmRCT will report a greater effect for the experimental group, effectively generating consistent false-positives where treatments have no effect.

In the case of cmRCTs, evidence of a trial effect is strong counterevidence against the validity, and thus the quality, of the evidence provided by the cmRCT. This is a case of one study invalidating another—evidence of the trial effect undermines the validity of the cmRCT. It also provides a cautionary tale against appraising the quality of evidence purely on the basis of the methodology.

In addition to illustrating the interaction of evidence, these interrelated problems in the evidence bases for alternative therapies (which also affect the evidence bases for conventional therapies) illustrate the two problems with relying on threshold levels of evidence. First, satisfying a

[100] The Zelen Method of post-randomization consent from the experimental group only should not be confused with the ‘Play the Winner’ method, also due to Zelen, in which patients are allocated based on the results of the treatments in the cohort so far [529].

simple criterion such as one or two positive RCTs neglects the importance of negative evidence. These challenges show that one should expect to find many false positive results in the medical literature.[101] So, positive results are overvalued. But furthermore, a hierarchy which assigns a high level of evidence where there are positive RCTs can be misled by false-positive generating methods like cmRCTs, by experimenter bias and by selective reporting. It is possible to manipulate or mine data to provide some positive results for ineffective therapies. Clearly, where much of the literature contradicts these findings, the level of evidence should be reduced. Threshold criteria should at the very least be revised to consider consensus amongst study-results, and to prioritise replication and independence within the evidence base. This is one of the steps taken by GRADE in improving from the levels of evidence tables.

The second challenge outlined above comes from the importance of mechanistic reasoning in undermining evidence from such trials. One important means of differentiating between evidence bases characterised by false positives and genuine evidence of an effect is the question of whether the effect is biologically plausible.

6.2.3: Biological Implausibility as Counter-evidence

Strong, high-quality evidence *against* the claim that a treatment is efficacious may be different from strong, high-quality evidence *for* the claim—there is an asymmetry in evidence quality. There can be strong evidence against the claim that a treatment is effective, despite the limitations of RCT evidence as negative evidence. Sources which provide strong reason to believe that a treatment will not be effective include mechanistic reasoning and physiological and pharmacological studies. Mechanistic reasoning allows us to work out plausible mechanisms by which a treatment could bring about an effect on the outcome of interest. Physiological and pharmacological studies can test whether such mechanisms are biologically plausible. If the effect hypothesised is biologically implausible, then this provides evidence that the treatment is not effective. Some authors have argued that a biologically plausible mechanism is weak evidence that a treatment will be effective, as there is no guarantee that the mechanism will work as anticipated in the body (e.g. [58,59,389]). But, as Bird ([116], p.6) has argued, even if a plausible mechanism is weak evidence for a hypothesis, evidence that the effect is implausible can be strong evidence against it. It is the combination of evidence against biological plausibility and the limitations of the positive evidence base which justifies rejecting the claims made by CAM proponents.

[101] Ioannidis argues that in some fields, the rate of positive results simply measures the degree of bias in that field (see e.g. [530,531]).

For instance, the use of multivitamin supplements for healthy adults without nutritional deficiencies is a case in which laboratory evidence can undermine the biological plausibility of beneficial effects. In particular, “megavitamins” (also known as “orthomolecular medicine” [532]) are vitamin supplements containing over five times [533] or over ten times [534] the recommended daily amount of vitamins (commonly vitamins C and E, and β-carotene). There are trials which suggest that large doses of vitamin E may reduce cardiovascular events (e.g. [535]), although these are widely contradicted in the current literature [536].

But physiological evidence alone can undermine such studies as evidence for a treatment effect. Mechanistic reasoning suggests that the body’s homeostatic mechanisms will react to surplus vitamin intake by metabolizing and excreting the unneeded substances, maintaining a stable level of the substance in the relevant organs.[102] Physiologic studies support the claim that homeostatic mechanisms maintain relatively stable vitamin levels in the brain despite large orally-administered doses of vitamins [540]. Water-soluble vitamins like vitamin C, niacin and folic acid are metabolized or excreted without substantial effects on vitamin concentration in the relevant organs [533]. Where this process is interrupted or unsuccessful, harms due to toxic accumulation result, which have been demonstrated in a number of studies (see e.g. [533,539,541,542]). Fat-soluble vitamins like vitamins A, D and K are stored, not used, and can build up to toxic levels (see e.g. [533,534,542-544]). In the case of microminerals such as fluorine, which are needed in small quantities in the body but are harmful at higher dosages, the introduction of megamineral therapies can similarly be shown to be harmful by physiologic studies which show increased concentration in the body, and interruption of normal metabolic pathways (e.g. overdosing with zinc prevents metabolisation of copper and iron) [534,539].

The importance of evidence of a biologically plausible mechanism is one way in which a more diverse evidence base is stronger than a homogeneous one. Where only epidemiological studies are provided, a plausible biological mechanism is not assured. But given that biological implausibility can be strong evidence against an efficacy claim, the absence of a plausible biological mechanism is an important lacuna in any evidence base. Hierarchies which do not at least require that there is *no evidence of biological implausibility* as a requirement for a high level of evidence, or strong

[102] In other cases, mechanistic reasoning can discredit a hypothesis immediately. A notable example discussed in detail by EBM proponent Ben Goldacre [537,538] is the claim (primarily associated with nutritionist Gillian McKeith) that spinach and other dark-leaved plants will oxygenate blood by photosynthesizing within the body. The hypothesis can easily be discredited by noting that photosynthesis requires light to occur, which renders the effect mechanistically implausible. Other examples include the consumption of shark cartilage to prevent cancer because sharks do not suffer from cancer, which was discredited by simple observations proving that sharks *do* get cancer ([539], p.87).

recommendation to use a treatment, are misguided. No threshold-based approach has required this mechanistic evidence be taken into account.

This claim is not equivalent to the Russo-Williamson Thesis, and does not commit one to the claim that mechanistic evidence is necessary to establish causation. Under this framework, *lack of evidence of biological plausibility* is a serious limitation of the evidence base, but not necessarily insurmountable by sufficiently compelling statistical evidence. However, *evidence of biological implausibility* is not, on this framework, surmountable, and undermines statistical evidence to the contrary.

6.3: GRADE and the Challenges to Levels of Evidence

The GRADE approach is the most sophisticated current hierarchical approach to evidence appraisal. GRADE includes a complex hierarchy of evidence bases, which is used at two stages in assessing the evidence for a recommendation. First, GRADE's hierarchy is used to assess the quality of the evidence base for each claim about a treatment's effect on each outcome of interest. Then, these assessments of evidence quality are combined to form an overall level of evidence for a claim about the treatment's risk-benefit ratio by taking the lowest quality level for any critical outcome. On the basis of the level of evidence for the risk-benefit claim and the balance of risks to benefits, a recommendation is made, and the strength of that recommendation is assessed. High level evidence for a claim that the benefits clearly outweigh the risks is needed to justify a strong recommendation. High level evidence is not sufficient to justify a strong recommendation alone—the cost-benefit ratio must also be clearly favourable.

Evidence quality is assessed by assigning an initial quality level on the basis of the methodology of the studies in the evidence base—high quality for RCT evidence, low quality for observational study evidence. Then, a number of up- and down-grading criteria are applied. The full criteria can be seen in Figure 9, above. The net result of applying all of these criteria is the final quality assessment for the evidence. For a full account of the methodology, see Section 2.1.

The GRADE approach explicitly recognizes that evidence interacts. Evidence that there is publication bias in the literature, for instance from a funnel plot [250,251] or a search of the grey literature [517,518], can downgrade the quality of an evidence base. Evidence of a threat of bias, such as the observer-expectancy bias identified in the exercise on prescription case, or the trial effects identified in cmRCT methodology, also downgrades evidence quality (see [203]). GRADE also recognizes that not all biases necessarily undermine the quality of evidence—for instance, where all plausible biases would reduce the size of the effect observed, and a positive effect is still observed, the biases do not detract from the quality of evidence for the claim that the treatment has a positive effect [202].[103]

[103] However, this presumably *would* detract from a claim about the absolute *effect size*. One of the flaws of the GRADE approach is that it does not adequately distinguish between the evidence for a claim that a treatment is effective/ineffective, and claims about the specific size of the treatment effect. Precise effect size estimates are needed to facilitate a proper comparison of harms and benefits in the recommendation stage of the GRADE appraisal.

However, there remain at least two serious issues which are currently unaddressed in the GRADE framework, which stem from the use of hierarchical appraisal:

1) The approach currently does not differentiate between evidence for and against a treatment effect hypothesis, and does not recognize that evidence of biological implausibility is strong counter-evidence against an effectiveness claim.
2) The approach focuses on evidence for an estimation of the average treatment effect in a population, and neglects evidence about the distribution and variation of treatment effects, which is crucial to making treatment recommendations.

Challenge (1) is the argument leveled against threshold-based and cumulative levels of evidence above. GRADE does not currently overcome this problem, but it has the capacity to do so, as is argued below. GRADE should be reformulated to address the problem. However, the GRADE system's emphasis on positive effectiveness claims is a major drawback. Challenge (2) is not a surmountable problem for a hierarchical approach to evidence. This challenge will be developed in full in Section 6.4. These problems are not unique to the GRADE approach. They also affect other hierarchies of evidence bases. GRADE is discussed here because it offers the most sophisticated approach with the most potential to ameliorate these problems.

With respect to challenge (1), GRADE has a useful set of tools to accommodate *some* asymmetry between high quality evidence for and against hypotheses. GRADE has separate up- and down-grading criteria, which could be used to allow different information to affect an evidence base depending on whether it supports or undermines the evidence base for the claim that a treatment is effective. Mechanistic reasoning and physiologic studies could be weak evidence *for* a treatment effect, but very strong evidence *against* one. GRADE could in principle accommodate this. There are two main ways in which this could be achieved. First, GRADE could introduce new downgrading criteria which downgrade the evidence base substantially where evidence of mechanistic and/or physiological implausibility is present (but not criteria which upgrade the evidence base because of a plausible mechanism). So, for instance, the quality of the evidence base for the claim that vitamin C megavitamin therapy is effective would be downgraded several levels because of physiological studies which show that vitamin C supplementation does not affect vitamin C levels in the brain and other relevant organs.

This solution is imperfect. GRADE remains focused on appraising evidence bases for claims that a treatment is effective, and overlooks appraisal of evidence bases against such claims. The focus is on appraising the evidence that a treatment works, and then checking whether other evidence (negative results, evidence of publication bias, evidence that the effect is not biologically plausible, etc.) undermines it. But a very low level of evidence that a treatment works is not the same as a high level of evidence that it does not work. Such an equation confuses absence of evidence with evidence of absence of the effect. Demonstrating that the evidence base for a treatment's effectiveness is very low quality would just *not recommending* the treatment. A high quality evidence base that the treatment is ineffective would be needed to justify *recommending against* the treatment. There clearly are cases in which there is high quality evidence that a treatment does not work, which are unconsidered in the GRADE approach (see Section 4.3.4).

Challenge (1) can be accommodated to some extent within the GRADE system. There is room for mechanistic evidence to play a role, especially as evidence to downgrade evidence bases. GRADE is better placed than simpler hierarchies of evidence bases to accommodate the insight that there may be methodologies which provide strong negative evidence but not strong positive evidence. However, at present, the approach only considers strong negative evidence to be evidence against a positive claim, not evidence for a negative claim.

This next section will consider challenge (2) based on the observation that many treatments have *heterogeneous effects*. The primary results of comparative clinical studies such as RCTs and non-randomized experimental studies only reveal information about the *differential average treatment effect* in a population. The evidence given a high initial ranking from RCTs is evidence about average treatment effects. But information about average treatment effects is an insufficient basis to make recommendations—information about variation and heterogeneity in effects is of critical importance to clinical recommendations and decision-making.

6.4: Heterogeneous Treatment Effects and GRADE

6.4.1: Clinical Trials provide Evidence for Average Treatment Effects

Hierarchies, including GRADE, prioritise evidence about the differential average treatment effect on some outcome in some population. This section defines 'differential average treatment effects', and shows that information about average treatment effects underdetermines the distribution of treatment effects. Clinical trials such as RCTs are primarily used to provide evidence for claims about the average treatment effect, and their primary results provide no evidence about individual treatment effects.

In conducting a controlled trial or study, researchers compare the incidence of the outcome of interest across several groups. In a simple two-armed controlled trial, there are two groups—the control group (a sub-population taking a control treatment) and the experimental group (a sub-population taking an experimental treatment). Researchers record the incidence of the outcome of interest in both groups—these are termed EER (Experimental Event Rate—the percentage experiencing the outcome of interest in the experimental-group) and CER (Control Event Rate—percentage of patients experiencing the outcome in the control-group).[104]

The researchers compare EER to CER. The Absolute Risk Reduction (ARR), the favoured measure of the efficacy of a treatment according to many EBM texts[105] (e.g. [1,22,93,547]), is defined as:

$$\text{ARR} = \text{CER} - \text{EER}$$

ARR measures the *difference in average risk* of an outcome between the groups. ARR is only a true *treatment effect* if other causes of differences in outcomes between the groups are ruled out, i.e. if the trial was not confounded. The proportion of ARR attributable to the treatment (i.e. not attributable to bias) is the *differential average treatment effect.* As noted in Chapter 2, internal validity measures confidence that the difference in average risk measured in a study is a true treatment effect. One does

[104] Alternatively, where the outcome of interest is not an event but performance on some measure, the average change in that measure is recorded to calculate the mean difference. The argument in this section applies equally to either approach. An example which considers this approach is provided in Section 6.4.2.

[105] Some EBM proponents prefer the NNT (Number-Needed-to-Treat) measure—but NNT is simply the inverse of ARR (see [9,545,546]).

not learn the absolute average effect size—only the difference in risk of the outcome compared to the control.[106]

Although the term ‘average treatment effect’ is standard terminology (see e.g. [1,457,481,482,548]), the term ‘average’ may suggest a method which determines the individual treatment effects, and takes the average of these. A natural response given the terminology would then be to call for greater reporting of the individual treatment effects. However, the average treatment effect is not discovered by summing individual treatment effects and dividing by the number of participants. Trials *do not* provide any information about individual treatment effects. There is information about the initial state and end state of each patient; however, it is unknown whether the outcome for each patient was due to the treatment. The control group is used to estimate what the incidence of the outcome *would probably have been* amongst the experimental group, had they been given the control treatment rather than the experimental treatment. But RCTs are unpaired—there is no identifiable matching individual in the control group to whom an individual in the experimental group corresponds. Therefore, what outcome any given *individual* would have experienced had she had the control rather than the experimental treatment cannot be determined. Comparative clinical studies such as RCTs record the overall effect, the difference in event-rate between the groups as a whole—an average treatment effect with no information about the individual effects which compose that average.

6.4.2: Average Treatment Effects are Insufficient to Justify Strong Recommendations

RCTs and other controlled clinical trials provide evidence for claims about the differential average treatment effect on some outcome in some population. This section will demonstrate that this kind of information alone is an inadequate basis for a clinical recommendation. First, the kinds of claims in which practitioners and patients are interested, claims about prognoses given treatments, are outlined. Then, the importance of variation in treatment effects is discussed, and approaches which consider the probability distribution of effects of different degrees, rather than simple point estimates of the average effect, are described. Heterogeneous treatment effects and effect modification are

[106] This is the case because the control intervention may also have had some effect on the risk of the outcome. The total effect of the treatment upon the outcome cannot be determined unless the effect of the control is already known. Even placebo treatments can have effects upon most outcomes of interest [229,230], so even these trials will not provide true estimates of absolute treatment effect (see [428-430]).

defined, and it is argued that information about variation and the causes of variation is crucially important to clinical recommendations and decision-making. This claim is supported by four case-studies from a range of medical specialties. Therefore, it is claimed, information about variation and the causes of variation in treatment effects is necessary for best practice, and basing practice upon information about differential average treatment effects alone is sub-optimal, no matter how strong or high-quality the evidence for that information. Strong recommendations on the basis of high-quality evidence for net average benefit alone are unjustified.

Medical practice focuses upon diagnosis, prognosis and treatment. A patient presents with symptoms and signs. The practitioners need to determine what is wrong, what can be expected to happen, and how this will differ if the patient is given various treatment options. They then need to implement whichever treatment plan the patient chooses. In common with the EBM movement, this discussion focuses primarily on treatment-prognosis claims, i.e. claims of the form:

‘If patient P is treated with regimen T, it is reasonable to expect her prognosis to be X.’

There are several important features of treatment-prognosis claims. First, these are *predictive* claims: they make a claim about likely future events. Second, they are *individualised* claims: they are about the prognosis of a specific patient, not a broader target population. They may or may not be *comparative* claims—that is, X may stand for a comparative claim, like ‘3-6 months longer life in slightly better health than if P was treated with C’, or an absolute claim like ‘3-6 months of reasonable quality life” or “a complete recovery within 2 weeks’. Absolute claims are more informative for the patient, and so are preferable, though comparative claims are also valuable where absolute predictions cannot be supported adequately.

Ultimately, practitioners are interested in a further claim, a ‘best prognosis’ claim of the form:

‘T is the treatment which maximises the desirability of the expected prognosis for P.’

However, inferring such a claim is beyond the scope of evidence appraisal. To support a ‘best prognosis’ claim, one must first decide upon the likely prognoses for each treatment option (including no treatment), and then weigh the value *for patient P* of each likely outcome. The patient must evaluate which outcomes are preferable, and then take into account the likelihoods of preferred outcomes and the risks given each option. Depending on her valuation of outcomes and her degree of risk-aversion, the ‘best prognosis’ claim *for her* can then be inferred. Even given identical information about their prognoses, two patients with different characteristics and preferences may disagree about the best treatment option for them.

So, the claims which clinicians want to base upon evidence are 'treatment-prognosis' claims—individualized and predictive. For EBM to succeed in applying evidence in clinical practice, EBM procedures must allow practitioners to appraise the evidence for and against these claims. As stressed in Chapter 1, this blocks retreats for the interpretation of hierarchies. The role of hierarchies in evidence appraisal must be to appraise evidence for these treatment-prognosis claims, not for some limited class of claims such as claims about average treatment effects in target populations, unless these directly relate to treatment-prognosis claims.

Treatment prognosis claims are claims about likely individual treatment effects. To make a strong prediction about likely individual effects requires information about the probability of different effects and effect-sizes. An estimate of the differential average treatment effect alone does not provide this. The distribution of individual treatment effects can vary, and is not necessarily well represented by the mean. Clinicians are interested in both the actual distribution of individual treatment effects in a study, and in a predicted distribution for future patients (i.e. a probability density function for the individual treatment effect variable[107]). Several features of these distributions provide important information relevant to clinicians' recommendations and patients' decision-making, including modality, skew and kurtosis.

For our purposes, the modality of a distribution refers to the number of peaks which occur when graphing a probability density function for individual treatment effects.[108] A unimodal distribution has a unique peak. Examples of unimodal distributions include the Gaussian (normal) distribution (see Fig.36, below). Bimodal distributions have two peaks, and distributions with more peaks are called 'multimodal' (see Fig.37, below).

[107] Note that where the treatment effect is discrete, not continuous, this will be a probability mass function. However, few discrete treatment effects are encountered in practice—even where the outcomes measured are discrete (e.g. stroke or no stroke in the next 12 months; >20% improvement on the ACR scale, etc.), the treatment effect is usually conceptualised as an effect on the *risk* or chance of meeting that criterion (e.g. stroke risk; likelihood of an ACR20 improvement), which is continuous. The points made here can be restated to apply equally to probability mass functions.

[108] In some definitions, modality refers to the number of unique modes—that is, the number of values which appear most frequently in a data-set. But it is more common to refer to distributions as multimodal if they have more than one peak. There are various ways of defining modality which preserve this commonplace use (see e.g. [549]).

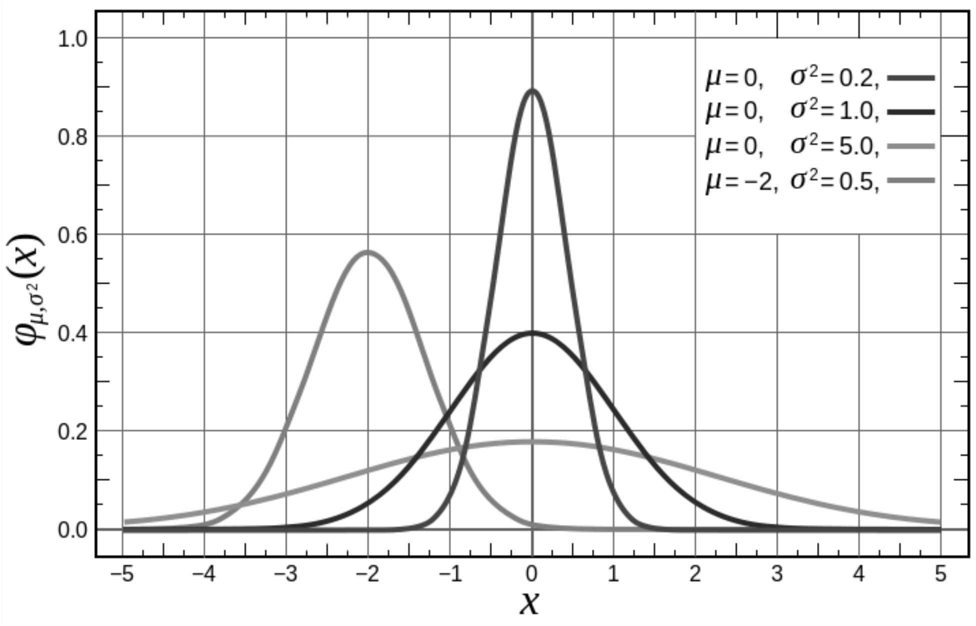


Figure 36: Creative commons image of four unimodal probability density functions. Mean is denoted by μ, while standard distribution is denoted by σ (and hence variance by σ²). In the medical treatment setting, the X-axis should be interpreted as the effect size, and the Y-axis as the probability of that effect size.

Multimodality in clinical medicine is likely to indicate that a treatment has different effects in different subgroups of the population. For instance, the treatment effects of temozolomide (see Section 6.4.3.3, below) could be represented through a multimodal distribution; one cluster of patients experience little effect, while another cluster experiences a greater effect. Treatments which work for a proportion of patients, but are ineffective in others, are likely to be best quantified through multimodal effect distributions (see Fig.37, below, for simple examples of such distributions).

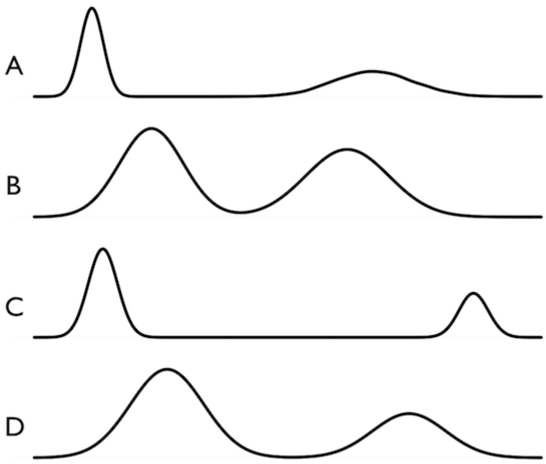


Figure 37: Examples of simple bimodal distributions adapted from Gregg [550]. Distribution A has two peaks, with half of all patients falling under each. However, because the variance of the second mode is greater than the first, the peak is flatter. This distribution may resemble one that is often expected in clinical practice—many patients receiving little or no effect with little variance (the left peak), with other patients experiencing a clear effect, but with greater variation (the right peak). Distribution B similarly has two peaks with equally many data points under each, but with more similar variances. In both of these cases, predicting an effect similar to the mean effect for any individual would be a mistake—the mean value falls equidistant between the two peaks, with almost no patients experiencing that effect. Distributions C and D are similar, but show the effect of different proportions of the sample falling under each peak, with a ratio of 2:1. Here the mean falls closer to the leftmost peak, but again in a region of the distribution with a low probability of occurring.

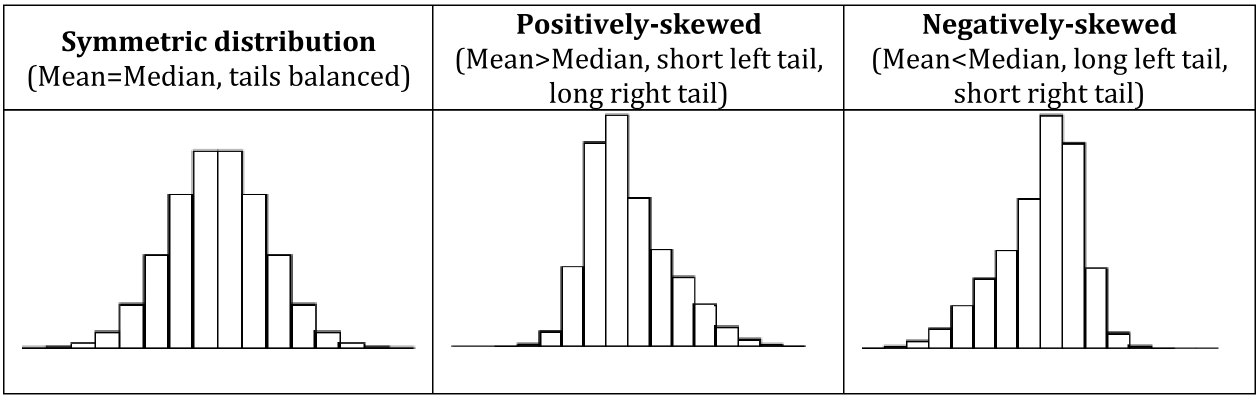
Measures of central tendency such as the mean effect provide much less useful information for clinicians where the distribution is multimodal. Few or even no patients may experience an effect similar to the mean, as is the case in each distribution in Fig.37. Knowing the average treatment effect in the population, then, is not clinically important where the distribution of treatment effects is systematically heterogeneous. Information about the distribution of treatment effects is a necessary precursor to a useable prediction on the basis of information about the average effect. Where the distribution is multimodal, a treatment prognosis claim will need to be nuanced—a simple point estimate of the most likely effect is unlikely to be a good prediction for any individual. Rather, a more complicated treatment prognosis claim is needed, taking into account the different modes and the different likely effect-sizes.

Even in unimodal distributions, several properties of the distribution are important in making clinical recommendations and decisions. First, the distribution may not be symmetrical.[109] Asymmetric distributions are *skewed*. Where the treatment effect distribution is skewed, the median effect typically does not match the mean.[110] In positively-skewed distributions, effects cluster below the mean, with a

[109] Of course, multimodal distributions can also exhibit asymmetry. Skew is hard to interpret in multimodal distributions—the same results for skew tests are usually consistent with multiple multimodal distributions.

[110] Contrary to the commonplace assertion in statistical textbooks (e.g. [551,552]), skew does not *imply* a specific relationship between mean, median and mode (e.g. positive skew implies that mode < median < mean) (see [553]).

longer tail above the mean—an intuitive interpretation is that most patients experience a lesser effect, with a minority of patients experiencing a much greater effect. In negatively-skewed distributions, most patients experience an effect greater than the mean, with a long tail below the mean, indicating that a minority of patients experience much lesser effects. Figure 38, below, depicts positive and negative skew.



Symmetric distribution
(Mean=Median, tails balanced)

Positively-skewed
(Mean>Median, short left tail, long right tail)

Negatively-skewed
(Mean<Median, long left tail, short right tail)

Figure 38: Symmetric and skewed frequency distributions, adapted from creative commons licensed images [554].

Skew is important in clinical decision-making. Patients are less likely to experience an effect close to the average treatment effect if the distribution is skewed than if the effects were symmetrically distributed about the mean. Information about skew, and the severity and direction of the skew, would help practitioners in predicting the actual likely effects for their patients, and the risks of exceptionally high or low effects. Depending on the direction of the skew, one might predict that patients are likely to experience a lesser or greater effect than the mean would indicate. There is also a greater risk of very large (in positively-skewed) or very small (in negatively skewed) effects, which should be taken into account. The more skewed the distribution, the more severe or common the risk. Skewed distributions typically have 'heavy tails' in one direction—more extreme outcomes, and extreme outcomes at a higher frequency, than would be expected if the effects were normally distributed.

Risks of exceptionally large or small effects may be very undesirable. For instance, in treating hypertension, a small effect is undesirable, but so is a very large effect due to the potential to induce hypotension by overtreating. This is often the case where a moderate value of a variable is ideal, and either extreme is pathological—e.g. BMI, heart-rate, vitamin concentrations, etc. Information about skew is particularly important when considering potential harms. The harms associated with positively vs. negatively skewed effect distributions are likely to be different—positively-skewed distributions are likely to have a greater risk of severe side-effects from overtreatment, while negatively-skewed distributions are likely to have a greater risk of minimal or negative effects. Depending on the risk-aversion and preferences of the patient, a treatment with a negatively skewed effect probability distribution may be preferable to one with a positively-skewed distribution, or vice versa—even if the average treatment effects are the same.

The evidence from the primary result of a randomized controlled trial does not provide information about skew. The average treatment effect is known, but the individual treatment effects are not. Hence the median effect and range is not calculable. So, the normal visual and statistical checks for skew cannot be performed (e.g. checking for discrepancy between mean and median; checking whether the mean lies close to the centre of the range; checking whether ±1 and ±2 standard deviations lie outside the range, etc.). The assumption that 68% of individual treatment effects will lie within 1 standard deviation of the mean (and that 95% will lie within 2 standard deviations) depends on the assumption that the treatment effects follow a Gaussian distribution—an assumption which is not always justified in clinical research.

A final important element of the distribution of treatment effects is kurtosis of the distribution. Degree of kurtosis can be interpreted as how closely the individual effects cluster together. In graphical terms, a distribution with positive kurtosis is more “peaked” than a Gaussian distribution, representing tightly clustered individuals, while negative kurtosis indicates a broader spread (see Fig.36, above, for examples of distributions exhibiting positive and negative kurtosis).[111] Again, the kurtosis of the distribution is clinically significant. Treatments with positive kurtosis in the effect distribution tend to have very consistent, predictable effects. Treatments with negative kurtosis are more variable. In some cases, treatments with consistent, predictable effects are preferable to more variable treatments. Where the clinician seeks to bring about an effect of a specific size, she would prefer to recommend a treatment that reliably achieves that effect. Again, this may be particularly salient where either extreme in a continuous variable is considered pathological. In other cases, where effects higher than the mean are desirable, clinicians and patients may prefer to risk a low effect for the chance of a high effect, especially where there is room for trying different treatments over time, as in chronic disease care. This may be particularly desired where high effects include complete remission. Which treatment is preferable may often be a matter of the patient’s circumstances and degree of risk-aversion.

Even in studies where the primary outcome measure is change from a baseline reading of a continuous variable, rather than the more commonplace risk of a binary event occurring in the

[111] Distributions with positive and negative kurtosis are also called ‘leptokurtic’ and ‘platykurtic’, respectively. The standard Pearsonian measure of kurtosis measures peakedness relative to the normal distribution. Without adjustments to give a zero value for kurtosis to the normal distribution, the terms ‘leptokurtic’ and ‘negative kurtosis’ are no longer equivalent. The unadjusted kurtosis of the normal distribution is 3.

population, information about median, range and skew is not usually available from clinical trial data. These are distributions of *differential treatment effects*, not of changes from baseline. Treatment effects are calculating by deducting what would happen if the patient received the control intervention from what actually happened given the experimental intervention (*cf.* [423]). But RCT methodology is an unpaired design—individuals in the experimental group are not matched with corresponding individuals from the control group. So, individual treatment effects cannot be calculated—there is no way of knowing whether a patient who experiences, say, a small change from baseline would have experienced a different change from baseline in the control group, etc.

Where the distribution of treatment effects exhibits multimodality, skew or positive or negative kurtosis, the treatment prognosis claims which are supported by the data are affected. Predicting an effect close to the average treatment effect is more or less justifiable depending on the distribution of effects. However, information about the distribution of treatment effects is not only important as a means of making predictions more nuanced. The distribution is also a feature of treatment effect which can be explained by *systematic heterogeneity*. Explaining why a treatment has greater or lesser effects in some patients than in others in terms of the features of the patient would allow greater precision in treatment prognosis claims. A treatment effect is called 'heterogeneous' to the extent that the effect on some outcome which the treatment produces varies within a population (*cf.* [457,481]). A purely homogeneous treatment always produces the same effect on the outcome. Treatment effects could be heterogeneous in many ways—for instance, a treatment might produce one of two different effects in a group of patients, or it might simply have a wide variance in treatment effects produced. *Systematic* heterogeneity is variation in the treatment effect which is caused by differences between the patients. Some patients experience different effects or effect-sizes to others *because* of some attributes of those patients. The next step in refining treatment prognosis predictions is to take into account information about the causes and predictors of different responses to treatment.

Consider the following model-case which will simulate the shortcomings of information about average treatment effects as a basis for treatment-prognosis claims. Suppose there is a painkiller, X. A trial compares X to placebo, and finds it reduces pain on average by 3 points on a 10-point pain-scale more than does the placebo.[112] Suppose this result reliably generalises to the target population. Very little is known about how individual treatment effects are distributed—only that the average is 3 points more than placebo. The average treatment effect underdetermines the distribution of individual

[112] For the purposes of this model case, assume the placebo has no measurable effects on pain-scale scores for any patient. In real-world cases, some degree of heterogeneity would be expected in *both* experimental and control treatment effect profiles.

effects. It could be the case that 100% of patients receive a 3-point decrease in pain-level—the painkiller's effect is perfectly homogeneous, unimodal and unskewed. On the other hand, it could be the case that 50% of patients receive a 6-point decrease in pain more than that of the placebo, while 50% receive no beneficial effects on pain—a bimodal distribution. In this case, there is serious *heterogeneity* of the treatment effect. Heterogeneity is particularly significant in the medical context because it is expected that most treatments will work in some patients, but not in others. As Guyatt et al. put it: *"few, if any, interventions are effective in all patients"* ([82], p.1292).

Consider a well-known instance of heterogeneous effects: side-effects. Even if a treatment has a single consistent primary effect, side-effects are generally heterogeneously distributed—some patients receive serious side-effects, while others do not. It is unlikely that the *risk* of these side-effects is the same across all patients. Some patients will be more susceptible to a side-effect than others. In a treatment with a heterogeneous side-effect profile, the expected cost-benefit ratio varies within the population, even if the primary treatment effect is quite homogeneous.

Many factors are relevant to variation in treatments' effects. Patients may differ in receptiveness to the treatment—for instance, metabolic factors may affect the uptake of active ingredients in a drug, and availability of the relevant receptors at the target-site may affect treatment-effectiveness [254,555]. The patient's baseline severity of the condition also modifies how effective a treatment can be (see [556]). If, for instance, a patient is at 3 on the pain-scale, the maximum benefit she can receive is a 3-point decrease in pain. Similarly, a patient whose stroke-risk is 5% cannot experience an absolute reduction in stroke risk greater than 5%. Kravitz, Duan & Braslow [481] identify examples of variation in baseline-risk, responsiveness to treatment and vulnerability to side-effects which can produce heterogeneous treatment effects.

A patient-attribute which causes a patient to experience a greater or lesser effect from a treatment is typically called an "effect modifier" in epidemiology [254] or is said to "interact" with the treatment effect by statisticians [557]. Victora, Bryce & Habicht distinguish between behavioural and biological effect modifiers. Behavioural effect modifiers under their framework are those which *"affect the dose of intervention that reaches participants"* ([254], p.402). These include any behaviours which influence the regularity and consistency with which patients take the prescribed regimen. Strict compliance with the precise regimen is carefully controlled in most clinical trials, but is rarer in practice [443,444]. Certain patients are much more likely to struggle with a complicated regimen, such as sufferers of ADHD or Alzheimer's disease [447,448]. It is unclear whether expectations of effects would be classed as 'behavioural', but studies have indicated that patients' expectations affect beneficial and harmful effects, and compliance (see e.g. [229,230]).

“Biological” effect modifiers refer to factors which affect the patient’s responsiveness to treatment [254]. Baseline risk is one such factor. A number of studies indicate that patients at low initial risk of an adverse outcome receive less benefit from treatments than patients at a high risk (see [556,558-560]). This may seem trivial, however in the light of preventive medicine strategies which have focused, since Rose, on preemptively treating large populations of low-risk individuals rather than attempting to identify high-risk patients [561,562], findings which show that low-risk patients do not benefit from an intervention are extremely important. For instance, cholesterol-lowering treatments have a much stronger effect on risk of mortality due to coronary heart disease (CHD) in patients at very high-risk of CHD, while the treatments showed no effect in moderate-risk groups and adverse effects in low-risk patients [558].

Other biological effect modifiers include both synergistic and antagonistic interactions with other substances in the target system [563]. Use of other prescription drugs, alcohol or narcotics can have a serious effect on the effects of many treatments [563,564].[113] Use of the herbal remedy St. John’s Wort is a particularly serious example, estimated to inhibit treatment effects for over 50% of common prescription drugs [454], including birth control pills, antiretrovirals and beta-blockers, and can be lethal in combination with certain opioids, LSD, MDMA and antidepressants (see [453,455]).

Environmental and dietary variations can also be important. Victora et al. describe how responsiveness to iron dietary supplements is modified by regional dietary variation—in areas where diets are rich in meats and ascorbic acid, iron absorption is enhanced and patients are more responsive to the treatment, while in areas where the diet is rich in phytates and polyphenols which inhibit iron uptake, the treatment is less effective ([254], p.404). There are numerous effects of living in a particular area which influence the individual patient’s exposures. For instance, McCormick et al. [567] showed that the prevalence of penicillin-resistant bacteria varies significantly between different sites in the USA, leading to geographical variation in responsiveness to penicillin-based treatments.

Call heterogeneity of treatment effects *‘systematic heterogeneity’* where there are effect modifier(s) which cause heterogeneity or allow heterogeneous responses to be predicted. The question of whether random heterogeneity occurs and can be meaningfully discussed can be set aside here, as systematic heterogeneity alone provides a sufficient challenge for hierarchies.

It may seem strange to argue that the *risk reduction* of a treatment could be heterogeneous, especially if risk is defined as the incidence in the target-population. But this is erroneous. *Average risk reduction* is the mean of the effects upon *individual risk* of the outcome. To take the individual’s risk as

[113] For instance, alcohol use is a behaviour which seriously increases the risks of methadone treatment—a particularly problematic interaction given that alcohol and heroin addiction are often co-morbidities [565,566].

equal to the average risk in a population ignores the reference class problem. Clearly, individuals are members of many populations (I am a member of the population of Brits, of English, of males, of philosophers, of blond-haired people, etc.). The average risk of an outcome is not necessarily the same in every population. What individual risk *is* in this case is beyond the scope of this thesis. But it is clear that an individual's risk need not match the average risk in any given reference class.

Let us take a model-case to illustrate before defending the thesis that such heterogeneity of risk reduction occurs in practice. Suppose the average baseline-risk of stroke in a population is 20%—20% of patients are expected to have a stroke in the next year if left untreated. Only 10% of the population suffer stroke when treated with Z. One hypothesis is that Z has uniformly reduced everyone's absolute risk of stroke by 10%—every individual was half as likely to have a stroke. But another is that Z has "*paradoxical effects*" [568,569]—in some patients, it prevents stroke, but in others, it causes stroke. One explanation is that some patients have their risk of stroke more than halved, but a minority of others have their risk raised, equaling out to an average effect of a 10% reduction across the target population. Paradoxical responses do occur. For instance, some antidepressants cause suicidal ideation in a minority of patients [568]. Carotid endarterectomy provides another example (see Section 6.4.3.1, below). Even where there is homogeneity on one outcome—perhaps, 'Appendectomy will cure your acute appendicitis'—there is often heterogeneity of effect on others, such as potential side-effects of the treatment; the risk of surgical mortality for appendectomy is systematically heterogeneous, with younger, healthier patients less likely to experience surgical complications than elderly patients or, for instance, patients with clotting disorders (*cf.* [569]).

6.4.3: Heterogeneous Treatment Effects in Practice

There are a number of cases of heterogeneous treatment effects and paradoxical effects which have been observed in clinical practice. These include carotid endarterectomy for carotid artery stenosis, temozolomide for gliobastoma multiforme, quinidine for atrial fibrillation, and benoxaprofen for chronic rheumatoid arthritis.

6.4.3.1: Carotid Endarterectomy

Perhaps the best-known case of heterogeneous treatment effects to philosophers is carotid endarterectomy for carotid artery stenosis. This case has been discussed in the context of problems of external validity (see [84,213,261,481,570])[114]. Carotid stenosis is the narrowing of a carotid artery due to plaque build-up. Carotid stenosis causes increased risk of stroke due to the potential for plaque emboli to break off and become lodged in blood-vessels in the patient's brain, causing reduced blood-flow (ischemia) to the area served by the vessel, and leading to transient (TIA) or permanent (stroke) ischemic attacks [571]. This risk is lower in patients with mild or asymptomatic stenosis [572].[115] Carotid endarterectomy involves the removal of the plaque, and has been repeatedly shown to reduce the risk of stroke *on average* compared to medical management alone (e.g. [575,576]).

However, carotid endarterectomy has paradoxical effects; the procedure itself can cause emboli to break off, and is associated with increased risk of stroke and death. Recent studies indicate that the reduction of stroke-risk resulting from carotid endarterectomy for moderate-stenosis (≤70% stenosis) and asymptomatic patients is *outweighed* by the increased risk due to the procedure [577-580]. As such, while the procedure may reduce risk *on average* by considerably reducing the stroke-risk of the highest-risk severe-stenosis patients, it also increases the risk of stroke for low-risk and asymptomatic patients. Despite these findings, until recently almost half of carotid endarterectomies were performed on moderately stenotic or asymptomatic patients [580-582].

A philosophical problem here is the assumption of a single *true treatment effect*. GRADE asks users to estimate 'the treatment effect' by drawing upon study data[116]—to "*Generate an estimate of [the] effect for each outcome*" ([195], p.385). But in carotid endarterectomy, the treatment has at least two different effects on the outcome of interest (stroke risk) which are in opposition. The treatment reduces the risk of future stroke caused by embolisation, but also increases the imminent stroke risk by undertaking a surgical procedure which can cause embolisation (iatrogenic stroke).[117]

There are two problematic consequences of overlooking heterogeneity in the carotid endarterectomy case. First, moderately-stenotic and asymptomatic patients are exposed to increased stroke-risk due to the failure to take account of heterogeneity, and treatment recommendations based

[114] There are a number of issues with generalising from the findings of large trials of carotid endarterectomy to estimates of the average treatment effect in the target population (see [213,261,570]), not least of which is the inclusion of only the best centres in multi-centre trials, and the use of more sophisticated diagnostic equipment (angiography rather than ultrasonography) to reduce false-positive diagnoses beyond that possible in practice [571]. These are *not* the problems discussed here.

[115] Symptoms include tinnitus, bruits, and TIAs [573]. Asymptomatic stenosis can be detected by imaging, most commonly through (non-invasive and inexpensive) colour-doppler duplex ultrasonography (*cf.* [574]).

[116] In reality, users are assessing the *differential* treatment effect with respect to some comparison treatment. So recommendations should really be phrased as, e.g. "I strongly recommend using X rather than Y", rather than simply "I strongly recommend using X".

[117] The framework of a "true treatment effect" permeates the EBM literature beyond GRADE publications. For instance, Bigby states: *"The confidence interval provides a range of values in which the "population" or true response to treatment lies."* ([335], p.1615), and similarly, Guyatt et al. in the *Users' Guides* state: "*This estimate is called a "point estimate" to remind us that although* the true value *lies somewhere in its neighborhood, it is unlikely to be exactly correct. Confidence intervals tell us the range within which* the true treatment effect *likely lies*" ([190], p.1803, emphasis added).

on the average treatment effect. Second, the estimated treatment effect for the severe-stenosis groups is diluted by the paradoxical effects on the other subgroups. The merits of the procedure for severe stenosis sufferers may be understated as a result, and the strength of recommendation to those patients should be increased. Both patients and doctors clearly benefit from the information that the treatment effect is severely heterogeneous on the outcome of interest (stroke risk), and that symptomaticity and degree of stenosis are strong predictors of effectiveness.

6.4.3.2: Quinidine

An example of this latter situation has been brought to philosophical attention by Williams ([583], p.116). The anti-arrhythmic quinidine was a prominent treatment for atrial fibrillation, an extremely common cardiac arrhythmia [279]. In the 1980s, several studies indicated that quinidine was associated with increased risk of torsades de pointes[118], leading to a higher death-rate among quinidine-takers than amongst other atrial fibrillation sufferers [584]. Consequently, quinidine is generally unused in modern practice.

However, more recent studies indicate that while quinidine severely increases risk of death for patients with a history of congestive heart failure (CHF), and other cardiac structural diseases[119], it also decreases the risk of cardiac death by 30% in patients without history of CHF or other structural heart diseases [586,587]. The average treatment effect which showed quinidine as a dangerous risk-increasing therapy for atrial fibrillation is the result of the beneficial effects on patients *without* structural heart diseases being swamped by the fourfold increase in rate of adverse cardiac events amongst patients *with* history of CHF and structural heart disease [587]. Not only does the average mask the effectiveness of the treatment for some patients, it understates the dangers for others. As a result of reliance upon average treatment effects and underappreciation of the role of heterogeneity, an effective treatment for some of the (very many) sufferers of atrial fibrillation is no longer available.

6.4.3.3: Temozolomide

Williams ([583], p.114) also cites the case of temozolomide as a treatment for glioblastoma. In a prominent study, the detected difference in median survival-time between the ‘temozolomide + radiation’ experimental group, and the radiation-only control group was small: 2.5 months [588]. This

[118] A form of ventricular tachycardia which can degenerate into lethal ventricular fibrillation.

[119] Atrial fibrillation has many causes, common amongst which are CHF and structural heart diseases [585]. But by no means has every atrial fibrillation sufferer’s condition resulted from structural heart disease. The recent studies indicated that a high proportion of the patients in quinidine trials had atrial fibrillation secondary to structural heart diseases [586]. Just one cause of heterogeneity—but an important one, as this example shows—is variation in the etiology of the patient’s condition, and underlying conditions.

may not outweigh the damages of the toxic side-effects of temozolomide chemotherapy for many patients (especially in combination with the toxicity of accompanying treatments). However, the study was recently re-analysed to take account of the activation of the MGMT gene—if the MGMT gene is inactive, damage to the tumour cells due to the therapy could not be repaired. In the MGMT-active subgroup, the difference in median survival-time was only 0.9 months, whereas the in the MGMT-inactive subgroup, the average median survival-time was 7 months longer than given radiation-therapy alone [589]. Coupled with a 23.3% absolute increase in two-year survival rates, this might sway MGMT-inactive patients towards taking temozolomide, and correspondingly discourage MGMT-active patients from the treatment (the meagre benefits would likely be outweighed by toxicity side-effects upon quality-of-life) (see also [590,591]).

Williams uses this case to criticize reliance on statistical significance testing and "*slavish*" adherence to the $p<0.05$ threshold [583]. But the cases also illustrate the practical reality of treatment effect heterogeneity, and the important lesson that average treatment effects radically underdetermine the distribution of those effects. This lesson is independent of how the evidence for average treatment effects arises, whether from RCTs or other sources, and how evidence about variation can be gathered—the point is that average treatment effects alone do not provide all of the information patients and practitioners need. Therefore, basing a recommendation on information about average treatment effects alone, no matter how strong the evidence for the estimate of the average treatment effect, is misguided.

6.4.3.4: Benoxaprofen

Worrall [75,76], Rawlins [97] and Black [148] consider NSAID benoxaprofen (aka. Opren/Oraflex) as a treatment for rheumatoid arthritis as an example of external validity issues. This case also illustrates the importance of information about heterogeneity. Benoxaprofen was promoted heavily because its long elimination half-life[120] meant it only required daily doses [120].

However, when studies into the pharmacokinetics of benoxaprofen in exclusively elderly study-populations were undertaken, it was found that the elimination half-life was much longer for elderly patients, whose renal function is predictably diminished in comparison with younger, healthier

[120] The time taken for the concentration of the active ingredients of the drug in the patient's bloodstream to drop to half the initial value.

patients[121] [594]. Subsequently, reports of fatal hepato-renal failure in elderly patients taking the standard dosage of benoxaprofen amassed, leading to the withdrawal of the drug ([120], pp.726-7).

Worrall [75,76] cites this case as an example of the problem of external validity—applying results from an unrepresentative trial-population to a different target-population. In addition, the case illustrates the problem of heterogeneous effects; differences in metabolism and renal function can lead to very different responsiveness to treatments and vulnerability to their side-effects. Moreover, failure to acknowledge heterogeneity again has two costs—lives lost due to failure to acknowledge heterogeneous pharmacokinetics, and the loss of an effective treatment for younger, healthier patients due to the subsequent scandal.

6.4.3.5: Further Cases

A number of additional cases can reinforce and extend this claim. For instance, research by Kent, Ruthazer & Selker [595] indicates that identifiable subgroups of patients at low-risk of intracranial haemorrhage caused by thrombolysis benefit from rtPA (recombinant tissue-type Plasminogen Activator) therapy even 3-6 hours after acute ischemic attack. This result counters the finding by the large ATLANTIS-B trial that rtPA is not beneficial more than 3 hours after stroke [596]. Their argument hinges on the heterogeneity of rtPA's effects which is overlooked in the average treatment effect reported by ATLANTIS-B. rtPA *is* beneficial even after 3 hours, but this benefit is cancelled out by the accompanying spike in risk for intracranial haemorrhage. For patients with low haemorrhagic risk, rtPA is (on average) beneficial and effective.

A similar result for patients treated with tPA following myocardial infarction (heart-attack) found that (contrary to the small average treatment effect reported by the GUSTO trial [597]) an identifiable subgroup of patients benefit greatly from tPA (25% of the patients receive 66% of the reported benefits of tPA). Again a major determinant of treatment-effectiveness was baseline risk of intracranial haemorrhage [598].

These case-studies illustrate clinically significant heterogeneity occurring in surgical and pharmaceutical interventions, across several fields including cardiology, oncology and rheumatology. Heterogeneity and the causes of variation in treatment effects and side-effects are topics which receive comparatively little research attention. These cases may, therefore, represent only a small subset of the treatments which have heterogeneous effects. One might hypothesise that there are many heterogeneous treatments assumed to be homogeneous (see [254,457,481]). At the very least, given

[121] In studies of patients with mean ages in the range 77-86, average elimination half-lives at the standard 600mg dosage ranged from 101-147.9 hours, far longer than the expected half-life of 30-35 hours found in studies of "normal" patients—generally younger, heavier and healthier [592-594].

the serious effects of overlooking heterogeneity in cases such as carotid endarterectomy and benoxaprofen, assuming that treatments effects are homogeneous unless proven otherwise is a risky strategy.

6.4.4: Heterogeneity—The Challenge to GRADE

Information about heterogeneity of treatment effects is crucial to ensuring the highest standard of care by making best-informed recommendations to patients. Individualized treatment-prognosis predictions require some assumption to be made about the distribution of effects. The models of treatment effect distribution could be evidence-based, or an assumption of a normal distribution could be made. The cases discussed above show that assuming homogeneous treatment effects can lead to poor recommendations to patients—effective treatments can be overlooked and ineffective or harmful treatments can be recommended. At the very least, strong recommendations are not justified on the basis of high-quality information about the average treatment effect alone. Predicting the average effect for a patient is not a truly evidence-based prediction unless the hypothesis that the patient is likely to receive roughly the average effect is based on high-quality evidence.

There are two ways in which GRADE has attempted to include information about variation in treatment effects within its appraisal procedure. First, the GRADE guidelines are explicit that GRADE recommendations must be stratified by population. So, if different populations respond differently to the same treatment, the GRADE appraisal procedure must be applied to the evidence for a treatment effect in each population, and recommendations must be justified separately to each group. Secondly, GRADE's down-grading criteria include "*inconsistency*" [197]. By inconsistency, they mean 'heterogeneity of study results' (as opposed to 'heterogeneity of treatment effects'). Heterogeneity of study results occurs when different trials indicate different directions or degrees of effect [197]. These findings may be troubling for many reasons—they could suggest that some of the trials were flawed, or, if the trials were performed in different populations, that there is systematic heterogeneity of *effects* related to the differences between the populations. However, GRADE does not directly consider evidence of heterogeneity of treatment effects as a factor in assessing evidence of a treatment effect, evidence for a treatment-prognosis prediction, or evidence for a recommendation.

Neither of GRADE's existing approaches to variation allow for full use of information about heterogeneity of treatment effects. In order to see why, it is first necessary to consider the effects information about the distribution of effects and systematic heterogeneity can have upon clinical recommendations. Using information about variation can have four main consequences:

(1) Reinforcement: Reinforcing a recommendation based on the average treatment effect in cases where the effect profile is shown to be homogeneous or unimodal and unskewed.

(2) Undermining: Preventing a strong recommendation due to evidence that the treatment effect is heterogeneous without information about the causes or predictors of different responses, where the risks or rates of undesirable effects are too great; or due to lack of information about heterogeneity. Where there is no evidence relating to heterogeneity, strong recommendations to individuals based on estimates of average treatment effects alone are almost always unjustified.[122]

(3) Supplementation or Replacement: Providing detailed information about the probabilities of different outcomes given a treatment option in addition to, or instead of, a recommendation so that patients can make informed decisions compatible with their level of risk-aversion. In this case, the probability distribution of effects is described to the patient instead of, or as well as, the average effect.

(4) Stratification: Making more nuanced recommendations where the effect profile is systematically heterogeneous by using predictors of different responses to treatments to stratify the recommendations by subpopulation.

Reinforcement results from increased confidence that individuals are likely to experience roughly the average treatment effect. Evidence demonstrating that most patients receive a similar effect reinforces the generalisation of an effect estimate from study results to the target population, and the plausibility of the particular treatment-prognosis claim which underpins a recommendation to the patient. So, evidence that treatment effects are homogeneous increases the quality of evidence for a treatment effect prediction, and can increase the strength of a recommendation. A recommendation may also be strengthened where homogeneous effects are desirable in and of themselves—for instance, in recommending painkillers, clinicians may prefer a drug with a smaller average effect if results are more consistent (see also Section 6.6).

“Undermining” occurs when one becomes aware of heterogeneous and potentially paradoxical effect profiles without the ability to stratify populations to mitigate the harmful responses. In cases like carotid endarterectomy and temozolomide, some patients receive only or primarily detrimental effects. If these were a sufficiently large proportion of the population and the risks were substantial, then one might recommend *against* using the treatment despite a net beneficial average treatment

[122] See section 6.5 for the further argument that population-level recommendations are similar undermined where there is a lack of information about variation.

effect. Recommendations can also be undermined where information about heterogeneity is unavailable. Information about heterogeneity is important to the strength of recommendations, and to whether a recommendation is even appropriate. Therefore, in the absence of information about the distribution of effects, one is uncertain about how strong a recommendation can be justified, or whether a recommendation is appropriate. Schunemann et al.'s ATS GRADE statement accepts that in the absence of important information, recommendations are affected:

> *"when the relevant information is not available, a guideline panel should be more cautious and, in most instances, opt to make a weak recommendation"* ([37], p.607)

As such, given that information about the distribution of effects is important, where that information is unavailable strong recommendations are usually undermined. A weak recommendation may still be justifiable in the absence of information about variation. A strong recommendation might still be justified where the effect is very large or universal[123], but only as a trade-off in terms of uncertainty.

"Supplementation" or "replacement" results from the fact that different patients (and indeed different subgroups of patients) can have different priorities and degrees of risk-aversion. The example of a chemotherapy drug which works in 25% of patients is instructive here. Some patients may prefer to risk the side-effects of the drug for a 25% chance of a cure. Others may reject the treatment, preferring a few months free of the drug toxicity to the chance of complete recovery. Factors such as age, duration of the condition, number of previous interventions attempted, existing health state and the predicted health state given recovery will influence the preferences and risk-aversion of patients. A simple recommendation, strong or weak, cannot take account of these factors adequately. It would be more respectful towards the patients' autonomy to offer the patients information about the chances of experiencing desirable and undesirable outcomes given the treatment, carefully communicated to reflect the strength of the evidence for that information. In this case, a simple recommendation to use or not use the treatment may be inappropriate, in which case it is replaced by conveying the relevant information. Alternatively, a recommendation may be given, but supplemented with the relevant information to allow the patient to weigh the risks according to her own values.

Finally, "stratification" occurs when there is heterogeneity in the treatment effects, and there is some information about the causes or predictors of different responses to treatment. Under such circumstances, recommendations can be stratified. Patients who are likely to receive more than the average treatment effect can be recommended to use the treatment more strongly than would be

[123] 'Universal' effects in this sense capture 'all or nothing' cases, in which either all (or no) patients formerly experienced some outcome, and now *some* do, or some patients formerly experienced some outcome, and now all (or none) do. However, it may be argued in this case that this information in itself is evidence of a homogeneous effect.

justified on the basis of the average effect alone. Patients who are likely to experience lesser effects or more likely to experience serious side-effects can be given weaker recommendations, or may be recommended not to use the treatment.

Now, the two concessions GRADE makes to variation in treatment effects can be considered. First, GRADE tells users to stratify recommendations by population if there is any variation in treatment effects. Prior to beginning the GRADE process, the guidelines require that:

> *"Recommendations must apply to specific settings and particular groups of patients whenever the benefits and harms differ across settings or patient groups."*
>
> ([12], p.1491)

They make it clear that both the practice context and patients' features such as baseline risk and age can affect a treatment's effects. The hope here is that by stratifying according to, for instance, symptomaticity and degree of stenosis in recommendations for carotid stenosis sufferers, GRADE users could take account of the variation in effects of carotid endarterectomy, and stratify their recommendations accordingly.

However, this concession to the importance of variation between different settings and populations offers no role for evidence about how treatment effects vary and what causes such variation. Users are told to stratify recommendations, but not told what evidence justifies stratification or how to appraise this evidence. The GRADE authors cannot intend users to only pay attention to studies performed in the specific population and within the same clinical context. Not only would this severely reduce the evidence available, but it would also exclude a great deal of evidence that *is* relevant. Trials performed in populations with differing characteristics—whether broader, narrower or different from the target population—can be informative where the differentiating factors are not relevant to a treatment's effects on an outcome, or where they provide a baseline for a more nuanced estimate.[124]

So, users need to employ information about how the specific population and context relates to trial populations and contexts to interpret and apply those results. This information is information

[124] For instance, the carotid endarterectomy trials performed by Barnett et al. [575] and Farrell et al. [576] were performed in populations including both moderately and severely-stenotic patients. Given the information that moderately-stenotic patients generally receive a detrimental effect, these trials provide a baseline estimate for the effect in the severely-stenotic patients; we know that the effect must be *at least* as large as the average discovered in the studies, and probably considerably larger. Schunemann et al. accept such reasoning in their presentation of the GRADE system: *"if only sicker patients receive an experimental intervention, yet they still fare better, it is likely that the actual treatment effect is larger than the data suggest."* ([37], p.610)

about heterogeneity, and about the causes, predictors and determinants of different responses. The information that a treatment's effects upon an outcome of interest are homogeneous would strengthen the recommendation in specific subgroups. Information about heterogeneity can strengthen it in some and weaken or change it in others, as we saw above. Therefore, simply directing users to stratify recommendations is not enough. Users of GRADE need to pay attention specifically to information about heterogeneity and, in order to make an evidence-based recommendation, to the *quality of the evidence* for this information. The GRADE framework provides no support for users to do this and no methodology for appraising evidence concerning heterogeneity or causes of variation.

The second approach to heterogeneity within the GRADE framework is to downgrade the quality of evidence-bases which exhibit heterogeneous study results. Heterogeneous study results *may* be a consequence of heterogeneous treatment effects. Where studies are performed in different populations and settings, if their results differ this may be due to systematic heterogeneity. However, this may also be due to bias in some of the studies (see Section 6.4.5, below, for further discussion on the interpretation of heterogeneous results in systematic reviews).

But even if heterogeneous results are due to heterogeneous treatment effects, downgrading the quality of the evidence base is an overly simplistic reaction. In some cases—those described above as undermining—the evidence base for a treatment-prognosis claim is lower quality due to the heterogeneity undermining the clinical significance of evidence about the average treatment effect. But in other cases, a stronger recommendation may be justified in some populations, and the evidence base for a treatment-prognosis predicting a larger treatment effect may be high-quality, while in other populations the evidence base may be high-quality for a prediction of a smaller treatment effect or a higher risk of side-effects. Information about heterogeneity *can* decrease the quality of an evidence base, but it can also increase it. Moreover, the upgrading criteria do not include the possibility of upgrading the quality of the evidence base where there is evidence that the treatment effect is homogeneous. As described above, the evidence base is enhanced by evidence of homogeneity because the predictions based on the average effect are more likely to be accurate and that information is more clinically significant when most patients experience roughly the average treatment effect.

Information about systematic heterogeneity and the causes of that heterogeneity is very important to clinical recommendations and decision-making, as the cases above have illustrated. Information about heterogeneity can reinforce or undermine recommendations, but it can also justify increased confidence in different recommendations for different populations. One challenge for GRADE is to take account of these different effects of evidence of systematic heterogeneity on the appraisal of evidence-bases and the strengths of recommendations. The attempts to take account of heterogeneity in the current GRADE approach are inadequate. But moreover, like evidence for estimates of the average treatment effect, not all evidence relating to heterogeneity is high quality. A second challenge, then, is to provide a system which can appraise the strength and quality of evidence of homogeneity, heterogeneity and the causes and predictors of variation in treatment effects. The next section will discuss whether hierarchical approaches could be used to assess evidence of heterogeneity, and argue that they are not appropriate for assessing such evidence.

6.4.5: Developing GRADE—Hierarchical approaches to Heterogeneity?

This section will argue that this criticism cannot be resolved adequately by the further use of hierarchies within GRADE. First, the problem is not solved by applying the GRADE hierarchy to evidence of heterogeneity in treatment effects. The reasons GRADE ranks RCT evidence as higher quality than observational study evidence do not apply when considering heterogeneity. Other sources of evidence are necessary for a full consideration of variation. Second, constructing a second different hierarchy designed to appraise the quality of evidence about heterogeneity is not a workable solution either. Evidence of heterogeneity should not appraised according to methodology, as a range of methods are needed to provide information about the range of concerns relevant to a full picture of variation in treatment effects.

6.4.5.1: Randomisation and Heterogeneity

One option for developing GRADE is simply to apply the existing GRADE hierarchy to evaluate evidence of heterogeneity and causes of variation. Under such a framework, an RCT evidence base is high-quality and an observational study evidence base is low quality, subject to upgrading and downgrading criteria. However, there is no reason to rank RCT evidence initially above observational study evidence when considering variation. The purported advantages of randomisation only apply to the evidence they provide for an estimate of the average treatment effect, and not to any secondary evidence they may provide for claims about systematic heterogeneity. The evidence of systematic heterogeneity which comes from evidence bases composed of RCTs is actually a form of observational evidence, as this section will show.

There are two main ways in which evidence from RCTs can be used to support claims about heterogeneity in treatment effects: between-trial and within-trial analyses. Between-trial analyses use meta-analysis techniques to test for associations between the reported average treatment effect found by studies and features of those studies—a technique called meta-regression (see e.g. [1,107,484]). For instance, to test the hypothesis that gender might interact with the effects of tocilizumab on rheumatoid arthritis, one could perform a meta-regression of trials of tocilizumab, testing for a relationship between the proportion of trial participants of a particular gender and the size of the treatment effect reported in the trial. Similarly, homogeneous study results amongst trials with heterogeneous populations suggest that there is little systematic heterogeneity.

Heterogeneous trial results could be due to variation in the trial protocols, bias in certain trials, or heterogeneous treatment effects [1,110,292,484]. To use between-trial meta-analyses as evidence for heterogeneity of treatment effects and for a causal or predictive relationship between a patient-attribute and treatment effects requires one to demonstrate that the findings are not confounded by variation in the trial protocols or bias amongst the trials. Statistical techniques to control for some causes of heterogeneous results have been developed (see e.g. [484]). A further problematic factor is the required assumption that the effects of the control treatment are homogeneous—another alternative explanation for, say, larger average treatment effects in trials of predominantly female patients compared to male patients, is that the control treatment(s) in these trials are less effective in women than men.

Within-trial analyses consist of various forms of subgroup analysis (see [477,484,599-601]). The trial population is stratified into subgroups according to some patient-feature. Subgroup analyses test for differences in the differential average treatment effect between certain subgroups. For instance, in Hegi et al.'s re-analysis [589] of Stupp et al.'s temozolomide trial [588], the investigators performed a subgroup analysis in which they stratified by MGMT-activation. Subgroup analyses involve dividing the trial population into subgroups—in this case, MGMT-active and MGMT-inactive—in both the experimental and control arms, and comparing these subgroups [477]. If the average effect differs significantly between the groups, then this may demonstrate heterogeneity of treatment effects, providing evidence for the hypothesis that MGMT-inactivation causes (or at least predicts) high responsiveness to temozolomide therapy. Various statistical techniques can be employed to analyse within-trial variation where data about individual patient features and outcomes is available [1,484]. Both NICE [484] and the Cochrane Collaboration [1] discuss and endorse different approaches to the analysis of variation within trials where individual patient data is available. Unlike between-trial analyses, which can be performed with aggregate trial data about the interaction variable, within-trial analyses require individual patient data concerning both the interaction variable and the outcome. Note, however, that subgroup analyses involve comparisons of subgroup average treatment effects, and do not facilitate direct discovery of individual treatment effects.

There are a number of challenges to both between-trial and within-trial analyses, and EBM proponents are divided about whether meta-analysis and subgroup analysis can provide high-quality evidence (see [1,216,300,477,478,601,602]). Hierarchies which rank meta-analysis highly rank these as a source of evidence for an estimate of an average treatment effect, not as a source of evidence about systematic heterogeneity. There are substantial difficulties in ruling out other causes of study-result heterogeneity in between-trial meta-regression. The quality of meta-analysis depends on the quality of the included studies—as the Cochrane Handbook puts it:

> "*If bias is present in each (or some) of the individual studies, meta-analysis will simply compound the errors, and produce a 'wrong' result that may be interpreted as having more credibility*"
>
> ([1], p.247).

Subgroup analyses are performed on sub-populations, and so can lack the statistical power to detect differences in treatment effects of the larger trial—although researchers can design trials in advance to provide adequately-powered subgroup analyses, at the cost of large increases in sample size (see [600,603]). There has been much concern about the potential for data-mining in subgroup analyses, and these concerns are also generalisable to between-trial meta-analysis. As the number of covariates tested for interaction with the average treatment effect increases, so does the probability of a false positive result (*cf.* [478,489,530,602]). Solutions to this problem include restricting the number of covariates tested, specifying which covariates will be tested in advance, only testing for covariates with a biomedical justification to suspect an interaction with treatment effects, and restricting the use of subgroup and between-trial analyses of variation to hypothesis-generation—i.e. using the findings to inform the research agenda and instigate new studies in the relevant sub-populations, rather than as evidence for systematic heterogeneity (see [570,601,602]).

However, the crucial point from the perspective of GRADE and hierarchies of evidence is that both between-trial and within-trial analyses of heterogeneity are observational studies, even when the data analysed originates from RCTs. As the Cochrane Collaboration Handbook puts it:

> *"Subgroup analyses are observational by nature and are not based on randomized comparisons."*
>
> ([1], §9.6.2)

Patients are not, and cannot be, randomised to different values of the interaction variable. For instance, in the subgroup analysis of Stupp et al.'s temozolomide trial, while patients were randomised to receive temozolomide or placebo, they obviously were not randomised to be MGMT-active or inactive. In the case of between-trial meta-analysis, trials are not randomised to have different proportions of patients from a specific sub-population. So, even where RCT data is used, analyses of systematic heterogeneity must be classified as observational studies.

Due to the serious challenges that present both techniques, one might argue that lowering the initial classification of the evidence-quality from these sources to "low" is justified, even if this prevents any evidence been classified as 'high-quality'. This would flatten the hierarchy. But there are research methodologies which are specifically designed to investigate heterogeneity and the causes and predictors of variation in treatment effects. Such studies deliberately and systematically attempt to control for the different biases which affect research into systematic heterogeneity. One such study was discussed in Chapter 5: Forsblad-d'Elia et al.'s study [125] of the predictors of drug adherence and responsiveness to treatment with tocilizumab for rheumatoid arthritis. Their study is not a controlled trial, but rather employed records of drug adherence and responsiveness to test for patient features which are associated with high and low responsiveness and drug adherence. It is an example of a cross-sectional analysis—data is gathered concerning the whole population of interest at one point, and analysed. Cross-sectional analysis has been omitted from traditional hierarchies of evidence, presumably because such studies are not designed to provide evidence for or against estimates of average treatment effects. But such studies can be and are used to investigate the distribution of treatment effects and correlates of different responses to treatment. Similarly, case-control designs can be used to investigate predictors of different responses—groups are identified in which different responses were found, and tested for baseline differences which could account for the difference.

Even if all such studies are classed initially as low-quality, the hierarchical approach is undermined—the system no longer ranks primarily according to methodology, but merely assumes all evidence is low-quality unless upgraded for some possibly non-methodological reason. Furthermore, the up- and down-grading criteria used by the GRADE system are not calibrated to assess either specifically designed observational studies, or meta-analysis and subgroup analysis. For instance, they class between-trial heterogeneity as reason to downgrade studies, which is inappropriate in this context, as between-trial heterogeneity is the data for many such meta-studies. The next section will present the argument that ranking evidence about systematic heterogeneity by study-design is inappropriate and unworkable.

There are a variety of methodologies designed to study variation and the causes of variation. Evidence can result indirectly from RCTs and other comparative trials. The existing hierarchies, including GRADE, are ill-equipped to assess evidence for variation and causes of variation as the existing justifications for the ranking fail in this context, and the hierarchies omit methodologies specifically designed to test for heterogeneity and predictors of heterogeneous responses.

6.4.5.2: A New Hierarchy for Heterogeneity?

An alternative approach would introduce a new hierarchy to assess evidence of heterogeneity and causes of variation. Evidence for an average treatment effect estimate would continue to be assessed using the GRADE hierarchy, and a novel hierarchy would assess evidence of heterogeneity. This approach resembles that used by many other hierarchy authors, such as the CEBM approach [15,103], Bhandari & Giannoudis' 2006 hierarchy [90], SORT [104], the NHMRC [33], and GREG [284], where multiple hierarchies are offered for different kinds of medical problem, such as treatment, prognosis, diagnosis and aetiology. No such hierarchy has yet been offered for questions of variation, although Roman, Silbersweig & Siu ([299], p.88) have included some relevant considerations such as observational studies of risk factors and inferences across populations from RCTs in their hierarchy (see Fig.34, above), and McAlister et al. included interstudy comparisons of various types and subgroup analyses in their hierarchy which focuses on comparisons between drugs of the same class (see Fig.35, above).

A novel hierarchy is not an appropriate tool for appraising evidence of systematic heterogeneity. The argument for this claim has three stages. First, there are many components to a full picture of the effect profile of a treatment. Sources of evidence for information about each component will vary. Secondly, the quality of evidence from a study of systematic heterogeneity depends more sensitively on the interaction tested and the data available, rather than the study design. Finally, it may often be necessary to combine multiple methods to find reliable information about each component of the effect profile. Methods may provide weak evidence alone, but much stronger evidence in combination with information from other sources. Successfully modelling treatment effect profiles may need to draw from several approaches.

For a full understanding of the distribution of treatment effects, one would ideally have information about the range of potential effects of a treatment, the composition of those effects (i.e. whether the net effect on some outcome of interest is composed of a number of effects on physiological systems, as in the case of carotid endarterectomy), the causes, predictors or determinants of different responses, and the strength of the association between these and the outcome. A full effect profile for a treatment will offer information about each of these, for each outcome of interest. The ultimate goal is a probabilistic model of a treatment's effects which allows predictions of the overall effect distribution, and if possible tailored predictions of individual effects based on a patient's attributes.

The studies appropriate to provide evidence for each part of an effect profile will differ. At the most basic level, information about what potential outcomes can result from a treatment is very valuable. For the most part, detailed effect profiles are currently only collected in terms of side-effects. Unfortunately, the collection of this data is unsystematic. As La Caze points out, many current systems merely rely upon data regarding side-effects collected during clinical trials designed to test for efficacy. As he puts it, they:

> *"rely on the serendipitous findings of randomized trials set up to test benefit hypotheses"*
>
> ([11], p.524)

There are numerous methods which could deliver a more systematic approach to recording both harms and benefits, including longitudinal studies and outcomes research. Guyatt et al. agree that:

> *"much of the evidence regarding the harmful effects of our therapies comes from observational studies"* ([82], p.1293).

They interpret this as a "*challenge*" to EBM (*ibid.*). However, the role of observational studies and outcomes research in collecting data about actual effects is not so much a challenge to EBM as it is a flaw in hierarchical approaches to evidence. Drawing upon this data could be a cornerstone of EBM, rather than a sticking point.

Information about the different *beneficial* effects of treatments, as well as the harms, is also very important. As Section 6.4.1 demonstrated, this information is not collected through RCTs. The data that is important here is relatively simple to acquire through longitundinal studies and outcome databases, in much the same way as information about harms. Information is needed about what outcomes patients have, given treatment, and about the proportions of patients experiencing various responses.[125] Even so, this information can represent a primary evidence base for clinicians—advising patients that '35% of patients taking X reach remission from rheumatoid arthritis' is genuinely useful and informative. Where the 'supplementation' or 'replacement' cases discussed above occur (i.e. where it is not appropriate to make a recommendation or to only offer a recommendation, because the treatment's effects are heterogeneous) this kind of information may be the extent of what the clinician can offer to the patient, or the core data upon which an effect model is build.

Building upon this data, information about the causes and predictors of effects is extremely important. This will come in many forms. First, physiologic studies can provide information about how

[125] These 'responses' could be conceived as discrete categories (high, medium and low response; remission or non-remission, etc.) or as continuous variables (e.g. pain level, quality of life). The desired result is a probability distribution over these, given treatment.

a treatment works and the various effects it has upon physiological systems. On this basis, mechanistic reasoning can be an important source of information about the likely threats to the expected causal pathway of an intervention. One can agree with EBM proponents that mechanistic reasoning provides weak evidence for the claim that a treatment will be 'effective', and yet accept that understanding of biomedical mechanisms gives important information about when to expect different responses. The benoxaprofen is a clear case of physiologic studies from pharmacokinetics demonstrating predictable variation. To be useful, this mechanistic reasoning may need to be very complex and supported by independent evidence at multiple stages, as Howick has argued [58,389]. Demonstrably high-quality evidence of this kind may be rare, but rarity should not discount its importance when available. This mechanistic approach can then be reinforced with pharmacological studies to test whether identified differences do translate to different levels of uptake and elimination, and with biostatistical studies such as subgroup analyses, outcomes research, and other observational studies.

There are methods from data science for generating predictions of likely outcomes and models for the distribution of likely treatment effects, beyond those considered in the EBM literature. They often do not fit into any relevant category in hierarchies. Perhaps the most important category is distributional modeling, in which models are created for plausible distributions of effect sizes, and then validated against data. Models can be formulated according to biological theory or extrapolated from existing data-sets as part of outcomes research. One contemporary technique for predictive modeling drawing upon raw observational data is machine learning (see e.g. [604,605]). Machine learning refers to a class of methods in which a program uses data sets of individual treatment variables and outcomes, along with data about patient attributes. The program generates predictions about the likelihood of different outcomes for future patients by an iterative process, following connectionist computing principles (see e.g. [606,607]). Examples include regression trees [604,605]. Whether such models will be valuable components of scientific inquiry and clinical prediction is beyond the scope of this thesis. There are sophisticated and potentially valuable methods for analyzing variation which go beyond those considered by approaches which focus on point estimates for average effects.

The quality and appropriateness of different sources of evidence and approaches to prediction will depend primarily on the kind of data available, and the interactions being tested. Sometimes one may be looking for system behaviours which would interfere with the delivery or intended effect of the treatment—behavioural effect modifiers. For instance, non-compliance with the treatment regimen could lead to under- or over-dosing, leading to decreased beneficial effects or increased toxicity. The megavitamins case offers a particularly striking example of this in action—Sauve et al.

report that parents often unwittingly react to the symptoms of chronic megavitamin overuse in their children (e.g. fatigue, malaise, anorexia) by increasing the dosage of vitamins ([534], p.1012). Investigating these intervening variables requires a pragmatic study and/or qualitative research (see [255,262,263]), and introduces new challenges such as the Hawthorne effect [231-233].

Biological effect modifiers may also interact with effects—for instance, the availability of uptake and receptor sites, the speed and efficiency of metabolism, etc. These variables will probably be most amenable to high-quality analysis by a combination of various pharmacological studies to identify patterns of physiological responses and corroborate a mechanism for a treatment effect. Genetic variation offers another set of analytic tools, with a range of methods including genome-wide association studies (GWAS) (e.g. [608]). The kinds of studies which can provide high-quality data will also depend sensitively on whether the interaction under study is independent from other potential interaction variables, and whether these variables can be controlled. For instance, age is correlated with rate of co-morbidities and baseline risk, while genetic traits may be less likely to covary with these phenotypic features, and more likely to correlate locally with other genes. The difficulty of controlling for these covariate interaction variables will depend upon the variables of interest and the data sets available. In particular, whether individual patient data is available, or whether studies must address aggregate or average data, affects the kinds of methods which are appropriate and the challenges to validity. As Dias et al. put it, one should expect:

> "*major differences in the quality of evidence from meta-regression that depend on the nature of the covariate in question, and the structure of the data*" ([484], p.11)

There is a range of information relevant to heterogeneity and the causes of variation. Which studies provide high-quality evidence for each component of this information will vary, and will depend upon the interaction being tested and the data available. As such, a single hierarchy based primarily on study methodology is unlikely to be a workable solution to appraise evidence for claims about heterogeneity. To access reliable information about effect distributions and heterogeneity, it may be necessary to draw upon multiple approaches in combination. In isolation, physiological and pharmacological studies, observational studies, meta-regression and mechanistic reasoning may each provide fairly low-quality evidence for claims about heterogeneity. There are serious limitations to each kind of study. Mechanistic reasoning fails to guarantee translation to practical effects. Statistical studies are prone to abuse through data-mining, and to spurious correlations and confounding. As one NICE statement notes, meta-analyses and subgroup analyses will:

> *"inherit all the difficulties of interpretation and inference that attach to non-randomised studies: confounding, correlation between covariates, and, most important,* the inability to infer causality from association.*"* ([484], p.11, emphasis in original).

However, the relationship between these sources of evidence can be synergistic. Numerous philosophical accounts of evidence for causal claims have argued that the strongest causal inferences result from multiple mutually-reinforcing evidence sources. Casuistic accounts of evidence in medicine, discussed by Upshur [609], integrate numerous sources of evidence to form an overall picture of the distribution of effects. Bradford-Hill's guidelines for causal inference (see [137,277,434]) outline eight mutually-reinforcing but individually unnecessary aspects of the support for a causal claim. Howick, Glasziou & Aronson [137] have adapted Bradford-Hill's guidelines to the modern context, classifying them into criteria which address *direct evidence* (i.e. results of statistical studies which establish a correlation), *mechanistic evidence* (i.e. evidence for a causal process—see also [58]), and *parallel evidence* which supports and corroborates the direct and mechanistic evidence. A similar account, albeit without the separation of parallel evidence, is offered by Russo, Williamson and colleagues [60-63] in their epistemic theory of causation (see Chapter 4), which requires both mechanistic and correlative evidence to support a causal claim. Drawing upon multiple sources to model effect profiles is consistent with this literature. Drawing upon particular sources according to their appropriateness to both the task and the available data is similarly a common-sense approach. By contrast, hierarchical approaches to this modeling are not adept at considering either the contextual appropriateness or integration of various sources.

6.4.6: Summary

Evidence about the distribution of treatment effects, and evidence about the predictors and causes of heterogeneity, is important to clinical decision-making and recommendations. In the presence of heterogeneous effects, evidence-based predictive modelling of treatment effects will need to draw upon information about variation and effect modifiers. Where effects are homogeneous, predictions based on evidence for average effect estimates can be strong—but the evidence that the effects are homogeneous is critical to this strength.

GRADE has only treated evidence which suggests that effects are heterogeneous as detracting from the quality of an evidence base, in the form of inconsistency in study results. However, as the cases discussed above and the arguments here show, evidence of heterogeneity can have various effects, including undermining the quality of an evidence base, but also including enhancing it, rendering a simple recommendation inappropriate, or justifying stratified recommendations.

A full system for appraising the evidence bases for recommendations must incorporate evidence of heterogeneity. GRADE's only attempt to do so is to require that recommendations are stratified. But GRADE offers no method for appraising the evidence which justifies these stratifications. A hierarchical approach to evidence of heterogeneity is not workable. Reusing the GRADE hierarchy is inappropriate, as the distinction between randomised and observational studies in the evidence base for claims about heterogeneity is not useful. Introducing a new hierarchy to appraise evidence of heterogeneity into a GRADE-like framework will not solve the problem either, as high quality evidence effect profiles and predictive models may require an evidence base containing evidence from a mixture of methods, and/or the quality of evidence from different kinds of study will depend more closely on the interaction tested and the data available than on the methods used.

This argument severs the link between the quality of evidence bases for claims about average treatment effects, for predictions about patients' likely outcomes, and for recommendations to patients. It breaks the connection between the quality of an evidence base as appraised by GRADE, and the strength and even the direction of a recommendation which can be justified. There could still be a role for a GRADE-like hierarchy within an evidence appraisal system—as a method of appraising the evidence base for a claim about the average treatment effect of a treatment on an outcome in some population. Given suitable additions to the up- and down-grading criterion to account for the role of mechanistic plausibility, commercial bias, and the interactions within the evidence base which GRADE does not currently consider (for instance, replications, the range of methodologies applied in the evidence base, the variation in populations and settings studied in the evidence base, etc.), GRADE *could* provide a reasonable method for appraising evidence for claims about average treatment effects. But the clinical relevance of this information always depends on the distribution and heterogeneity of the treatment effects. A radical overhaul of the GRADE process would be needed to take full account of evidence of variation. This overhaul would require a non-hierarchical approach to the assessment of the quality of evidence bases for treatment-prognosis claims and of the strength of recommendation which can be justified. These are precisely the claims which interest practitioners. Therefore, an adequate adaptation of GRADE for use in Evidence-Based Medicine would have little use for a hierarchical component to appraisal.

6.5: Restricting the Scope of Hierarchical Appraisal

This section evaluates a final option which might salvage the GRADE approach to evidence—restricting the scope of the application of the method to questions of public health and health policy. However, given some basic assumptions about distributive justice in policy and the goals of public health, the same problems will arise. Here, it is argued that ignoring evidence about heterogeneity and the causes of variation is similarly unacceptable in health policy and public health.

The argument above concerned using hierarchies to appraise evidence for treatment-prognosis claims and individualised recommendations. In health policy and public health, practitioners are less concerned with individualised predictions. They make population-level decisions. In health policy, these decisions include whether to license a treatment for a particular purpose, whether to compensate for a treatment under national or personal insurance plans, and what can be permissibly paid for a treatment. This requires evidence about the cost and effects of the treatment within the target population. The net effect of a policy could be calculated by multiplying the average treatment effect by the number of patients treated. In public health, practitioners are often interested in interventions applied to large populations creating net benefits across the population, such as vaccination programmes, screening and behavioural campaigns. One response to the criticism in this chapter, then, is to restrict the application of hierarchies to health policy and public health (see e.g. [23,25,30]). This section shows that a modified form of the argument from Section 6.4 applies also to hierarchies used in health policy and public health, and therefore this restriction fails to solve the problem.

In health policy, a number of approaches could be used to decide which policy is best. These include utility maximisation (the best policy is that which yields the greatest net benefits, or the greatest benefits per unit of cost), maximin (the best policy is that which leaves the worst-off in the most favourable circumstances), and prioritarianism (the best policy prioritises helping those who are worse off, weighing their outcomes more importantly in the calculation). There are other rules beside these, however considering these will illustrate the importance of the general problem.

Suppose one wants to maximise utility in decisions about treatment licensing. RCT evidence indicates that carotid endarterectomy is *on average* beneficial, so one licenses the treatment. However, if the evidence about the variation in benefits from carotid endarterectomy had been taken into account, the policy-maker would see that greater benefit results from a policy which licenses the surgery for symptomatic, highly-stenotic patients, but not for asymptomatic moderately-stenotic patients, as this removes the increased risk of infarction in this subpopulation. It is only by taking account of variation and the causes of variation here that the policy-maker can maximise utility.

Similarly, maximin and prioritarian rules primarily require information about distribution of effects to determine the best policy. Information about the average treatment effects underdetermines both the distribution of the treatment effects within the population, and the distribution of wellbeing (i.e. the end results after treatment). Consider the hypothetical painkillers from Section 6.4.2. If one painkiller reduces pain evenly across the population, and another has no effect for half of the population and a large effect on the other half, but both have the same average treatment effect, then the policy-maker needs to be aware of this. The different drugs are appropriate to different circumstances. The causes of variation are critically important, especially where these are related to the risk or severity of the condition. Those at high-risk and with severe conditions are worse off, and so prioritarian rules favour policies which attend to their needs. If the heterogeneous drug removes pain completely for mild pain sufferers but has no effect on severe pain sufferers, a prioritarian policy-maker would reject it in favour of the homogeneous treatment. However, if it reduces pain substantially for the severe-pain sufferers but not for the mild-pain sufferers, prioritarian and maximin rules favour it.[126]

Crucially, this is even true in some cases where the average treatment effects differ. One such case is outlined in Figure 39, below. In this case, half the patients experience mild pain, half severe. Drug A has homogeneous effects, and would be preferred under prioritarian rules to Drug B, even though they have identical average effects, because A leaves the severe pain sufferers better off. Yet Drug C is preferred over Drugs A and B despite having a lower average effect because it substantially decreases severe pain, leaving the worst off in the best circumstances of the three. Given a policy decision where only one drug can be funded, or only one drug can be given to all patients, C is preferable on a maximin account. A pure utility maximization account would favour A or B, but a preferable policy here would be to use B for mild pain sufferers and C for severe pain sufferers. Here, the average effect has no effect on the decision.

[126] This discussion assumes that only one of the painkillers can be licensed for all. Of course, a more promising approach would identify the subgroup in which the heterogeneous painkiller is effective and use it there, using the alternative homogeneous painkiller in all other patients.

Drug	Effect on Mild Pain	Effect on Severe Pain	Average Effect
A	-3	-3	-3
B	-4	-2	-3
C	0	-5	-2.5

Figure 39: A table showing the importance of systematic heterogeneity in treatment effects for policy-makers. Three painkillers, A,B, & C each have different effects upon mild and severe pain sufferers. For the purpose of this example, assume that every pain-sufferer experiences either mild or severe pain, and that 50% of sufferers experience mild pain and 50% experience severe pain. The values are given in the effect of the painkiller on a 10-point pain scale, where 10 is the worst pain and 0 is no pain. Assume that mild pain registers 4 on the scale, and severe pain registers 8.

Under whichever rule we adopt, there are cases in which policy-makers will make better decisions if they have access to information about heterogeneity and the causes and predictors of different responses. Moreover, wherever this information is omitted or unavailable, the evidence that the policy adopted is the best policy is weak because the policy-maker has no evidence to rule out alternative policies.

The same reasoning applies in public health cases. If a vaccination or screening programme helps those at low risk while exposing those at high risk of a undesirable outcome to additional risk, then *even if* it reduces the incidence of the outcome overall, it would be rejected under a number of decision rules such as maximin. Even when simple utility maximisation is used, information about heterogeneity and causes of variation can improve the public health interventions constructed. Targeting those who benefit the most will be, in some cases, a more effective and more cost-effective strategy, especially where it is cheap and easy to identify those most likely to benefit (*cf.* [254]).

Therefore, similar issues to those raised above arise when applying hierarchies to public health and health policy decisions. Information about average treatment effects underdetermines the distribution of treatment effects. Where the distribution of treatment effects is relevant to the decision, information about average treatment effects underdetermines which decision should be made. There will be cases in which the distribution of effects is relevant to both health policy and public health decisions. Therefore, information about average treatment effects alone is not enough even in domains of medicine where interventions and policies are targeted at the population level, not the individual.

6.6: Conclusion

This chapter has argued that there is a very limited role, if any, for hierarchical appraisal of evidence in clinical practice. Atomistic hierarchies of evidence fail to take account of the fact that other information can affect the strength and quality of the evidence a study provides for a particular claim. Suitably sophisticated hierarchies of evidence bases can resolve this problem to some extent.

However, most hierarchies of evidence bases are too simple to be tenable, equating the quality of evidence bases with the strength of a recommendation. These overlook the importance of negative evidence, and the possibility that strong, high-quality evidence *against* a hypothesis may be different in properties and methodology to strong, high-quality evidence *for* that hypothesis. In particular, evidence that a treatment effect is not biologically plausible may be extremely strong counter-evidence against the hypothesis that a treatment is effective. Most hierarchies of evidence bases have no capacity to deal with this asymmetry.

GRADE is the most sophisticated hierarchy of evidence bases currently available. There is some machinery available within GRADE to account for differences between positive and negative evidence, in the form of separate up- and down-grading criteria. It might be possible to develop GRADE to account for both the criticisms directed at atomistic hierarchies of evidence, and the criticism of simple hierarchies of evidence bases. This would require substantial reworking of GRADE. Even so, there are some challenges which remain unaddressed. For instance, downgrading the quality of a positive evidence base where there is evidence of biological implausibility does not fully account for the role of negative evidence. That is, a low-quality evidence base for an effectiveness claim is not the same as a high-quality evidence base for an ineffectiveness claim.

But the GRADE hierarchy and its less sophisticated alternatives are each susceptible to a further objection—that they do not account for variation in treatment effects. Information about the distribution of treatment effects and the causes and predictors of different responses (both beneficial and harmful) to treatment can be very important in clinical practice, as the case studies in Section 6.4.3 illustrated. Modelling the likely effects of a treatment where effects are heterogeneous is challenging. Predicting an effect similar to the average is an overly simplistic response liable to underestimate both benefits for some patients and risks for others. But the evidence considered by GRADE and other hierarchies is just that—evidence for estimates of the average treatment effect in some population.

A hierarchical approach does not allow for the integration of multiple sources of evidence about variation and systematic heterogeneity to model the likely distribution of treatment effects and predict likely outcomes for individuals. In clinical practice, and therefore in Evidence-Based Medicine, the most important claims are just such claims about likely effects—treatment-prognosis claims and recommendations to patients. A hierarchical approach to the analysis of variation is not workable, and thus neither is a hierarchical approach to the appraisal of evidence bases for the key claims that interest practitioners and EBM proponents—at least where heterogeneity of treatment effects is a factor. Where effects are homogeneous, the evidence for this homogeneity is a crucial component of the support for a prediction.

Therefore, hierarchies of evidence have systematically overlooked or undervalued evidence which is of high clinical importance in making predictions and recommendations. Hierarchical approaches are not suitably adaptable to account for this omission. Whether an account of evidence appraisal which focuses on integration of multiple sources, on predictive modelling of effects, or a mixed approach which follows different lines of appraisal depending on whether a treatment has heterogeneous or homogeneous effects, is preferable is beyond the scope of this thesis. Here, it is only claimed that hierarchical approaches to evidence appraisal alone will underperform for clinical applications.

Moreover, this problem is not resolved by restricting the application of hierarchies to population-level questions. As Section 6.5 showed, the same problems of variation and heterogeneity are again applicable at this level.

Chapter Seven:

Conclusions:
Heuristics and EBM after Hierarchies

Contents

The only prominent interpretation of hierarchies which remains after the criticisms developed in Chapters 5 and 6 is the heuristic interpretation defended by Howick et al. in their *CEBM* hierarchies [15,16,103]. This chapter will show that defenses of hierarchies as heuristics depend upon empirical questions which have not been studied. The success of a heuristic depends on whether it allows users to achieve their goals more often than rival approaches. A heuristic interpretation could be defensible if empirical evidence confirms that using hierarchies improves clinical practice. But given the concerns raised in Chapter 6, it is not satisfactory to assume that hierarchies are an effective decision-aid unless proven otherwise. The burden of proof is on the proponents of hierarchies.

Finally, the case for hierarchies as heuristics is more sympathetic if there is no pragmatic alternative to hierarchical approaches. If a holistic effect-modeling approach to evidence such as that considered in Chapter 6 cannot be implemented in practice given constraints on time and resources, then hierarchies may be the best available solution. Section 7.2 will argue that hierarchies are not the only plausible practical approach to evidence, sketching systems which take a more holistic, non-hierarchical approach without considerably adding to the complexity of appraisal systems. Furthermore, given current trends towards guidelines-based medicine in EBM, a heuristic may no longer be necessary for clinicians to access and use evidence in practice.

7.1: HIERARCHIES AS HEURISTICS

In their guidance document for the 2011 'CEBM Levels of Evidence' [15], Howick et al. state that their hierarchy should be:

> *"used as a heuristic that clinicians and patients can use to answer clinical questions quickly and without resorting to pre-appraised sources"* ([16], p.1).

They define heuristics as:

> *"rules of thumb that helps* [sic] *us make a decision in real environments, and are often as accurate as a more complicated decision process"* (*ibid.*).

A heuristic is a decision-aid—a shortcut which allows users to omit certain complexities of a full evidence appraisal and form a more rapid opinion of the evidence base for a claim. In particular, heuristics work under normal conditions, but may fail in unusual circumstances [317,363]. A heuristic interpretation introduces modifiers into the interpretation—'*Under normal conditions*, evidence X is *probably* Y-quality evidence for hypothesis H'. As such, Howick et al. accept that hierarchies will not always provide an accurate appraisal of the evidence base for a claim. They do not specify whether it may under- or over-estimate evidence quality, or both. They repeatedly emphasise the probabilistic modification in interpretation, for instance:

> *"The* likely *strongest evidence is* likely *to be found furthest to the left of the Levels, and each column to the right represents* likely *weaker evidence"* ([16], p.3, emphasis added).

Justifications for heuristics and decision-aids are usually purely empirical [317,363]. There may be theoretical reasons to suspect that complexities can usually be ignored, but ultimately a heuristic is only as good as the evidence that it achieves its goal. Howick ([59], p.171) has given examples of working heuristics and decision-aids, such as the Ottawa Ankle Rules for radiography [610]. In each case, medical and philosophical justification for the rules is extraneous—what matters is that patients are better off when the rule of thumb is followed than where clinicians use their own judgment.

A heuristic approach could solve the problems raised in Chapter 6 in one of two ways. There could be empirical evidence that *under normal conditions*, systematic heterogeneity of treatment effects does not occur or does not significantly affect the recommendations clinicians should make to individuals. If this were the case, then arguments about heterogeneity would not undermine the justification for a 'normal conditions' modified interpretation of a hierarchy. Alternatively, if it could be shown that using hierarchies achieves the goals of EBM in practice, despite omitting information about variation, then a hierarchy heuristic would be justified. In this case, the argument is not avoided, but a purely instrumental justification for the continued use of hierarchies is possible. Each empirical claim is reviewed here in turn, arguing that neither question has been adequately studied to support an empirical justification for the continued use of hierarchies as a heuristic.

7.1.1: The Prevalence of Systematic Heterogeneity in Practice

As yet, studies of the prevalence of heterogeneous responses to treatment are uncommon. A number of cases of heterogeneity, in which information about heterogeneity clearly affects the evidence-base for clinically-relevant claims, were highlighted in Chapter 6. These examples come from a range of fields, and from both surgical and pharmaceutical interventions. However, these case-studies alone do not demonstrate that heterogeneity is commonplace. Moreover, the argument that hierarchical approaches lead to an incorrect appraisal of the evidence base depends upon an assumption about what the correct appraisal of that evidence would be. For instance, in the case of temozolomide, it was claimed that the correct appraisal of the evidence base is that there is strong evidence in favour of recommending temozolomide to MGMT-inactive patients, and some evidence against recommending the treatment for MGMT-active patients. But one could equally defend a hard-line hierarchical approach and claim that the evidence amassed in favour of this view is unreliable and misleading. The only recourse here is to appeal to common sense, current oncological practice and clinical consensus. In other cases, such as carotid endarterectomy, while the literature may approach consensus with the view of the evidence presented here, clinical practice does not yet broadly concur [580]. There is, as in the case of empirical justifications for hierarchical rankings, circularity in the argument that cases of heterogeneous treatment effects provide counterexamples against hierarchical appraisals of evidence.

However, Howick et al. accept that common sense and clinical judgement play important roles within evidence appraisal. Their heuristic approach requires clinical judgement of whether a hierarchy is likely to be applicable:

> "*the Levels* [must] *be interpreted with a healthy dose of common sense and good judgment*"
>
> ([16], p.4).

A heuristic approach is not intended to be a full replacement for a detailed appraisal by experts, and therefore a consensus of experts whose judgements are firmly based in evidence does provide a challenge to their account—*if* these cases occur in the course of normal practice.

Explicating 'normal conditions' is particularly challenging due to the variation across medical specialisations and fields. What is commonplace in public health may be rare in oncology. Treatments in some fields may be more prone to importantly heterogeneous effects than in others. Studies in some fields may be more prone to certain biases than in other fields—for instance, in fields where it is not generally possible to blind patients and practitioners to treatment allocation in trials, there may be a greater risk of treatment bias amongst studies than in fields where blinding is relatively straightforward.[127] As a result, it may be the case that a hierarchy as a heuristic is defensible in one field because there is low heterogeneity of treatment effects amongst the common interventions, but not in another field where most treatments have strongly heterogeneous effects. Measuring the heterogeneity of treatment effects within fields is challenging, and no major attempts to do so appear in the literature.

There is little *prima facie* reason to assume that systematic heterogeneity is not commonplace amongst commonly-used medical interventions across a wide range of fields. Many authors accept that heterogeneity is real and widespread. As previously mentioned, Guyatt et al. state that *"few, if any, interventions are effective in all patients"* ([82], p.1292). Similarly, Straus & McAlister accept that:

> *"the universal occurrence of biological variation hampers attempts to extrapolate evidence"* ([612], p.838).

At the very least, the side-effect profiles of most interventions are widely-regarded to be heterogeneous [613]. It seems strange to presume that side-effects will exhibit heterogeneity while primary beneficial effects remain homogeneous.[128] Heterogeneity of treatment effects *could prove to be* scarce, and treatment effects could tend to be normally distributed around the mean effect. However, given the clinically-significant information discovered about prominent treatments such as carotid endarterectomy and temozolomide, it is unwise to assume that treatments' effects are homogeneous unless proven otherwise. Studies of variation in treatment effects should be prioritised, not only because they bear on clinical decision-making as explained in Chapter 6, but because information about the scale of heterogeneity within normal clinical practice is critical to the questions

[127] Where surgical interventions are the norm, for instance, blinding may be impossible—although some trials have been performed with sham surgery controls this is now broadly judged as unethical (e.g. [611]). Similarly, where procedures are expensive or invasive, blinding may be impractical (see e.g. [69,187,272]).

[128] This discrepancy is likely due in part to the way data about primary effects and side-effects is collected. Primary effects are measured as average treatment effects through comparative clinical studies, while side-effect data is generally amassed from longitudinal observation and/or reports from trials, so is more likely to be reported as a frequency within a population rather than an average effect (see [11]).

of which approaches to evidence appraisal are workable and whether hierarchies provide a sufficiently reliable heuristic.

7.1.2: Empirical Success

The second potential argument for hierarchies as heuristics is that they achieve the goals of EBM in practice, even if they lack theoretical justification. No studies have been reported which test whether the use of hierarchies improves the integration of research evidence into practice, the recommendations which clinicians make, or patients' outcomes. Tobin [614] reports that no studies validating hierarchies have been undertaken. No studies appear in a systematic search of PubMed for "hierarchy of evidence". However, a limited number of studies have been performed evaluating the effect of EBM training as a whole, which might be marshalled in support of this claim.

Most of the attempts to provide empirical support for hierarchies revert to claims that empirical evidence demonstrates that observational studies provide lower-quality evidence than do RCTs (e.g. [138]—studies commonly cited include [369-373,615]), which are unconvincing (see Section 4.1.1). A similar tactic is to demonstrate that explicitly evidence-based guidelines score higher on quality criteria than do consensus-based guidelines. A comparison of eight breast cancer guidelines assessed using five different quality scales found that explicitly evidence-based guidelines scored higher on quality indexes than consensus-based guidelines, but that there was little disagreement in recommendations between the guidelines [616]. However, this evidence only favors explicit appraisal of evidence, not a particular approach to appraisal such as hierarchies.

Some studies evaluate whether clinicians who are given EBM training display better appraisal skills. Saint et al. found that epidemiological training had little effect on whether clinicians read and appraise trial data, reporting that most clinicians:

> *"rely on journal editors to ensure that what they are reading is methodologically sound"*
>
> ([617], p.883)

A systematic review of studies of EBM and critical appraisal training found that students showed a significant increase in knowledge of appraisal procedures but not in attitudes, skills or behaviours [618].[129] The same reviewers found no studies assessing the effect of critical appraisal on patient care [619]. Straus et al. [10] report studies which claim that 72% of clinicians actively use evidence-based guidelines, but less than half of these reported understanding critical appraisal tools [620]. Another

[129] By contrast, the use of journal clubs in medical education fared better in a similar systematic review [302].

study showed that, when viewed from the perspective of a junior doctor, following EBM literature searching procedures led to different treatment or diagnostic decisions in 23% of cases [41]. Of course, this provides little if any evidence that quality of care improved as a result.

Straus & McAlister [612] identified a lack of empirical evidence for EBM and hierarchies as one of the common criticisms in the literature and in clinician surveys. They admit that no trials of EBM have been performed, and argue that there are methodological and ethical limitations which prevent such trials. Their objections seem to be to studies which would compare clinicians who use evidence with a control who were not allowed to access evidence. However, clearly studies which compared different approaches to evidence-appraisal would be easier to perform in an ethical way. They claim that outcomes research (e.g. [580,621,622]) which shows that *"patients who receive proven efficacious therapies have better outcomes than those who do not"* ([612], p.839) provides an empirical basis for EBM. However, they only marshal a couple of examples in favor of this claim, and it is not clear why this would support the particular approaches to evidence endorsed by the EBM movement, including hierarchies of evidence.

One of the cases which Straus & McAlister cite as evidence in support of EBM is Wong, Findlay & Suarez-Almazor's study of carotid endarterectomy [580], which shows that 20% of carotid endarterectomies were performed inappropriately, and almost half in patients where surgical appropriateness was unclear, and concludes that:

> *"high complication rate possibly negated any overall surgical benefit in the large group of asymptomatic patients."* ([580], p.891).

This kind of study, ranked as low-quality by most hierarchical approaches including Howick et al.'s *Levels of Evidence* [15], provides precisely the kind of information which Chapter 6 argued is overlooked and underemphasized in hierarchical approaches.

This exhausts the evidence which has been cited by EBM proponents in support of their appraisal methods. There seems to be no convincing evidence to disagree with Tobin [614] or Every-Palmer & Howick's [243] conclusions that there is little evidence that hierarchies work in practice, or with Buetow et al.'s statement that the effectiveness of EBM is unknown ([360], p.401). Certainly, there is insufficient evidence to conclude that hierarchies are or are not good methods of accessing high-quality evidence, integrating that evidence into practice, improving recommendations to patients, or improving patients' outcomes. Evidence that hierarchies improve performance on at least one of these measures must be provided before an empirical case for hierarchies as heuristics will be at all convincing.

A complementary problem for a heuristic justification for hierarchies is the lack of a clear comparator. Straus & McAlister [612] claim that studies of whether EBM improves patient care are unethical because the control group would have to practice without drawing upon research evidence at all. But this 'no evidence' control would make a straw man of the anti-hierarchy position. Opponents of EBM or of hierarchies of evidence are not suggesting ignoring evidence in practice. Rather, they defend alternative models of evidence appraisal (e.g. [55,154,609]). Hierarchies of evidence could be tested against clinicians' own informal approaches to evidence appraisal. If hierarchies were shown to correlate with improved results in this comparison, then this would support an empirical case that a systematic method of appraisal improves performance. Of course, however, other systematic approaches to evidence appraisal could be formulated which might outperform hierarchies on those same measures. An argument that a hierarchy improves practice compared to unsystematic appraisal does not entail that a hierarchy is the best critical appraisal approach. Section 7.2 will elaborate on what such rival critical appraisal approaches might involve.

Given the arguments of Chapter 6, the burden of proof is surely on the proponents of hierarchies to provide evidence that their systems achieve their goals. A heuristic defence without empirical support is no defence at all.

7.1.3: User-friendliness and other criteria for a workable Heuristic

There are other empirical considerations which are relevant to the question of whether hierarchies are effective heuristics. Are hierarchies easy to use? Are they significantly less resource-intensive than a full appraisal of the evidence (i.e. are they actually good shortcuts)? Is the output of a hierarchical heuristic replicable and consistent across users? If hierarchical approaches are not sufficiently simple, fast and consistent, then they are unlikely to be strong contenders for a pragmatic appraisal procedure within clinical practice.

The limited evidence which is presently available casts some doubt on the user-friendliness of hierarchies. First, the breadth of hierarchies available and the range of interpretative assumptions consistent with most hierarchies as illustrated in Chapter 2 suggest that many hierarchies could be interpreted quite differently by different users. This would reduce the consistency of hierarchies' outputs. One study of Evidence-Based Program Registers (EBPRs)[130] found remarkably low

[130] EBPRs are databases of programs which meet specified evidentiary standards, e.g. the What Works Clearinghouse [623] and the National Registry of Evidence-Based Programs and Practices [624]. These institutions use explicit criteria to assess the evidence-base for programs, and list programs which meet those criteria as an accessible resource for policy-makers (see [625,626]).

consistency between ratings of the evidence base for healthcare programs across 20 different EBPRs, most of which used hierarchies of evidence to perform their critical appraisal [625].

A lack of guidance in the use of hierarchies undermines the simplicity of the approach. Where the intended interpretation is more carefully specified, as in the case of GRADE, approaches are often complex and require detailed explanation. For instance, the GRADE Working Group recommend one-to-three day seminar courses to learn the principles of applying their approach [627]. The recommended introductory series for new users of GRADE consists of six papers (see [48]), while the full guidelines for using GRADE runs to twenty articles (see [195]). The SIGN developers state that a seminar series is often required for users to understand and apply their hierarchical critical appraisal procedure ([134], p.26). The learning approaches to critical appraisal identified by Coomarasamy & Khan [619] similarly require a large time commitment.

Moreover, GRADE incorporates a number of up- and down-grading criteria which must be considered. For a truly evidence-based application of GRADE, users would need to check for evidence of each of these criteria, which quickly increases the burden of appraisal to resemble that needed in a full holistic literature review. GRADE requires separate appraisals for each outcome of interest, which again suggests that using their approach will be quite resource-intensive, as Howick et al. agree:

> "*what GRADE has gained in accuracy, it may have lost in simplicity and efficiency*"
>
> ([16], p.1).

Alexander et al. [206] conducted a review of all WHO guidelines published in 2007-2012. The WHO has formally adopted GRADE [4,38], although only 37% of guidelines actually used the system. Alexander et al. found that 33% of strong recommendations made by the WHO were based on evidence graded as "low-quality" by GRADE, while a further 22.5% of strong recommendations were based on "very low-quality" graded evidence. They do not offer potential explanations for this. However, there are a number of possible explanations, each of which challenges the usefulness or applicability of GRADE. Perhaps the system was too difficult to apply even for WHO guideline developers. Alternatively, the results of GRADE may have low replicability such that the WHO guideline developers and Alexander et al. disagree about the correct GRADE appraisal. Finally, the WHO guideline developers may have disagreed with the GRADE appraisal and overridden it, perhaps due in part to evidence from omitted sources. Whatever the reason, the WHO's experience with GRADE suggests that the system is often not being used and not used according to the GRADE Working Group's intended procedure, even by an organization which has formally committed to it.

Of course, the GRADE approach is not designed to be a quick or user-friendly heuristic. Howick et al. [15,16] noted this problem and developed their *CEBM* hierarchy to fill the gap. But little evidence is available to validate the consistency or user-friendliness of other hierarchical approaches. In addition to a lack of evidence that hierarchies achieve their goals, there is a lack of evidence that they are a practical, pragmatic solution.

In the absence of a full empirical justification, the only argument which remains for using a hierarchical heuristic is an absence of alternatives. If there are no plausibly practical critical appraisal approaches which could fill the current role of hierarchies, then there is an argument for hierarchies as a stopgap in lieu of a theoretically or empirically justified system.

7.2: ALTERNATIVES TO HIERARCHIES

This section suggests that there is no reason to believe that a practical heuristic cannot be designed which at least takes account of the role of information about heterogeneity and takes a more holistic approach than traditional hierarchies. Two outline designs are sketched here—a non-hierarchic adaptation of GRADE, and a system based on multiple-realisability of a strong evidence-base for a clinical claim. These are only a couple of alternatives—others include checklist-based systems like CONSORT [154] and STROBE [313], or QATs as Stegenga has called them [152]. Finally, it is suggested that given the current philosophy and goals of the Evidence-Based Medicine movement, it may not be necessary to supply a heuristic for evidence-appraisal at all—clinical practice may move from an 'evidence-based' approach to a 'guideline-based' approach, removing the need for quick and simple appraisal heuristics.

A GRADE-like system could be developed which rejects hierarchical appraisal. Hierarchies were involved in GRADE in the appraisal of the evidence-base for hypotheses about treatment effects on each outcome. The problems for GRADE were the omission of information about heterogeneity, and the initial assignment of a high quality level to RCT evidence and a low quality level to observational study evidence. The system could be adjust to remove these problems by removing the initial hierarchical appraisal altogether.

Under such a system, the user initially assigns a very-low quality level to the evidence base. She can then upgrade or downgrade this estimate according to a range of criteria which relate to the composition of the evidence base. These criteria would include the presence of randomised trials or systematic reviews with clearcut results, but could also include criteria relating to evidence that the treatment effect is homogeneous in the target population, criteria relating to evidence of publication bias, evidence for a plausible causal mechanism, replications of key studies, independence amongst the studies forming the evidence base, and so on. Evidence of heterogeneity in treatment effects would downgrade the evidence base *unless* there was also evidence that the patients in the target population were generally high-responders to the treatment, in which case the strength of the evidence-base is upgraded. Similarly, inconsistency amongst study results would detract from the strength of the evidence base unless evidence that this variation in results was caused by heterogeneous treatment effects was provided.

This model may be slightly more complicated than the original GRADE approach. It may be quite resource-intensive, requiring users to read a number of studies within the medical literature to apply the method fully. Therefore, there may need to be accompanying guidance on priorities where resources are scarce. The approach has a number of advantages, however; it allows for heterogeneity to be both good and bad within an evidence base; it allows for more judgment and common sense and does not reinforce the tendency to omit any evidence not explicitly considered; it acknowledges the asymmetry between what counts as strong evidence for the claim that a treatment works (up-grading criteria) and strong evidence against this claim (down-grading criteria).

A simpler approach might rely on a series of descriptions of evidential situations which are classified by strength. This approach accepts that a high-quality, strong evidence base is multiply-realisable—there are many ways to reach a high level of evidence for a hypothesis. One such high-level evidence base is an evidence base composed of RCTs or systematic reviews consistently supporting the claim that the treatment is effective, plus evidence that the treatment's relevant effects are homogeneous in the target population. This suffices (as far as a heuristic goes) for a high level of evidence. Other high level evidential situations would include unanimous RCT or systematic review support, evidence that the treatment effect is heterogeneous, with evidence that the target population tends to benefit more than the average patient (i.e. that the target population possesses a predictor or cause of high response). One could envision other situations, including situations where the evidence base consists of observational studies that show a large homogeneous effect, etc. The clinician can then skim these 'high-quality evidence situations' and determine whether any applies. A particular advantage of this approach is that it allows clinicians to quickly identify what evidence is missing from their current knowledge of the evidence base which would allow them to claim a high level of evidence. For instance, if the clinician knows that there is RCT evidence supporting the claim that the treatment works, and knows that the treatment has heterogeneous effects, then such tables would inform her that she needs to search for evidence concerning the causes and predictors of high responsiveness to treatment and check whether her patient(s) fall into those categories.

Finally, a GRADE-like approach could be used which adopts criteria more reminiscent of Bradford-Hill's guidelines [277,434]. Howick, Glasziou & Aronson [137] have already suggested such a move. Their three-part reworking of Bradford-Hill's guidelines includes direct evidence from statistical studies, mechanistic evidence of a plausible causal process from treatment to effect, and parallel evidence which supports the direct and mechanistic evidence. The exact details of such an appraisal technique, and its integration with GRADE, are left vague in Howick, Glasziou & Aronson's paper. But the suggestion is again a plausible, holistic picture which allows for evidence to interact and for multiple strands of evidence to be involved in justifying a hypothesis or recommendation.

Each of these approaches are loose sketches of a more complex full appraisal technique. They could be pared down or elaborated to fit with time constraints. The point here is not to suggest the right system to replace hierarchies, but to substantiate the claim that there *are* plausible pragmatic alternatives to hierarchies which take a more holistic view. Again, empirical research would have to arbitrate on the question of which approaches are workable and which lead to the best standards of care.

But a fully-realised alternative heuristic may not be necessary for the most part in modern EBM. As emphasised in Chapter 3, the approach defended by many EBM proponents has evolved since its introduction in the early 1990s. Some proponents like Gordon Guyatt no longer believe that individual practitioners need to be fully-fledged evidence appraisers within their practice [73,138]. The GRADE Working Group explicitly take this view:

> *"It is not practical for individual clinicians and patients to make unaided judgments for each clinical decision."* ([12], p.1490)

Karanicolas, Kunz & Guyatt state that *"systematic reviews … lie at the heart of EBM"* ([138], p.1067). Contemporary EBM approaches tend to place the emphasis on trained, experienced evidence-appraisers working for bodies like the Cochrane Collaboration, NICE, or the WHO to perform complete and detailed appraisals and produce guidelines. The WHO recently echoed this view, stating:

> *"Assessing the evidence and developing evidence summaries is a specialized task that is best done by a methodological expert"* ([4], p.37)

The future of medicine for the ordinary practitioner may not be one of 'Evidence-Based Medicine' in the traditional mode, so much as 'Guidelines-Based Medicine', where the guidelines themselves are evidence-based. If this is the case, then appraisal heuristics are only really necessary where guidelines are missing. In these situations, it might be hoped that practitioners would be able to dedicate more time to a fuller assessment of the evidence base. But for the most part, they will be engaged in applying the findings of evidence-based guideline groups. If any hierarchical search heuristic is needed, it may be one more resembling Brian Haynes' '4S' and '5S' approaches [101,145,146], which rank different resources for pre-appraised evidence.

The arguments in chapters 5 and 6 show that guideline developers should be taking a more holistic approach to evidence appraisal than hierarchical approaches allow. Guidelines should not be produced on the basis of hierarchies. Heuristics are designed to help practitioners access and use evidence. But given the EBM movement's new focus on guidelines in practice, a heuristic may only be necessary in those cases where guidelines are unavailable or outdated. It is in these circumstances alone, and subject to empirical evidence that a hierarchy is fit for purpose, that hierarchies of evidence may have a role left within Evidence-Based Medicine.

7.3: Conclusion

This thesis presented an argument that hierarchies of evidence are a poor basis for assessing the evidence for claims made in clinical practice. The EBM movement focuses most of its attention on claims clinicians make in practice—predictions about treatments' likely effects on patients, and recommendations to use or not to use those treatments. So, hierarchies of evidence are a poor method for evidence appraisal in EBM.

The central controversial claim in this argument is that hierarchies of evidence are a poor basis for evidence appraisal. Many arguments have been put forward to attempt to demonstrate that hierarchies entail false claims about evidence quality and strength. However, for the most part these arguments have addressed particular hierarchies interpreted in a specific way. There is a great deal more variation and flexibility in hierarchical approaches to evidence appraisal than has been generally recognised. In order to present a convincing argument that hierarchical appraisal is not fit for purpose, and to avoid making a straw man of EBM proposals, the range of potential hierarchy interpretations needs to be mapped and addressed.

To facilitate this, a new framework for the analysis of hierarchies is necessary. Chapter 2 presented this framework, drawing upon existing literature from Bluhm [55] and La Caze [11], but ultimately offering a novel framework including six dimensions of interpretative assumptions. Assumptions in each of these dimensions are needed to extract information about evidence from a hierarchy. But authors have offered hierarchies which are conducive to different interpretations. Proponents and critics have presented a range of interpretations. The interpretative assumptions have major consequences for hierarchical appraisal of evidence, and thus for any policy or practice based on such an appraisal. Nevertheless, awareness of the variation in hierarchies and ways of interpreting those hierarchies is rare in the EBM literature.

Drawing upon this framework, historical trends in the interpretation of hierarchies defended by EBM proponents can be identified. Chapter 3 argued that the development of EBM hierarchies was characterised by an initial conservative approach—applying hierarchies only to epidemiological evidence, and with conditions. In the late 1990s and early 2000s, interpretative assumptions were strengthened—hierarchies ranked *all* evidence, including evidence from laboratory studies, clinical experience and mechanistic reasoning. Non-epidemiological evidence was subordinated to the lowest level of hierarchies, and many of the qualifications were removed. Since the development of GRADE from 2004 [12], the general trend has been back towards more sophisticated interpretations.

However, to date most philosophical work has focused on the interpretation promulgated during the more radical 1998-2004 period, which Worrall calls the "*hard-line*" position [187], and Brody, Miller & Bogdan-Lovis call "*crude*" EBM [123].

The framework allows systematic identification of those interpretations already undermined by particular criticisms. The majority of philosophical criticism of hierarchies only applies to hierarchies interpreted in a particular way—most notably to those hard-line interpretations which offer hierarchies which include all evidence in their scope, and make absolute claims about evidence quality from individual studies, without conditions or modifiers. Chapter 4 showed that the patchwork of well-criticised interpretations of hierarchies does not cover the whole space of interpretations. Relative rankings, modest interpretations like La Caze's [11], heuristic interpretations and more sophisticated hierarchies of evidence bases like GRADE, generally survive criticism (subject to some enhancements).

However, in Chapters 5 to 7, these remaining interpretations were criticised. Chapter 5 demonstrated that relative rankings are either so strong as to provide obviously false information, or are too weak to be helpful. Modest interpretations such as La Caze's are too modest to be functional. As such, relative rankings are a poor basis for evidence appraisal in clinical practice—they do not give clinicians reliable and useful information about the properties of evidence or evidence bases. Therefore, they fail to promote EBM's goals of developing a useful system for evidence appraisal.

In Chapter 6, most of the remaining interpretative approaches were criticised. The chapter presented arguments against hierarchies of evidence, hierarchies of evidence bases, and sophisticated versions such as GRADE, in turn. First, it was argued that hierarchies which appraise the evidence from individual studies in isolation are not defensible. It is a fundamental assumption of such hierarchies that the properties of evidence (quality, strength, etc.) can be assessed by looking at a study in isolation. But Section 6.1 showed that the strength and quality of evidence for some hypothesis from a study depends on information from both within and outside the study. Internal validity, external validity and clinical usefulness can each depend on other information. To appraise the evidence for a claim, clinicians and guideline developers will need to look at multiple studies in combination to understand the way evidence interacts within an evidence base. Atomistic hierarchies provide a poor basis for evidence appraisal in clinical practice because they cannot take such meta-evidence and inter-study interaction into account.

Hierarchies of evidence bases might solve the problems of atomistic appraisal. But hierarchies of evidence bases are also problematic. Simple 'levels of evidence' fail to consider the importance of evidence *against* a causal claim, and the possibility that strong, high-quality negative evidence may be different to strong, high-quality evidence for a hypothesis. For instance, Section 6.2 draws upon the example of the ‘paradox’ of evidence-based alternative medicine to argue that evidence of biological implausibility is strong evidence against a causal hypothesis, even if evidence of a plausible biological mechanism is not strong evidence for a causal hypothesis. This example also illustrates the importance of negative evidence, and the potential for it to be overlooked and underappreciated. Therefore, levels of evidence are not a good approach to the appraisal of evidence in clinical practice. They lack the machinery to assess relevant properties of evidence bases such as replications, variation within the evidence base, meta-evidence, and evidence of biological implausibility. They have no capacity to consider the asymmetry of positive and negative evidence.

A more sophisticated position could potentially include these considerations. GRADE provides the most sophisticated attempt to appraise evidence bases through a primarily hierarchical procedure. GRADE’s up- and down-grading machinery allows it—at least in principle—to consider different factors as enhancing an evidence base to those which detract from one. The up- and down-grading criteria already include some elements of meta-evidence, such as evidence of risk of bias, evidence of publication bias, and inconsistency in study results. GRADE could potentially be refined to accommodate many of the criticisms brought against atomistic hierarchies and levels of evidence tables.

However, GRADE, like other hierarchies, focuses exclusively on evidence for estimates of the average treatment effect of an intervention on an outcome in a population. Hierarchies assess the evidence that studies provide for estimates of differential average treatment effects. But as a number of case studies demonstrated, information about the distribution of effects and the causes and predictors of effect heterogeneity is important for clinicians and patients. This information has been omitted or underappreciated in EBM’s appraisal procedures. GRADE’s few concessions to the importance of variation view information about variation as undermining the quality of an evidence base and the strength of recommendations.

But information about variation has several roles. While it may undermine a recommendation, it could also reinforce some recommendations, render a straightforward recommendation inappropriate, or justify stratifying recommendations according to patients’ attributes. As the relevance of the distribution of effects and causes of heterogeneous responses shows, information about average treatment effects alone is an insufficient basis to support individualised predictions and recommendations. Suitably sophisticated hierarchies might be able to appraise the level of evidence for an estimate of an average treatment effect, but this does not facilitate an appraisal of the evidence base for a prediction or recommendation. Even GRADE is a poor method for appraising the evidence for clinical predictions. It omits important information, and draws overly hasty connections between average effect, expected effects, and recommendations.

The issues of variation and heterogeneity cannot be adequately accommodated by formulating a hierarchy to assess evidence of variation. Therefore, while hierarchies of evidence could have a minor role to play in EBM's appraisal procedures in assessing evidence for estimates of an average treatment effect, a hierarchical procedure alone should not be used to assess evidence bases for clinically-relevant claims. The same argument also affects the use of hierarchies to appraise claims in public health and health policy.

Finally, despite these theoretical arguments which show that hierarchical approaches omit important information, it might be argued that the use of hierarchies is justified because they succeed in achieving EBM's goals in practice. Such a heuristic justification would require empirical evidence that hierarchies are user-friendly, and that they succeed in improving patient care. The evidence for these claims is currently very weak, as this chapter showed. Moreover, there are plausible alternative appraisal heuristics which could rival hierarchical approaches. But, as the approach of many EBM proponents is now focused on practitioners applying evidence-based guidelines rather than appraising evidence bases for themselves, the need for an appraisal heuristic within an EBM approach is questionable. There is an important role for future empirical research into the success and appropriateness of different approaches to evidence appraisal and integration, especially in comparing clinicians' appraisal of evidence to guideline developers'.

A more useful model for clinical practice may be that of effect distribution modelling. Evidence is marshalled to support a predictive model of likely effects. Studies which measure central tendency such as RCTs have a role within such research programmes, but it makes little sense to *rank* the contributions of different methodologies. However, the argument of this thesis does not endorse a particular alternative model of evidence appraisal for predictions and recommendations. Alternatives sketched here should be subjected to further philosophical scrutiny. This is one of the areas showing promise for further research. Philosophical work in modelling could be applied to these programmes, and research into machine learning may also be important in the development of new predictive tools.

This thesis has only focused on hierarchies which appraise evidence for claims about treatments in clinical practice and policy. There are other hierarchies used in medicine. Hierarchies have been formulated to appraise evidence for claims about diagnostic tests, differential diagnoses, harms, screening, cost-effectiveness, and decision analyses (e.g. [33,90,103,104,340]). Daly et al. [189] produced a hierarchy of qualitative evidence. Philosophical analysis may be able to contribute to the assessment of these tools, and there is potential for further work in this area. The framework outlined in Chapter 2 should provide a useful basis for this work.

There are also applications of hierarchies beyond clinical medicine. Hierarchies continue to be produced in epidemiology, as well as in healthcare fields such as dentistry (e.g. [208]) and nursing (e.g.

[192,290]). Although it is expected that the arguments in this thesis will generalise to these settings, they are beyond the scope of the claims made here. Similarly, evidence-based policy more broadly (see e.g. [163,548]) falls outside of the purview of this thesis, although some of the arguments made here may be applicable.

Bibliography

1. Higgins, J.P.T. & Green, S. (Eds.). (2011) *Cochrane handbook for systematic reviews of interventions* (5.1.0 ed.). available at: http://handbook.cochrane.org, accessed 01/04/15: The Cochrane Collaboration.
2. National Institute for Health and Care Excellence (NICE). (2004) “Reviewing and grading the evidence” *NICE: Guideline Development Methods* (Vol. 7). London: National Institute for Health and Care Excellence.
3. National Institute for Health and Care Excellence (NICE). (2009) *The guidelines manual*. London: National Institute for Health and Care Excellence.
4. World Health Organisation (WHO). (2012) *WHO Handbook for Guideline Development*. Geneva, Switzerland: WHO Press.
5. U.S. Preventive Services Task Force. (2008) *U.S. Preventive Services Procedure Manual*. AHRQ Publication No. 08-05118-EF.
6. Australian National Health and Medical Research Council (ANHMRC). (1999) *A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines*. Commonwealth of Australia: available at: http://www.health.gov.au/nhmrc/publicat/synopses/cp30syn.html, accessed 01/04/15.
7. Australian National Health and Medical Research Council (ANHMRC). (2011) *Procedures and requirements for meeting the 2011 NHMRC standard*. Melbourne: National Health and Medical Research Council.
8. Guyatt, G.H., *et al.* (2002) *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: AMA Press.
9. McAlister, F.A., *et al.* (2000) “Users' guides to the medical literature: XX. Integrating research evidence with the care of the individual patient. Evidence-Based Medicine Working Group”. *JAMA,* 283(21), 2829-2836.
10. Straus, S.E. & Sackett, D.L. (Eds.). (2005) *Evidence-Based Medicine: How to Practice and Teach EBM* (3rd ed.). Edinburgh: Elsevier Churchill Livingstone.
11. La Caze, A. (2009) “Evidence-Based Medicine Must Be...”. *Journal of Medicine and Philosophy,* 34, 509-527.
12. Atkins, D., *et al.* (2004) “Grading quality of evidence and strength of recommendations”. *BMJ,* 328(7454), 1490.
13. Brozek, J.L., *et al.* (2009) “Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions”. *Allergy,* 64(5), 669-677.
14. Balshem, H., *et al.* (2011) “GRADE guidelines: 3. Rating the quality of evidence”. *J Clin Epidemiol,* 64(4), 401-406.
15. Howick, J., *et al.* (2011) “The Oxford 2011 Levels of Evidence”. *Oxford Centre for Evidence-Based Medicine, available at www.cebm.net*, accessed 01/04/15.
16. Howick, J., *et al.* (2011) “Explanation of the 2011 Oxford Centre for Evidence-Based Medicine (OCEBM) levels of evidence (Background Document)”. *OCEBM, available at http://www.cebm.net/index.aspx?o=5653*, accessed 01/04/15.
17. Lewis, M. (2003) *Moneyball: the art of winning an unfair game*. New York: W.W. Norton & Co.
18. Christie, A. (1920) *The Mysterious Affair at Styles*. London: The Bodley Head.
19. Guyatt, G.H., Sackett, D.L. & Cook, D.J. (1993) “Users' Guides to the Medical Literature: II. How to Use an Article About Therapy or Prevention: A. Are the Results of the Study Valid?”. *Journal of the American Medical Association,* 270(21), 2598-2601.
20. Sackett, D.L. & Rosenberg, W.M. (1995) “On the need for evidence-based medicine”. *J Public Health Med,* 17(3), 330-334.
21. Davidoff, F., *et al.* (1995) “Evidence based medicine”. *BMJ,* 310(6987), 1085-1086.

22. Sackett, D.L., *et al.* (1997) *Evidence-based medicine: how to practice and teach EBM*. New York ; Edinburgh: Churchill Livingstone.
23. Briss, P.A., *et al.* (2000) "Developing an evidence-based Guide to Community Preventive Services--methods. The Task Force on Community Preventive Services". *Am J Prev Med,* 18(1 Suppl), 35-43.
24. Harris, R.P., *et al.* (2001) "Current methods of the US Preventive Services Task Force: a review of the process". *Am J Prev Med,* 20(3 Suppl), 21-35.
25. Kohatsu, N.D., Robinson, J.G. & Torner, J.C. (2004) "Evidence-based public health: an evolving concept". *American Journal of Preventive Medicine,* 27(5), 417-421.
26. National Institute for Health and Care Excellence (NICE). (2014) *Developing NICE guidelines: the manual [PMG20]*. London, RCP: available at http://www.nice.org.uk/article/PMG20, accessed 08/07/15.
27. National Institute for Health and Care Excellence (NICE). (2011) *CG117: Tuberculosis: clinical diagnosis and management of tuberculosis, and measures for its prevention and control*. London: RCP.
28. National Institute for Health and Care Excellence (NICE). (2006) *Hypertension: Management of Hypertension in Adults in Primary Care, NICE Clinical Guideline 34*. London: National Institute for Health and Care Excellence.
29. Weightman, A., *et al.* (2005) *Grading evidence and recommendations for public health interventions: developing and piloting a framework*: Health Development Agency London, UK.
30. US Department of Health and Human Services: Public Health Service: Agency for Health Care Policy and Research. (1992) "Acute pain management: operative or medical procedures and trauma, Part 1. Agency for Health Care Policy and Research". *Clin Pharm,* 11(4), 309-331.
31. Canadian Task Force on the Periodic Health Examination. (1979) "The Periodic Health Examination". *CMAJ,* 121, 1193-1252.
32. Canadian Task Force on Preventive Health Care. (2014) *Procedure Manual*. Available at: http://canadiantaskforce.ca/methods/, accessed 13/2/15.
33. Merlin, T., Weston, A. & Tooher, R. (2009) "Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'". *BMC Med Res Methodol,* 9, 34.
34. Institute for Clinical Systems Improvement. (2003) *Evidence Grading System*. available at: http://www.icsi.org, accessed 04/05/13.
35. Guyatt, G., *et al.* (2006) "Grading strength of recommendations and quality of evidence in clinical guidelines: report from an american college of chest physicians task force". *Chest,* 129(1), 174-181.
36. Gronseth, G.S., Woodroffe, L.M & Getchius, T.S. (2011) *Clinical Practice Guideline Process Manual*. AAN: available at: http://tools.aan.com/globals/axon/assets/9023.pdf, accessed 01/04/15.
37. Schunemann, H.J., *et al.* (2006) "An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations". *Am J Respir Crit Care Med,* 174(5), 605-614.
38. GRADE Working Group. (2014) *Organizations that have endorsed or that are using GRADE*. Available at: http://www.gradeworkinggroup.org/society/index.htm: accessed 12/05/15.
39. EBM Working Group. (1992) "Evidence-Based Medicine: a new approach to teaching the practice of medicine". *JAMA,* 268(17), 2420-2425.
40. Rosenberg, W. & Donald, A. (1995) "Evidence Based Medicine: an approach to clinical problem-solving". *British Medical Journal,* 310, 1122.
41. Sackett, D.L. & Straus, S.E. (1998) "Finding and applying evidence during clinical rounds: the "evidence cart"". *JAMA,* 280(15), 1336-1338.
42. Wagoner, B., *et al.* (2004) "Guide to Research Methods: The Evidence Pyramid" *SUNY Downstate Medical Center EBM Tutorial*. Available at: http://library.downstate.edu/EBM2/2100.htm, accessed 1/06/14.
43. Greenhalgh, T. (1997) *How to read a paper: the basics of evidence based medicine*. London: BMJ Pub. Group.

44. Eddy, D.M. (2005) "Evidence-based medicine: a unified approach". *Health Aff (Millwood),* 24(1), 9-17.
45. Atkins, D.*, et al.* (2004) "Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group". *BMC Health Serv Res,* 4(1), 38.
46. Petrisor, B.A., Keating, J. & Schemitsch, E. (2006) "Grading the evidence: levels of evidence and grades of recommendation". *Injury,* 37(4), 321-327.
47. Jaeschke, R.*, et al.* (2008) "Use of GRADE grid to reach decisions on clinical practice guidelines when consensus is elusive". *BMJ,* 337, a744.
48. Guyatt, G.H.*, et al.* (2008) "GRADE: an emerging consensus on rating quality of evidence and strength of recommendations". *BMJ,* 336(7650), 924-926.
49. Gillis, A.M. & Skanes, A.C. (2011) "Canadian Cardiovascular Society atrial fibrillation guidelines 2010: implementing GRADE and achieving consensus". *Can J Cardiol,* 27(1), 27-30.
50. Andrews, J.C.*, et al.* (2013) "GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations". *Journal of Clinical Epidemiology,* 66(7), 719-725.
51. Worrall, J. (2002) "What Evidence in Evidence-Based Medicine?". *Philosophy of Science,* 69(3), S316-330.
52. Worrall, J. (2007) "Why there's no cause to randomize". *British Journal of Philosophy of Science,* 58(3), 451.
53. Upshur, R.E. (2002) "If not evidence, then what? Or does medicine really need a base?". *J Eval Clin Pract,* 8(2), 113-119.
54. Upshur, R.E. (2009) "Making the grade: assuring trustworthiness in evidence". *Perspect Biol Med,* 52(2), 264-275.
55. Bluhm, R. (2005) "From hierarchy to network: a richer view of evidence for evidence-based medicine". *Perspect Biol Med,* 48(4), 535-547.
56. Bluhm, R. (2010) "Evidence-based medicine and philosophy of science". *J Eval Clin Pract,* 16(2), 363-364.
57. Goldenberg, M.J. (2006) "On evidence and evidence-based medicine: lessons from the philosophy of science". *Soc Sci Med,* 62(11), 2621-2632.
58. Howick, J., Glasziou, P. & Aronson, J.K. (2010) "Evidence-based mechanistic reasoning". *J R Soc Med,* 103(11), 433-441.
59. Howick, J. (2011) *The philosophy of evidence-based medicine*. Oxford: Wiley-Blackwell.
60. Russo, F. & Williamson, J. (2007) "Interpreting causality in the health sciences". *International studies in the philosophy of science,* 21(2), 157-170.
61. Illari, P.M. (2011) "Mechanistic evidence: disambiguating the Russo–Williamson thesis". *International studies in the philosophy of science,* 25(2), 139-157.
62. Clarke, B.*, et al.* (2013) "The evidence that evidence-based medicine omits". *Prev Med,* 57(6), 745-747.
63. Clarke, B.*, et al.* (2013) "Mechanisms and the evidence hierarchy". *Topoi*, 1-22.
64. Guyatt, G.H. (1991) "Evidence-based medicine". *ACP Journal Club: Supplement 2 to Annals of Internal Medicine,* 114, A-16.
65. Cochrane Collaboration. (2011) *Evidence-based healthcare and systematic reviews*. Available at: http://www.cochrane.org/about-us/evidence-based-health-care, accessed 12/04/15.
66. Sackett, D.L. (1997) "Evidence-based medicine". *Semin Perinatol,* 21(1), 3-5.
67. Marchevsky, A.M. (2005) "Evidence-based medicine in pathology: an introduction". *Semin Diagn Pathol,* 22(2), 105-115.
68. Zarkovich, E. & Upshur, R.E. (2002) "The virtues of evidence". *Theor Med Bioeth,* 23(4-5), 403-412.
69. Bluhm, R. (2010) "The epistemology and ethics of chronic disease research: further lessons from ECMO". *Theor Med Bioeth,* 31(2), 107-122.
70. Sehon, S.R. & Stanley, D.E. (2003) "A philosophical analysis of the evidence-based medicine debate". *BMC Health Serv Res,* 3(1), 14.

71. Timmermans, S. & Mauck, A. (2005) "The promises and pitfalls of evidence-based medicine". *Health Aff (Millwood),* 24(1), 18-28.

72. La Caze, A. (2008) "Evidence based medicine can't be...". *Social Epistemology,* 22(4), 353-370.

73. Daly, J. (2005) *Evidence-based medicine and the search for a science of clinical care*. Berkeley, CA ; London: University of California Press.

74. Bluhm, R. (2009) "Some observations on" observational" research". *Perspectives in Biology and Medicine,* 52(2), 252-263.

75. Worrall, J. (2007) "Evidence in Medicine and Evidence-Based Medicine". *Philosophy Compass,* 2/6, 981-1022.

76. Worrall, J. (2010) "Evidence: philosophy of science meets medicine". *J Eval Clin Pract,* 16(2), 356-362.

77. Lacchetti, C. & Guyatt, G. (2002) "Surprising results of randomized controlled trials", In Guyatt, G.H. & Rennie, D. (Eds.), *Users' guide to the medical literature* (pp. 247-265). Chicago, IL: AMA Press.

78. Sackett, D.L., *et al.* (1996) "Evidence based medicine: what it is and what it isn't". *BMJ,* 312(7023), 71-72.

79. Masterman, M. (1970) "The nature of a paradigm", In Lakatos, I. & Musgrave, A. (Eds.), *Criticism and the Growth of Knowledge* (pp. 59-90). Cambridge, UK: Cambridge U.P.

80. Kuhn, T.S. (1962) *The structure of scientific revolutions*. Chicago: University of Chicago Press.

81. Norman, G.R. (1999) "Examining the assumptions of evidence-based medicine". *J Eval Clin Pract,* 5(2), 139-147.

82. Guyatt, G.H., *et al.* (2000) "Users' Guides to the Medical Literature: XXV. Evidence-based medicine: principles for applying the Users' Guides to patient care. Evidence-Based Medicine Working Group". *JAMA,* 284(10), 1290-1296.

83. Guyatt, G.H. & Rennie, D. (2008) "The philosophy of evidence-based medicine", In Guyatt, G.H. & Rennie, D. (Eds.), *Users' Guides to the Medical Literature* (pp. 9-16). New York: McGraw Hill Medical.

84. Haynes, R.B. (2002) "What kind of evidence is it that Evidence-Based Medicine advocates want health care providers and consumers to pay attention to?". *BMC Health Services Research,* 2, 3.

85. Montori, V.M. & Guyatt, G.H. (2008) "Progress in evidence-based medicine". *JAMA,* 300(15), 1814-1816.

86. Djulbegovic, B., Guyatt, G.H. & Ashcroft, R.E. (2009) "Epistemologic inquiries in evidence-based medicine". *Cancer Control,* 16(2), 158-168.

87. Solomon, M. (2009) *Just a Paradigm: Evidence-Based Medicine Meets Philosophy of Science*: expansion of a paper presented at SPSP 2009, available at: http://philpapers.org/rec/SOLJAP, accessed 09/03/13.

88. Shahar, E. (1998) "Evidence-based medicine: a new paradigm or the Emperor's new clothes?". *J Eval Clin Pract,* 4(4), 277-282.

89. Couto, J.S. (1998) "Evidence-based medicine: a Kuhnian perspective of a transvestite non-theory". *J Eval Clin Pract,* 4(4), 267-275.

90. Bhandari, M. & Giannoudis, P.V. (2006) "Evidence-based medicine: what it is and what it is not". *Injury,* 37(4), 302-306.

91. Bhandari, M. & Tornetta, P. (2003) "Evidence-based orthopaedics: a paradigm shift". *Clin Orthop Relat Res*(413), 9-10.

92. Bhandari, M., Zlowodzki, M. & Cole, P.A. (2004) "From eminence-based practice to evidence-based practice: a paradigm shift". *Minn Med,* 87(4), 51-54.

93. Sackett, D.L. (1991) *Clinical epidemiology: a basic science for clinical medicine* (2nd ed.). Boston ; London: Little, Brown.

94. Fletcher, R.H., Fletcher, S.W. & Wagner, E.H. (1982) *Clinical epidemiology : the essentials*. Baltimore ; London: Williams & Wilkins.

95. Sackett, D.L., Haynes, R.B. & Tugwell, P. (1985) *Clinical epidemiology : a basic science for clinical medicine*. Boston: Little, Brown.

96. Sackett, D.L. (2002) "Clinical epidemiology: what, who, and whither". *J Clin Epidemiol,* 55(12), 1161-1166.

97. Rawlins, M. (2008) "De Testimonio: On the Evidence for Decisions about the use of Therapeutic Interventions". *Lancet,* 372, 2152-2161.
98. Sauve, S., *et al.* (1995) "The Critically Appraised Topic: a practical approach to learning critical appraisal". *Annales CRMCC,* 28(7), 396-398.
99. Quin, G. (2001) "Best BETs". *Evid Based Med,* 6(3), 70.
100. Haynes, B. (1991) "The origins and aspirations of the ACP Journal Club". *ACP J Club,* A18.
101. DiCenso, A., Bayley, L. & Haynes, R.B. (2009) "ACP Journal Club. Editorial: Accessing preappraised evidence: fine-tuning the 5S model into a 6S model". *Ann Intern Med,* 151(6), JC3-2, JC3-3.
102. Anonymous. (1995) "Evidence-Based Everything". *Bandolier,* 12, 91-98.
103. Howick, J., *et al.* (2009) "CEBM Levels of Evidence". *Centre for Evidence-Based Medicine, www.cebm.net*, accessed 01/04/15.
104. Ebell, M.H., *et al.* (2004) "Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature". *Am Fam Physician,* 69(3), 548-556.
105. Edwards, A.G., Russell, I.T. & Stott, N.C. (1998) "Signal versus noise in the evidence base for medicine: an alternative to hierarchies of evidence?". *Fam Pract,* 15(4), 319-322.
106. Barratt, H. (2009) *The hierarchy of research evidence – from well-conducted meta-analysis down to small case series, publication bias.* Public Health Action Support Team (PHAST), Department of Health. Available at: http://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/hierarchy-research-evidence: accessed 15/07/15.
107. Cochrane Collaboration. (2015) *Glossary*. Available at: http://community.cochrane.org/glossary/, accessed 12/07/15.
108. Cook, D.J., Sackett, D.L. & Spitzer, W.O. (1995) "Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis". *J Clin Epidemiol,* 48(1), 167-171.
109. Moher, D., *et al.* (1999) "Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement". *Lancet,* 354(9193), 1896-1900.
110. Haidich, A. (2010) "Meta-analysis in medical research". *Hippokratia,* 14(Suppl 1), 29.
111. Sackett, D.L. (2000) "The sins of expertness and a proposal for redemption". *BMJ,* 320(7244), 1283.
112. Haynes, R.B., Devereaux, P.J. & Guyatt, G.H. (2002) "Clinical expertise in the era of evidence-based medicine and patient choice". *ACP J Club,* 136(2), A11-14.
113. Chambers, D.W. (2010) "Evidence-based dentistry". *J Am Coll Dent,* 77(4), 68-80.
114. Antman, E.M., *et al.* (1992) "A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction". *JAMA,* 268(2), 240-248.
115. Hayward, R.S., *et al.* (1995) "Users' Guides to the Medical Literature: VIII. How to Use Clinical Practice Guidelines A. Are the Recommendations Valid?". *JAMA,* 274(7), 570-574.
116. Bird, A. (2011) "What can philosophy tell us about Evidence-Based Medicine? An assessment of Jeremy Howick'sThe Philosophy of Evidence-based Medicine". *International Journal of Person Centered Medicine,* 1(4), 642-648.
117. Broadbent, A. (2013) *Philosophy of epidemiology*: Palgrave Macmillan.
118. Marchevsky, A.M. & Wick, M.R. (2010) "Evidence levels for publications in pathology and laboratory medicine". *Am J Clin Pathol,* 133(3), 366-367.
119. Jenicek, M. (2002) *Foundations of evidence-based medicine*. Boca Raton: Parthenon.
120. Abraham, J. (1994) "Bias in Science and Medical Knowledge: The Opren Controversy". *Sociology,* 28(3), 717-736.
121. Concato, J. (2012) "When to randomize, or 'Evidence-based medicine needs medicine-based evidence'". *Pharmacoepidemiol Drug Saf,* 21 Suppl 2, 6-12.
122. Sackett, D.L. (2000) "The fall of "clinical research" and the rise of "clinical-practice research"". *Clin Invest Med,* 23(6), 379-381.
123. Brody, H., Miller, F.G. & Bogdan-Lovis, E. (2005) "Evidence-based medicine: watching out for its friends". *Perspect Biol Med,* 48(4), 570-584.

124. La Caze, A. (2011) “The role of basic science in evidence-based medicine”. *Biol Philos,* 26, 81-98.

125. Forsblad-d’Elia, H., *et al.* (2014) “Drug adherence, response and predictors thereof for tocilizumab in patients with rheumatoid arthritis: results from the Swedish biologics register”. *Rheumatology,* 54(7), 1186-1193.

126. Maini, R., *et al.* (2006) “Double-blind randomized controlled clinical trial of the interleukin-6 receptor antagonist, tocilizumab, in European patients with rheumatoid arthritis who had an incomplete response to methotrexate”. *Arthritis & Rheumatism,* 54(9), 2817-2829.

127. Vickrey, B.G. (1999) “Getting oriented to patient-oriented outcomes”. *Neurology,* 53(4), 662-662.

128. Ebell, M.H., *et al.* (1999) “Finding POEMs in the medical literature”. *J Fam Pract,* 48(5), 350-355.

129. Guyatt, G.H., *et al.* (2011) “GRADE guidelines: 8. Rating the quality of evidence--indirectness”. *J Clin Epidemiol,* 64(12), 1303-1310.

130. Phillips, R., *et al.* (1998) “Levels of Evidence”. *Centre for Evidence-Based Medicine,* www.cebm.net.

131. Ball, C.M. & Phillips, R. (2002). “Levels of Evidence - March 2002”, *Evidence Based On Call*. Retrieved from http://www.eboncall.org/content/levels.html

132. Eaves, F.F. (2011) “Got evidence? Stem cells, bias, and the level of evidence ladder: commentary on: "ASAPS/ASPS position statement on stem cells and fat grafting"”. *Aesthet Surg J,* 31(6), 718-722.

133. Harbour, R. & Miller, J. (2001) “A new system for grading recommendations in evidence based guidelines”. *BMJ,* 323(7308), 334-336.

134. Scottish Intercollegiate Guidelines Network (SIGN). (2008) *SIGN 50: A Guideline Developer's Handbook*. available at: www.sign.ac.uk, accessed 18/01/13.

135. Sackett, D.L., *et al.* (2000) *Evidence-Based Medicine: How to Practice and Teach EBM*. London: Churchill-Livingstone.

136. Giacomini, M.K., Cook, D.J. & Evidence-Based Medicine Working Group. (2000) “Users' guides to the medical literature: XXIII. Qualitative research in health care A. Are the results of the study valid?”. *JAMA,* 284(3), 357-362.

137. Howick, J., Glasziou, P. & Aronson, J.K. (2009) “The evolution of evidence hierarchies: what can Bradford Hill's 'guidelines for causation' contribute?”. *J R Soc Med,* 102(5), 186-194.

138. Karanicolas, P.J., Kunz, R. & Guyatt, G.H. (2008) “Point: evidence-based medicine has a sound scientific base”. *Chest,* 133(5), 1067-1071.

139. Porzsolt, F., *et al.* (2003) “Evidence-based decision making--The six step approach”. *EBM Notebook,* 8, 165-166.

140. Schardt, C. & Mayer, J. (2010) *Introduction to Evidence-Based Practice*. Creative Commons: Available at: http://www.hsl.unc.edu/Services/Tutorials/EBM/index.htm, accessed 03/02/13.

141. Sharp, D. (1996) “Evidence-Based Medicine: How to Practice and Teach EBM [Review]”. *Lancet,* 348, 1297.

142. Chesson, A.L., Jr., *et al.* (1999) “Practice parameters for the treatment of restless legs syndrome and periodic limb movement disorder. An American Academy of Sleep Medicine Report. Standards of Practice Committee of the American Academy of Sleep Medicine”. *Sleep,* 22(7), 961-968.

143. Haynes, R.B., *et al.* (1985) “Computer searching of the medical literature. An evaluation of MEDLINE searching systems”. *Ann Intern Med,* 103(5), 812-816.

144. Haynes, R.B. (1988) “Computer literature searching for busy clinicians”. *Can Fam Physician,* 34, 435-440.

145. Haynes, R.B. (2001) “Of studies, syntheses, synopses, and systems: the "4S" evolution of services for finding current best evidence”. *ACP J Club,* 134(2), A11-13.

146. Haynes, R.B. (2005) “Of studies, syntheses, synopses, summaries, and systems: the" 5S" evolution of information services for evidence-based health care decisions”. *ACP J Club,* 145(3), A8-A8.

147. Hunt, D.L., Haynes, R.B. & Browman, G.P. (1998) "Searching the medical literature for the best evidence to solve clinical questions". *Ann Oncol,* 9(4), 377-383.
148. Black, N. (1996) "Why we need observational studies to evaluate the effectiveness of health care". *British Medical Journal,* 312, 1215-1218.
149. Goldenberg, M.J. (2009) "Iconoclast or creed? Objectivism, pragmatism, and the hierarchy of evidence". *Perspect Biol Med,* 52(2), 168-187.
150. Stegenga, J. (2011) "Is meta-analysis the platinum standard of evidence?". *Studies in history and philosophy of science part C: Studies in history and philosophy of biological and biomedical sciences,* 42(4), 497-507.
151. Cook, D.J., *et al.* (1992) "Rules of evidence and clinical recommendations on the use of antithrombotic agents". *Chest,* 102(4 Suppl), 305S-311S.
152. Stegenga, J. (2015) "Herding QATs: Quality Assessment Tools for Evidence in Medicine" *Classification, Disease and Evidence* (pp. 193-211): Springer.
153. Zwarenstein, M., *et al.* (2008) "Improving the reporting of pragmatic trials: an extension of the CONSORT statement". *BMJ,* 337, a2390.
154. Schulz, K.F., Altman, D.G. & Moher, D. (2010) "CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials". *BMJ,* 340, c332.
155. Campbell, D.T. & Stanley, J.C. (1963) *Experimental and Quasi-Experimental Designs for Research*. Boston: Hougton Miffling Co.
156. Cook, T.C. & Campbell, D.T. (1976) "The design and conduct of quasi-experiments and true experiments in field settings", In Dunnete, M. (Ed.), *Handbook of Industrial and Organizational Psychology*. Skokie, IL: Rand McNally.
157. Goldenberg, M.J., Borgerson, K. & Bluhm, R. (2009) "The nature of evidence in evidence-based medicine: guest editors' introduction". *Perspect Biol Med,* 52(2), 164-167.
158. Upshur, R.E.G., Van Den Kerkhof, E.G. & Goel, V. (2001) "Meaning and measurement: an inclusive model of evidence in health care". *Journal of Evaluation in Clinical Practice,* 7(2), 91-96.
159. Upshur, R.E. & Colak, E. (2003) "Argumentation and evidence". *Theor Med Bioeth,* 24(4), 283-299.
160. Sturmberg, J.P. (2009) "EBM: a narrow and obsessive methodology that fails to meet the knowledge needs of a complex adaptive clinical world: a commentary on Djulbegovic, B., Guyatt, GH & Ashcroft, RE (2009) Cancer Control, 16, 158–168". *Journal of Evaluation in Clinical Practice,* 15(6), 917-923.
161. Kelly, T. (2014) "Evidence", In Zalta, E.N. (Ed.), *Stanford Encyclopedia of Philosophy* (Vol. Fall 2014). http://plato.stanford.edu/archives/fall2014/entries/evidence/, accessed 15/07/15.
162. Clarke, B. (2013) *What is the difference between data and evidence?* Available at: https://www.ucl.ac.uk/sts/staff/clarke/data-evidence_distinction, accessed 14/07/15.
163. Cartwright, N., *et al.* (2008) *Evidence-based policy: where is our theory of evidence?* London: Contingency and Dissent in Science Project, Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science.
164. Bogen, J. & Woodward, J. (1988) "Saving the phenomena". *Philosophical Review*, 303-352.
165. Goodman, K.W. (2003) *Ethics and evidence-based medicine : fallibility and responsibility in clinical science*. Cambridge: Cambridge University Press.
166. Conee, E. & Feldman, R.D. (2004) *Evidentialism: Essays in Epistemology: Essays in Epistemology*. Oxford: Oxford University Press.
167. Urbach, P. & Howson, C. (1993) "Scientific Reasoning: The Bayesian Approach". Peru, Illinois: Open Court Publishing.
168. Talbott, W. (2015) "Bayesian Epistemology", In Zalta, E.N. (Ed.), *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/sum2015/entries/epistemology-bayesian/, accessed 15/07/15.
169. Jeffrey, R. (1992) *Probability and the Art of Judgment*. Cambridge, UK: Cambridge University Press.
170. Earman, J. (1992) *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.

171. Popper, K.R. (1959) *The logic of scientific discovery*. London: Hutchinson.

172. Popper, K.R. (1963) *Conjectures and refutations : the growth of scientific knowledge*. London: Routledge & Kegan Paul.

173. Bird, A. (2010) "Eliminative abduction: examples from medicine". *Studies in History and Philosophy of Science Part A,* 41(4), 345-352.

174. Bird, A. (2011) "The epistemological function of Hill's criteria". *Prev Med,* 53(4), 242-245.

175. Glasziou, P. (2001) "Which methods for bedside Bayes?". *ACP J Club,* 135(3), A11-12.

176. Richardson, W.S., *et al.* (1999) "Users' guides to the medical literature: XV. How to use an article about disease probability for differential diagnosis. Evidence-Based Medicine Working Group". *JAMA,* 281(13), 1214-1219.

177. Richardson, W.S. & Wilson, M.C. (2008) "The process of diagnosis", In Guyatt, G.H. & Rennie, D. (Eds.), *Users Guides to the Medical Literature* (pp. 399-406). Chicago, IL: AMA Press.

178. Penston, J. (2005) "Large-scale randomised trials--a misguided approach to clinical research". *Med Hypotheses,* 64(3), 651-657.

179. Ziliak, S.T. & McCloskey, D.N. (2008) *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*: University of Michigan Press.

180. Sackett, D.L. (2001) "Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!)". *CMAJ,* 165(9), 1226-1237.

181. Feldman, R.D. (2009) "Evidentialism, higher-order evidence, and disagreement". *Episteme,* 6(03), 294-312.

182. Kelly, T. (2010) "Peer disagreement and higher order evidence", In Goldman, A. & Whitcomb, D. (Eds.), *Social epistemology: Essential readings* (pp. 183-217). Oxford: Oxford University Press.

183. Christensen, D. (2010) "Higher-Order Evidence". *Philosophy and Phenomenological Research,* 81(1), 185-215.

184. Lasonen-Aarnio, M. (2014) "Higher-Order Evidence and the Limits of Defeat". *Philosophy and Phenomenological Research,* 88(2), 314-345.

185. Bartlett, J.G., *et al.* (1998) "Community-acquired pneumonia in adults: guidelines for management. The Infectious Diseases Society of America". *Clin Infect Dis,* 26(4), 811-838.

186. Petticrew, M. & Roberts, H. (2003) "Evidence, hierarchies, and typologies: horses for courses". *J Epidemiol Community Health,* 57(7), 527-529.

187. Worrall, J. (2008) "Evidence and ethics in medicine". *Perspect Biol Med,* 51(3), 418-431.

188. Nordenström, J. (2007) *Evidence-based medicine in Sherlock Holmes' footsteps*. Malden, Mass.: Blackwell Pub.

189. Daly, J., *et al.* (2007) "A hierarchy of evidence for assessing qualitative health research". *J Clin Epidemiol,* 60(1), 43-49.

190. Guyatt, G.H., *et al.* (1995) "Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group". *JAMA,* 274(22), 1800-1804.

191. Blunt, C.J. (2015) *Hierarchies of Evidence Database*. Available at: http://cjblunt.com/hierarchies-evidence, accessed 24/07/15.

192. Evans, D. (2003) "Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions". *J Clin Nurs,* 12(1), 77-84.

193. Sackett, D.L. (1981) "How to read clinical journals: IV. To determine etiology or causation". *CMAJ,* 124, 985-990.

194. Khan, K. (2011) *Systematic reviews to support evidence based medicine: How to review and apply findings of healthcare research*: CRC Press.

195. Guyatt, G.H., *et al.* (2011) "GRADE guidelines: 1. Introduction--GRADE evidence profiles and summary of findings tables". *Journal of Clinical Epidemiology,* 64, 383-394.

196. Guyatt, G.H., *et al.* (2011) "GRADE guidelines: 2. Framing the question and deciding on important outcomes". *J Clin Epidemiol,* 64(4), 395-400.

197. Guyatt, G.H., *et al.* (2011) "GRADE guidelines: 7. Rating the quality of evidence--inconsistency". *J Clin Epidemiol,* 64(12), 1294-1302.

198. Guyatt, G.H., *et al.* (2011) "GRADE guidelines: 5. Rating the quality of evidence--publication bias". *J Clin Epidemiol,* 64(12), 1277-1282.
199. Guyatt, G.H., *et al.* (2008) "Grades of recommendation for antithrombotic agents: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition)". *Chest,* 133(6 Suppl), 123S-131S.
200. Goldet, G. & Howick, J. (2013) "Understanding GRADE: an introduction". *Journal of Evidence-Based Medicine,* 6(1), 50-54.
201. Guyatt, G.H., *et al.* (2011) "GRADE guidelines: 6. Rating the quality of evidence--imprecision". *J Clin Epidemiol,* 64(12), 1283-1293.
202. Guyatt, G.H., *et al.* (2011) "GRADE guidelines: 9. Rating up the quality of evidence". *J Clin Epidemiol,* 64(12), 1311-1316.
203. Guyatt, G.H., *et al.* (2011) "GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias)". *J Clin Epidemiol,* 64(4), 407-415.
204. Guyatt, G.H., *et al.* (2012) "GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes". *J Clin Epidemiol*.
205. Andrews, J.C., *et al.* (2013) "GRADE guidelines: 15. Going from evidence to recommendation—determinants of a recommendation's direction and strength". *Journal of Clinical Epidemiology,* 66(7), 726-735.
206. Alexander, P.E., *et al.* (2014) "World Health Organization recommendations are often strong based on low confidence in effect estimates". *Journal of Clinical Epidemiology,* 67(6), 629-634.
207. LaForce, F.M. (1987) "Immunizations, immunoprophylaxis, and chemoprophylaxis to prevent selected infections. US Preventive Services Task Force". *JAMA,* 257(18), 2464-2470.
208. Niederman, R. & Richards, D. (2010) "The challenge, access, risks, and deficits of evidence-based dentistry". *J Am Coll Dent,* 77(4), 20-24.
209. Khan, K.S., Dinnes, J. & Kleijnen, J. (2001) "Systematic reviews to evaluate diagnostic tests". *European Journal of Obstetrics & Gynecology and Reproductive Biology,* 95(1), 6-11.
210. Schunemann, H.J. & Bone, L. (2003) "Evidence-based orthopaedics: a primer". *Clin Orthop Relat Res,* 413, 117-132.
211. Guyatt, G.H., Sackett, D.L. & Cook, D.J. (1994) "Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group". *JAMA,* 271(1), 59-63.
212. Slack, M.K. (2001) "Establishing the internal and external validity of experimental studies". *American journal of health-system pharmacy,* 58(22), 2173-2181.
213. Rothwell, P.M. (2005) "External validity of randomised controlled trials: "To whom do the results of this trial apply?"". *Lancet,* 365, 82-93.
214. Rothwell, P.M. (2010) "Commentary: External validity of results of randomized trials: disentangling a complex concept". *Int J Epidemiol,* 39(1), 94-96.
215. Glasziou, P., *et al.* (1998) "Applying the results of trials and systematic reviews to individual patients". *ACP J Club,* 129(3), A15-16.
216. McAlister, F.A., *et al.* (1999) "Users' Guides to the Medical Literature: XIX. Applying clinical trial results B. Guidelines for determining whether a drug is exerting (more than) a class effect". *JAMA,* 282(14), 1371-1377.
217. Bucher, H.C., *et al.* (1999) "Users' guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. Evidence-Based Medicine Working Group". *JAMA,* 282(8), 771-778.
218. Sackett, D.L. (2011) "Clinician-trialist rounds: 4. why not do an N-of-1 RCT?". *Clin Trials,* 8(3), 350-352.
219. Friedman, G.D. (1994) *Primer of epidemiology* (4th ed.). New York ; London: McGraw-Hill.
220. Cook, T.D., Campbell, D.T. & Day, A. (1979) *Quasi-Experimentation: design and analysis issues for field settings*. Chicago, IL: Rand McNally.
221. Messick, S. (1989) "Validity", In Linn, R.L. (Ed.), *Educational Measurement* (pp. 13-103). Washington, DC: American Council on Education.
222. Calder, B.J., Phillips, L.W. & Tybout, A.M. (1982) "The concept of external validity". *Journal of Consumer Research,* 9, 240-244.

223. Glasziou, P., Del Mar, C. & Salisbury, J. (2003) *Evidence-based medicine workbook : finding and applying the best evidence to improve patient care*. London: BMJ.

224. Petrie, A. & Sabin, C. (2013) *Medical statistics at a glance*: John Wiley & Sons.

225. Sedgwick, P. (2013) "Selection bias versus allocation bias". *BMJ,* 346, f3345.

226. Kunz, R., Vist, G. & Oxman, A.D. (2002) "Randomisation to protect against selection bias in healthcare trials". *Cochrane Database of Methodology Reviews,* 4, available at: http://onlinelibrary.wiley.com/doi/10.1002/14651858.MR000012/full, accessed 07/09/15.

227. Borgerson, K. (2009) "Valuing evidence: bias and the evidence hierarchy of evidence-based medicine". *Perspect Biol Med,* 52(2), 218-233.

228. Worrall, J. (2010) "Do We Need Some Large, Simple Randomized Trials in Medicine?", In Suárez, M., Dorato, M. & Rédei, M. (Eds.), *EPSA Philosophical Issues in the Sciences* (pp. 289-301): Springer Netherlands.

229. Kaptchuk, T.J., *et al.* (2000) "Do medical devices have enhanced placebo effects?". *Journal of Clinical Epidemiology,* 53(8), 786-792.

230. Olshansky, B. (2007) "Placebo and Nocebo in Cardiovascular Health: Implications for Healthcare, Research, and the Doctor-Patient Relationship". *Journal of the American College of Cardiology,* 49(4), 415-421.

231. Adair, J.G. (1984) "The Hawthorne effect: A reconsideration of the methodological artifact". *Journal of applied psychology,* 69(2), 334.

232. De Amici, D., *et al.* (2000) "Impact of the Hawthorne effect in a longitudinal clinical study: the case of anesthesia". *Control Clin Trials,* 21(2), 103-114.

233. McCarney, R., *et al.* (2007) "The Hawthorne Effect: a randomised, controlled trial". *BMC Med Res Methodol,* 7(1), 30.

234. Braunholtz, D.A., Edwards, S.J.L. & Lilford, R.J. (2001) "Are randomized clinical trials good for us (in the short term)? Evidence for a "trial effect"". *Journal of Clinical Epidemiology,* 54(3), 217-224.

235. Rosenthal, R. & Jacobson, L. (1968) "Pygmalion in the classroom". *The Urban Review,* 3(1), 16-20.

236. Rosenthal, R. (1966) *Experimenter effects in behavioral research*. New York: Appleton Century Crofts.

237. Sackett, D.L. (1979) "Bias in analytic research". *J Chronic Dis,* 32(1), 51-63.

238. Mahoney, M.J. (1977) "Publication prejudices: An experimental study of confirmatory bias in the peer review system". *Cognitive therapy and research,* 1(2), 161-175.

239. Easterbrook, P.J., *et al.* (1991) "Publication bias in clinical research". *Lancet,* 337(8746), 867-872.

240. Stern, J.M. & Simes, R.J. (1997) "Publication bias: evidence of delayed publication in a cohort study of clinical research projects". *BMJ: British Medical Journal,* 315(7109), 640.

241. Dwan, K., *et al.* (2008) "Systematic review of the empirical evidence of study publication bias and outcome reporting bias". *PLoS One,* 3(8), e3081.

242. Hopewell, S., *et al.* (2009) "Publication bias in clinical trials due to statistical significance or direction of trial results". *Cochrane Database Syst Rev,* 1(1).

243. Every-Palmer, S. & Howick, J. (2014) "How evidence-based medicine is failing due to biased trials and selective publication". *Journal of Evaluation in Clinical Practice,* 20(6), 908-914.

244. Gøtzsche, P.C. (1989) "Multiple publication of reports of drug trials". *European journal of clinical pharmacology,* 36(5), 429-432.

245. Tramèr, M.R., *et al.* (1997) "Impact of covert duplicate publication on meta-analysis: a case study". *BMJ,* 315(7109), 635-640.

246. Newell, D.J. (1992) "Intention-to-treat analysis: implications for quantitative and qualitative research". *Int J Epidemiol,* 21(5), 837-841.

247. Lexchin, J. & Light, D.W. (2006) "Commercial bias in medical journals: Commercial influence and the content of medical journals". *BMJ: British Medical Journal,* 332(7555), 1444.

248. Rising, K., Bacchetti, P. & Bero, L. (2008) "Reporting bias in drug trials submitted to the Food and Drug Administration: review of publication and presentation". *PLoS Med,* 5(11), e217.

249. Howick, J. (2008). "Double-Blinding: the Benefits and Risks of being in the dark". In Fennell, D. (Ed.), *Contingency and Dissent in Science*. Centre for the Philosophy of Natural and Social Science: London School of Economics.
250. Sterne, J.A. & Egger, M. (2001) "Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis". *Journal of Clinical Epidemiology,* 54(10), 1046-1055.
251. Egger, M., *et al.* (1997) "Bias in meta-analysis detected by a simple, graphical test". *BMJ,* 315(7109), 629-634.
252. Collins, H. & Pinch, T. (2008) *Dr. Golem: how to think about medicine*. Chicago: University of Chicago Press.
253. Collins, H.M. & Pinch, T. (1998) *The golem: What you should know about science*: Cambridge University Press.
254. Victora, C.G., Habicht, J. & Bryce, J. (2004) "Evidence-Based Public Health: Moving Beyond Randomized Trials". *American Journal of Public Health,* 94, 400-405.
255. Godwin, M., *et al.* (2003) "Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity". *BMC Med Res Methodol,* 3(1), 28.
256. Ho, P.M., Peterson, P.N. & Masoudi, F.A. (2008) "Evaluating the evidence: is there a rigid hierarchy?". *Circulation,* 118(16), 1675-1684.
257. Cartwright, N. & Munro, E. (2010) "The limitations of randomized controlled trials in predicting effectiveness". *J Eval Clin Pract,* 16(2), 260-266.
258. Cartwright, N. (2011) "A philosopher's view of the long road from RCTs to effectiveness". *Lancet,* 377(9775), 1400-1401.
259. Cartwright, N. (2011) *Predicting "it will work for us":(way) beyond statistics*: Oxford University Press.
260. Sokka, T. & Pincus, T. (2003) "Most patients receiving routine care for rheumatoid arthritis in 2001 did not meet inclusion criteria for most recent clinical trials or american college of rheumatology criteria for remission". *J Rheumatol,* 30(6), 1138-1146.
261. Rothwell, P.M. (1995) "Can overall results of clinical trials be applied to all patients?". *Lancet,* 345, 1616-1619.
262. Schwartz, D. & Lellouch, J. (1967) "Explanatory and pragmatic attitudes in therapeutical trials". *J Chronic Dis,* 20(8), 637-648.
263. MacRae, K.D. (1989) "Pragmatic versus explanatory trials". *Int J Technol Assess Health Care,* 5(3), 333-339.
264. Thorpe, K.E., *et al.* (2009) "A pragmatic–explanatory continuum indicator summary (PRECIS): a tool to help trial designers". *Journal of Clinical Epidemiology,* 62(5), 464-475.
265. Jonas, W.B. (2001) "The evidence house: how to build an inclusive base for complementary medicine". *Western Journal of Medicine,* 175(2), 79.
266. Dans, A.L., *et al.* (1998) "Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. Evidence-Based Medicine Working Group". *JAMA,* 279(7), 545-549.
267. Felson, D.T., *et al.* (1995) "American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis". *Arthritis & Rheumatism,* 38(6), 727-735.
268. InfoPOEM Inc. (2004) *Improving access to knowledge: an overview of InfoPOEMs*. available at: https://www.infopoems.com/resources/InfoPOEMsOverview_PPT.pdf, accessed 04/08/15.
269. Bartlett, R.H., *et al.* (1985) "Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study". *Pediatrics,* 76(4), 479-487.
270. O'Rourke, P.P., *et al.* (1989) "Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistent pulmonary hypertension of the newborn: a prospective randomized study". *Pediatrics,* 84(6), 957-963.
271. UK Collaborative ECMO Trial Group. (1996) "UK collaborative randomised trial of neonatal extracorporeal membrane oxygenation". *Lancet,* 348(9020), 75-82.
272. Truog, R.D. (1992) "Randomized controlled trials: lessons from ECMO". *Clin Res,* 40(3), 519-527.
273. Royall, R.M. (1991) "Ethics and statistics in randomized clinical trials". *Statistical Science,* 6(1), 52-62.

274. Fitelson, B. (1999) “The plurality of Bayesian measures of confirmation and the problem of measure sensitivity”. *Philosophy of Science*, S362-S378.
275. Eells, E. & Fitelson, B. (2000) “Measuring confirmation and evidence”. *Journal of Philosophy,* 97(12), 663-672.
276. Eells, E. & Fitelson, B. (2002) “Symmetries and asymmetries in evidential support”. *Philosophical Studies,* 107(2), 129-142.
277. Worrall, J. (2011) “Causality in medicine: getting back to the Hill top”. *Prev Med,* 53(4-5), 235-238.
278. Petrisor, B. & Bhandari, M. (2007) “The hierarchy of evidence: Levels and grades of recommendation”. *Indian J Orthop,* 41(1), 11-15.
279. Fuster, V., *et al.* (2006) “ACC/AHA/ESC 2006 guidelines for the management of patients with atrial fibrillation: full text A report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the European Society of Cardiology Committee for Practice Guidelines (Writing Committee to Revise the 2001 Guidelines for the Management of Patients With Atrial Fibrillation) Developed in collaboration with the European Heart Rhythm Association and the Heart Rhythm Society”. *Europace,* 8(9), 651-745.
280. Anonymous. (2004) “Hierarchy of Evidence and Grading of Recommendations”. *Thorax,* 59, 13-14.
281. Schunemann, H.J., Fretheim, A. & Oxman, A.D. (2006) “Improving the use of research evidence in guideline development: 9. Grading evidence and recommendations”. *Health Res Policy Syst,* 4, 21.
282. Ogilvie, R.I., *et al.* (1993) “Report of the Canadian Hypertension Society Consensus Conference: 3. Pharmacologic treatment of essential hypertension”. *CMAJ,* 149(5), 575-584.
283. Shekelle, P.G., *et al.* (1999) “Clinical guidelines: developing guidelines”. *BMJ,* 318(7183), 593-596.
284. Mason, J. & Eccles, M. (2003) *Guideline Recommendation and Evidence Grading (GREG)*. Newcastle, UK: University of Newcastle Centre for Health Services Research.
285. Guyatt, G.H., *et al.* (1998) “Grades of recommendation for antithrombotic agents”. *Chest,* 114(5 Suppl), 441S-444S.
286. Cook, D.J., *et al.* (1995) “Clinical recommendations using levels of evidence for antithrombotic agents”. *Chest,* 108(4 Suppl), 227S-230S.
287. Guyatt, G.H., *et al.* (2001) “Grades of recommendation for antithrombotic agents”. *Chest,* 119(1 Suppl), 3S-7S.
288. Sackett, D.L. (1986) “Rules of evidence and clinical recommendations on the use of antithrombotic agents”. *Chest,* 89(2 Suppl), 2S-3S.
289. Glover, J. (2011) *EBM Pyramid: EBM Page Generator*. Available at: www.ebmpyramid.org, accessed 08/05/13.
290. Melnyk, B.M. & Fineout-Overholt, E. (2005) *Evidence-Based Practice in Nursing and Healthcare: A Guide to Best Practice*. Philadelphia: Lippincoot, Williams & Wilkins.
291. Costantino, G., Montano, N. & Casazza, G. (Forthcoming) “When should we change our clinical practice based on the results of a clinical study? The hierarchy of evidence”. *Internal and emergency medicine,* preview available at: http://www.researchgate.net/publication/274722246_When_should_we_change_our_clinical_practice_based_on_the_results_of_a_clinical_study_The_hierarchy_of_evidence, accessed 22/07/15.
292. Olkin, I. (1995) “Statistical and theoretical considerations in meta-analysis”. *Journal of Clinical Epidemiology,* 48(1), 133-146.
293. Devereaux, P. & Yusuf, S. (2003) “The evolution of the randomized controlled trial and its role in evidence-based decision making”. *Journal of internal medicine,* 254(2), 105-113.
294. Cardarelli, R., Virgilio, R.F. & Taylor, L. (2007) “Evidence-based medicine, part 2. An introduction to critical appraisal of articles on therapy”. *The Journal of the American Osteopathic Association,* 107(8), 299-303.
295. Lang, E. (2004) “The Why and How of Evidence-Based Medicine”. *McGill Journal of Medicine,* 8, 90-94.

296. Greenhalgh, T. (1997) “How to read a paper. Getting your bearings (deciding what the paper is about)”. *BMJ,* 315(7102), 243-246.

297. Brozek, J.L., *et al.* (2009) “Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies”. *Allergy,* 64(8), 1109-1116.

298. Sackett, D.L. (1989) “Rules of evidence and clinical recommendations on the use of antithrombotic agents”. *Chest,* 95(2 Suppl), 2S-4S.

299. Roman, S.H., Silberzweig, S.B. & Siu, A.L. (2000) “Grading the evidence for diabetes performance measures”. *Eff Clin Pract,* 3(2), 85-91.

300. Hahn, S., *et al.* (2000) “Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies”. *Statistics in medicine,* 19(24), 3325-3336.

301. Mackway-Jones, K., *et al.* (1998) “The best evidence topic report: a modified CAT for summarising the available evidence in emergency medicine”. *J Accid Emerg Med,* 15(4), 222-226.

302. Ebbert, J.O., Montori, V.M. & Schultz, H.J. (2001) “The journal club in postgraduate medical education: a systematic review”. *Med Teach,* 23(5), 455-461.

303. Petrisor, B.A. & Bhandari, M. (2006) “Principles of teaching evidence-based medicine”. *Injury,* 37(4), 335-339.

304. Sullivan, B.M. & Cambron, J.A. (2008). *Overview of Study Designs in Clinical Research [Presentation]*. Paper presented at the EBP@NUHS CH5 Study Design.

305. van Tulder, M., *et al.* (2000) “Exercise therapy for low back pain: a systematic review within the framework of the cochrane collaboration back review group”. *Spine (Phila Pa 1976),* 25(21), 2784-2796.

306. Kerry, R., *et al.* (2012) “Causation and evidence-based practice: an ontological review”. *Journal of Evaluation in Clinical Practice,* 18(5), 1006-1012.

307. Barton, S. (2000) “Which clinical studies provide the best evidence? The best RCT still trumps the best observational study”. *BMJ,* 321(7256), 255-256.

308. Haigh, R. (2012) “Being Economical with the Evidence”. *Group Analysis,* 45, 70-83.

309. Gross, P.A., *et al.* (1994) “Purpose of quality standards for infectious diseases. Infectious Diseases Society of America”. *Clin Infect Dis,* 18(3), 421.

310. Evans, W.K., *et al.* (1997) “Lung cancer practice guidelines: lessons learned and issues addressed by the Ontario Lung Cancer Disease Site Group”. *J Clin Oncol,* 15(9), 3049-3059.

311. Elamin, M.B. & Montori, V.M. (2012) “The Hierarchy of Evidence: From Unsystematic Clinical Observations to Systematic Reviews”, In Burneo, J.G. (Ed.), *Neurology: an evidence-based approach* (pp. 11-24): Springer Science & Business Media, LLC.

312. Edlund, W., *et al.* (2004) *Clinical Practice Guideline Process Manual, Appendix 9*. American Academy of Neurology: available at: www.aan.com, accessed 01/04/13.

313. von Elm, E., *et al.* (2007) “Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies”. *BMJ,* 335(7624), 806-808.

314. Schulz, K.F. (1994) “Subverting randomization in controlled trials”. *Journal of the American Medical Association,* 274(18), 1456-1458.

315. Altman, D.G. & Schulz, K.F. (2001) “Concealing treatment allocation in randomised trials”. *British Medical Journal,* 323, 446.

316. Schulz, K.F., *et al.* (2010) “CONSORT 2010: Changes and testing blindness in RCTs”. *Lancet,* 375, 1144-1146.

317. Gigerenzer, G. (1991) “How to make cognitive illusions disappear: Beyond “heuristics and biases””. *European review of social psychology,* 2(1), 83-115.

318. Rychetnik, L., *et al.* (2002) “Criteria for evaluating evidence on public health interventions”. *J Epidemiol Community Health,* 56(2), 119-127.

319. Glasziou, P., Vandenbroucke, J. & Chalmers, I. (2004) “Assessing the quality of research”. *BMJ,* 328, 39-41.

320. Paul, J.R. (1958) *Clinical epidemiology*. Chicago: University of Chicago Press.

321. Feinstein, A.R. (1985) *Clinical epidemiology : the architecture of clinical research*. Philadelphia ; London: Saunders.
322. Feinstein, A.R. (1968) "Clinical epidemiology. I. The populational experiments of nature and of man in human illness". *Ann Intern Med,* 69(4), 807-820.
323. Feinstein, A.R. (1977) *Clinical biostatistics*. Saint Louis: Mosby.
324. Feinstein, A.R. (1987) *Clinimetrics*. New Haven ; London: Yale University Press.
325. Cochrane, A.L. (1972) *Effectiveness and efficiency : random reflections on health services*. London: Royal Society of Medicine Press.
326. McGowan, J.E.*, et al.* (1992) "Guidelines for the use of systemic glucocorticosteroids in the management of selected infections". *Journal of Infectious Disease,* 165, 1-13.
327. Hoogendoorn, W.E.*, et al.* (1999) "Physical load during work and leisure time as risk factors for back pain". *Scand J Work Environ Health,* 25(5), 387-403.
328. Greer, N.*, et al.* (2000) "A practical approach to evidence grading". *Joint Commission Journal on Quality Improvement,* 26, 700-712.
329. Cochrane, A.L. (1979) "Personal Papers: Forty years back: a retrospective survey". *Br Med J,* 2(6205), 1662-1663.
330. Chalmers, I. (1993) "The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care". *Ann N Y Acad Sci,* 703, 156-163; discussion 163-155.
331. Eddy, D.M. (1990) "Practice policies: where do they come from?". *JAMA,* 263(9), 1265, 1269, 1272 passim.
332. Shin, J.H. & Haynes, R.B. (1991) "Does a problem-based, selfdirected undergraduate medical curriculum promote continuing clinical competence?". *Clin Res,* 39, 143A.
333. Oliver, R.M., Wehby, J.H. & Reschly, D.J. (2011) "Teacher classroom management practices: effects on disruptive or aggressive student behavior". *Campbell Systematic Reviews,* 4.
334. Oxman, A.D., Sackett, D.L. & Guyatt, G.H. (1993) "Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group". *JAMA,* 270(17), 2093-2095.
335. Bigby, M. (1998) "Evidence-based medicine in a nutshell. A guide to finding and using the best evidence in caring for patients". *Arch Dermatol,* 134(12), 1609-1618.
336. Sackett, D.L. (1995) "Applying overviews and meta-analyses at the bedside". *J Clin Epidemiol,* 48(1), 61-66; discussion 67-70.
337. Guyatt, G.H.*, et al.* (2008) "Advanced Topics in the Validity of Therapy Trials: The Principle of Intention-to-Treat", In Guyatt, G.H.R., D., (Ed.), *Users' Guides to the Medical Literature* (3rd ed., pp. 167-178). Chicago: AMA Press.
338. Montori, V.M. & Guyatt, G.H. (2001) "Intention-to-treat principle". *CMAJ,* 165(10), 1339-1341.
339. Straus, S.E. (2011) *Evidence-based medicine: how to practice and teach it* (4th ed.). Edinburgh: Elsevier Churchill Livingstone.
340. Gray, J.A.M. (1997) *Evidence-Based Healthcare*. NYC: Churchill-Livingstone.
341. Guyatt, G.H.*, et al.* (2000) "Practitioners of evidence based care. Not all clinicians need to appraise evidence from scratch but all need some skills". *BMJ,* 320(7240), 954-955.
342. Sackett, D.L. (2010) "Clinician-trialist rounds: 1. Inauguration, and an introduction to time-management for survival". *Clin Trials,* 7(6), 749-751.
343. Sackett, D.L. (2011) "Clinician-trialist rounds: 2. time-management of your clinical practice and teaching". *Clin Trials,* 8(1), 112-114.
344. Dans, A.L. & Dans, L.F. (2000) "The need and means for evidence-based medicine in developing countries". *ACP J Club,* 133(1), A11-12.
345. Lomas, J. (1997) *Improving research dissemination and uptake in the health sector: beyond the sound of one hand clapping*. Hamilton, Ontario: McMaster University Centre for Health Economics and Policy Analysis.
346. Straus, S.E. & Jones, G. (2004) "What has evidence based medicine done for us?". *BMJ,* 329(7473), 987-988.
347. Aberle, D.F. (1966) *The Peyote Religion Among the Navaho*. Chicago: Aldine.
348. Hurwitz, B. (2004) "How does evidence based guidance influence determinations of medical negligence?". *BMJ,* 329(7473), 1024-1028.

349. West, S., *et al.* (2002) "Systems to rate the strength of scientific evidence". *Evid Rep Technol Assess (Summ)*(47), 1-11.
350. Guyatt, G.H., *et al.* (2011) "GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology". *J Clin Epidemiol,* 64(4), 380-382.
351. Guyatt, G.H., *et al.* (2006) "An emerging consensus on grading recommendations?". *Evidence Based Medicine,* 11(1), 2-4.
352. Coleman, K., *et al.* (2009) *NHMRC additional levels of evidence and grades for recommendations for developers of guidelines*. Melbourne: Australian National Health and Medical Research Council (ANHMRC).
353. Hillier, S., *et al.* (2011) "FORM: an Australian method for formulating and grading recommendations in evidence-based clinical guidelines". *BMC Med Res Methodol,* 11(1), 23.
354. National Institute for Health and Care Excellence (NICE). (2009/2015) *CG78: Borderline personality disorder: treatment and management*. London: RCP.
355. Australian Cancer Network Melanoma Guidelines Revision Working Party. (2008) *Clinical practice guidelines for the management of melanoma in Australia and New Zealand: Evidence-based best practice guidelines*. Wellington, NZ: Cancer Council Australia, Australian Cancer Network, and Sydney and New Zealand Guidelines Group.
356. Guyatt, G.H. (2008) "How to use a patient management recommendation", In Guyatt, G.H. & Rennie, D. (Eds.), *Users' Guides to the Medical Literature* (pp. 595-618). Chicago: American Medical Association Press.
357. Swanson, J.A., Schmicz, D. & Chung, K.C. (2010) "How to practice evidence based medicine". *Plast Reconstr Surg,* 126, 286.
358. Haynes, R.B. (2006) *Clinical epidemiology : how to do clinical practice research* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.
359. Miles, A., *et al.* (2000) "New perspectives in the evidence-based healthcare debate". *Journal of Evaluation in Clinical Practice,* 6, 77-84.
360. Buetow, S., *et al.* (2006) "Taking stock of evidence-based medicine: opportunities for its continuing evolution". *J Eval Clin Pract,* 12(4), 399-404.
361. Gallagher, E.J. (1999) "P< 0.05: threshold for decerebrate genuflection". *Academic Emergency Medicine,* 6(11), 1084-1087.
362. Cartwright, N. (2007) "Are RCTs the Gold Standard?". *Biosocieties,* 2, 11-20.
363. Tversky, A. & Kahneman, D. (1974) "Judgment under uncertainty: Heuristics and Biases". *Science,* 185, 1124-1131.
364. Kahneman, D. & Frederick, S. (2002) "Representativeness revisited: Attribute substitution in intuitive judgment.", In Gilovich, T, Griffin, D & Kahneman, D (Eds.), *Heuristics and Biases* (pp. 49-81). New York: CUP.
365. MacMahon, S. & Collins, R. (2001) "Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies". *Lancet,* 357(9254), 455-462.
366. Bhandari, M., *et al.* (2004) "Hierarchy of evidence: differences in results between non-randomized studies and randomized trials in patients with femoral neck fractures". *Arch Orthop Trauma Surg,* 124(1), 10-16.
367. Ioannidis, J.P., *et al.* (2001) "Comparison of evidence of treatment effects in randomized and nonrandomized studies". *JAMA: the journal of the American Medical Association,* 286(7), 821-830.
368. Chalmers, T.C., *et al.* (1983) "Bias in treatment assignment in controlled clinical trials". *N Engl J Med,* 309(22), 1358.
369. Kunz, R. & Oxman, A.D. (1998) "The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials". *BMJ: British Medical Journal,* 317(7167), 1185.
370. Colditz, G.A., Miller, J.N. & Mosteller, F. (1989) "How study design affects outcomes in comparisons of therapy. I: Medical". *Statistics in medicine,* 8(4), 441-454.
371. Concato, J., Shah, N. & Horwitz, R.I. (2000) "Randomized, controlled trials, observational studies, and the hierarchy of research designs". *N Engl J Med,* 342(25), 1887-1892.

372. Concato, J. (2004) "Observational versus experimental studies: what's the evidence for a hierarchy?". *NeuroRx,* 1(3), 341-347.

373. Benson, K. & Hartz, A.J. (2000) "A comparison of observational studies and randomized, controlled trials". *Am J Ophthalmol,* 130(5), 688.

374. Odgaard-Jensen, J.*, et al.* (2011) "Randomisation to protect against selection bias in healthcare trials". *The Cochrane Library*.

375. Echt, D.S.*, et al.* (1991) "Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial". *N Engl J Med,* 324(12), 781-788.

376. Morganroth, J., Bigger, J.T., Jr. & Anderson, J.L. (1990) "Treatment of ventricular arrhythmias by United States cardiologists: a survey before the Cardiac Arrhythmia Suppression Trial results were available". *Am J Cardiol,* 65(1), 40-48.

377. Moore, T.J. (1995) *Deadly medicine: why tens of thousands of heart patients died in America's worst drug disaster*: Simon & Schuster.

378. Bucher, H.C.*, et al.* (2002) "Advanced Topic in Applying the Results of Therapy Trials: Surrogate Outcomes", In Guyatt, G. & Rennie, D. (Eds.), *Users' Guides to the Medical Literature*.

379. Califf, R.M.*, et al.* (1997) "A randomized controlled trial of epoprostenol therapy for severe congestive heart failure: The Flolan International Randomized Survival Trial (FIRST)". *Am Heart J,* 134(1), 44-54.

380. Packer, M.*, et al.* (1991) "Effect of oral milrinone on mortality in severe chronic heart failure. The PROMISE Study Research Group". *N Engl J Med,* 325(21), 1468-1475.

381. Claridge, J.A. & Fabian, T.C. (2005) "History and development of evidence-based medicine". *World journal of surgery,* 29(5), 547-553.

382. Weatherall, D. (2001) "Foreword", In Greenhalgh, T. (Ed.), *How To Read A Paper: the basics of Evidence-Based Medicine* (pp. ix-xiii). London: BMJ Books.

383. Crowley, P. (1981) "Corticosteroids in pregnancy: the benefits outweigh the costs". *Journal of Obstetrics & Gynecology,* 1(3), 147-150.

384. Stiell, I.G.*, et al.* (1992) "A study to develop clinical decision rules for the use of radiography in acute ankle injuries". *Ann Emerg Med,* 21(4), 384-390.

385. Stiell, I.*, et al.* (1995) "Multicentre trial to introduce the Ottawa ankle rules for use of radiography in acute ankle injuries". *BMJ,* 311(7005), 594-597.

386. de Dombal, F.T.*, et al.* (1972) "Computer-aided diagnosis of acute abdominal pain". *Br Med J,* 2(5804), 9-13.

387. Grove, W.M.*, et al.* (2000) "Clinical versus mechanical prediction: a meta-analysis". *Psychol Assess,* 12(1), 19-30.

388. Grossman, J. & Mackenzie, F.J. (2005) "The randomized controlled trial: gold standard, or merely standard?". *Perspectives in Biology and Medicine,* 48(4), 516-534.

389. Howick, J. (2011) "Exposing the vanities—and a qualified defense—of mechanistic reasoning in health care decision making". *Philosophy of Science,* 78(5), 926-940.

390. Pocock, S.J. & Elbourne, D.R. (2000) "Randomized trials or observational tribulations?". *New England Journal of Medicine,* 342(25), 1907-1909.

391. Pocock, S.J. (1993) "Statistical and ethical issues in monitoring clinical trials". *Statistics in medicine,* 12(15-16), 1459-1469.

392. Leibovici, L. (2001) "Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial". *BMJ,* 323(7327), 1450-1451.

393. Tonelli, M.R. (2011) "Not a philosophy of clinical medicine: a commentary on 'The Philosophy of Evidence-based Medicine' Howick, J. (2001)". *Journal of Evaluation in Clinical Practice,* 17(5), 1013-1017.

394. Papineau, D. (1994) "The virtues of randomization". *The British journal for the philosophy of science,* 45(2), 437-450.

395. Cartwright, N. (1989) *Nature's Capacities and their Measurement*: Clarendon Press Oxford.

396. Pearl, J. (2000) *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.

397. Urbach, P. (1985) "Randomization and the design of experiments". *Philosophy of Science*, 256-273.

398. Urbach, P. (1994) “Reply to David Papineau”. *British journal for the philosophy of science*, 712-715.

399. Giere, R.N. (1979) “Understanding scientific reasoning”. *New York: Holt, Rinehart and Winston*.

400. Schulz, K.F. & Grimes, D.A. (2002) “Allocation concealment in randomised trials: defending against deciphering”. *The Lancet,* 359(9306), 614-618.

401. Roberts, C. & Torgerson, D.J. (1999) “Understanding controlled trials: baseline imbalance in randomised controlled trials”. *BMJ: British Medical Journal,* 319(7203), 185.

402. Vickers, A.J. & Altman, D.G. (2001) “Analysing controlled trials with baseline and follow up measurements”. *BMJ,* 323(7321), 1123-1124.

403. Peto, R. (1999) “Failure of randomisation by “sealed” envelope”. *Lancet,* 354(9172), 73.

404. Schulz, K.F.*, et al.* (1995) “Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials”. *JAMA,* 273(5), 408-412.

405. Brown, S.*, et al.* (2005) “Minimization—reducing predictability for multi-centre trials whilst retaining balance within centre”. *Statistics in medicine,* 24(24), 3715-3727.

406. Taves, D.R. (1974) “Minimization: a new method of assigning patients to treatment and control groups”. *Clin Pharmacol Ther,* 15(5), 443.

407. Scott, N.W.*, et al.* (2002) “The method of minimization for allocation to clinical trials: a review”. *Control Clin Trials,* 23(6), 662-674.

408. Howick, J. (2012) “Saying things the "right" way: avoiding "nocebo" effects and providing full informed consent”. *Am J Bioeth,* 12(3), 33-34.

409. Fergusson, D.*, et al.* (2004) “Turning a blind eye: the success of blinding reported in a randomised sample of randomised, placebo controlled trials”. *British Medical Journal,* 328, 432.

410. Hróbjartsson, A.*, et al.* (2007) “Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding”. *Int J Epidemiol,* 36(3), 654-663.

411. Lundh, A.*, et al.* (2012) “Industry sponsorship and research outcome”. *Cochrane Database Syst Rev,* 12.

412. Huss, A.*, et al.* (2008) “Source of funding and results of studies of health effects of mobile phone use: systematic review of experimental studies”. *Cien Saude Colet,* 13(3), 1005-1012.

413. Davidson, R.A. (1986) “Source of funding and outcome of clinical trials”. *J Gen Intern Med,* 1(3), 155-158.

414. Friedman, L.S. & Richter, E.D. (2004) “Relationship between conflicts of interest and research results”. *J Gen Intern Med,* 19(1), 51-56.

415. Lexchin, J.*, et al.* (2003) “Pharmaceutical industry sponsorship and research outcome and quality: systematic review”. *BMJ,* 326(7400), 1167-1170.

416. Sismondo, S. (2008) “Pharmaceutical company funding and its consequences: a qualitative systematic review”. *Contemporary Clinical Trials,* 29(2), 109-113.

417. Bhandari, M.*, et al.* (2004) “Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials”. *Canadian Medical Association Journal,* 170(4), 477-480.

418. Dickersin, K. & Min, Y. (1993) “Publication bias: the problem that won't go away”. *Ann N Y Acad Sci,* 703(1), 135-148.

419. Dickersin, K., Min, Y. & Meinert, C.L. (1992) “Factors influencing publication of research results”. *JAMA: the journal of the American Medical Association,* 267(3), 374-378.

420. Rennie, D. (1997) “Thyroid storm”. *JAMA-Journal of the American Medical Association-International Edition,* 277(15), 1238-1243.

421. Dong, B.J.*, et al.* (1997) “Bioequivalence of generic and brand-name levothyroxine products in the treatment of hypothyroidism”. *JAMA: the journal of the American Medical Association,* 277(15), 1205-1213.

422. Howick, J. (2012) *Is EBM a new Kuhnian Paradigm?* [Presentation] Available at: http://prezi.com/a_plwmcguwom/is-ebm-a-new-kuhnian-paradigm (19/07/12), accessed 15/03/13

423. Rubin, D.B. (1974) “Estimating causal effects of treatments in randomized and nonrandomized studies”. *Journal of educational Psychology,* 66(5), 688.

424. Varadhan, K.K., *et al.* (2010) "Antibiotic therapy versus appendectomy for acute appendicitis: a meta-analysis". *World journal of surgery,* 34(2), 199-209.

425. Fitzmaurice, G.J., *et al.* (2011) "Antibiotics versus appendectomy in the management of acute appendicitis: a review of the current evidence". *Canadian Journal of Surgery,* 54(5), 307.

426. Smith, G.C. & Pell, J.P. (2006) "Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials". *International Journal of Prosthodontics,* 19(2), 126.

427. Ney, P.G., Collins, C. & Spensor, C. (1986) "Double blind: double talk or are there ways to do better research". *Med Hypotheses,* 21(2), 119-126.

428. Howick, J. (2009) "Questioning the methodologic superiority of 'placebo' over 'active' controlled trials". *Am J Bioeth,* 9(9), 34-48.

429. Howick, J. (2009) "Reviewing the unsubstantiated claims for the methodological superiority of 'placebo' over 'active' controlled trials: reply to open peer commentaries". *Am J Bioeth,* 9(9), W5-7.

430. Howick, J. (2009) "Placebo misery. Escaping from placebo prison". *BMJ,* 338, b1898.

431. Anderson, J.A. (2006) "The ethics and science of placebo-controlled trials: Assay sensitivity and the Duhem–Quine thesis". *Journal of Medicine and Philosophy,* 31(1), 65-81.

432. Temple, R. & Ellenberg, S.S. (2000) "Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues". *Annals of Internal Medicine,* 133(6), 455-463.

433. Russo, F. (2012) "Philosophy of medicine: between clinical trials and mechanisms". *Metascience,* 21(2), 387-390.

434. Schunemann, H., *et al.* (2011) "The GRADE approach and Bradford Hill's criteria for causation". *J Epidemiol Community Health,* 65(5), 392-395.

435. Mant, D. (1999) "Can randomised trials inform clinical decisions about individual patients?". *Lancet,* 353(9154), 743-746.

436. Feinstein MD, A.R. & Horwitz MD, R.I. (1997) "Problems in the "evidence" of "evidence-based medicine"". *Am J Med,* 103(6), 529-535.

437. Upshur, R.E. (2005) "Looking for rules in a world of exceptions: reflections on evidence-based practice". *Perspect Biol Med,* 48(4), 477-489.

438. Fries, J.F. & Krishnan, E. (2004) "Equipoise, design bias, and randomized controlled trials: the elusive ethics of new drug development". *Arthritis Res Ther,* 6(3), R250-R255.

439. McMurdo, M.E., Witham, M.D. & Gillespie, N.D. (2005) "Including older people in clinical research: Benefits shown in trials in younger people may not apply to older people". *BMJ: British Medical Journal,* 331(7524), 1036.

440. Masoudi, F.A., *et al.* (2003) "Most hospitalized older persons do not meet the enrollment criteria for clinical trials in heart failure". *American Heart Journal,* 146(2), 250-257.

441. Van Spall, H.G., *et al.* (2007) "Eligibility criteria of randomized controlled trials published in high-impact general medical journals". *JAMA: the journal of the American Medical Association,* 297(11), 1233-1240.

442. Schain, W.S. (1994) "Barriers to clinical trials: Part II: Knowledge and attitudes of potential participants". *Cancer,* 74(S9), 2666-2671.

443. Dresser, R. (2002) "The ubiquity and utility of the therapeutic misconception". *Social philosophy and policy,* 19(02), 271-294.

444. Appelbaum, P.S., *et al.* (1987) "False hopes and best data: consent to research and the therapeutic misconception". *Hastings Center Report,* 17(2), 20-24.

445. Vist, G.E., *et al.* (2008) "Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate". *Cochrane Database Syst Rev,* 3.

446. Wells, R.E. & Kaptchuk, T.J. (2012) "To tell the truth, the whole truth, may do patients harm: the problem of the nocebo effect for informed consent". *American Journal of Bioethics,* 12(3), 22-29.

447. Swanson, J. (2003) "Compliance with stimulants for attention-deficit/hyperactivity disorder". *CNS drugs,* 17(2), 117-131.

448. Small, G. & Dubois, B. (2007) “A review of compliance to treatment in Alzheimer's disease: potential benefits of a transdermal patch”. *Curr Med Res Opin,* 23(11), 2705-2713.

449. Schneider, L.S.*, et al.* (1997) “Eligibility of Alzheimer's disease clinic patients for clinical trials”. *Journal of the American Geriatrics Society,* 45(8), 923.

450. Boyd, C.M.*, et al.* (2005) “Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases”. *JAMA: the journal of the American Medical Association,* 294(6), 716-724.

451. Blumer, J.L. (1999) “Off-label uses of drugs in children”. *Pediatrics,* 104(Supplement 3), 598-602.

452. Cuzzolin, L., Zaccaron, A. & Fanos, V. (2003) “Unlicensed and off-label uses of drugs in paediatrics: a review of the literature”. *Fundamental & clinical pharmacology,* 17(1), 125-131.

453. Henderson, L.*, et al.* (2002) “St John's wort (Hypericum perforatum): drug interactions and clinical outcomes”. *Br J Clin Pharmacol,* 54(4), 349-356.

454. Markowitz, J.S.*, et al.* (2003) “Effect of St John's wort on drug metabolism by induction of cytochrome P450 3A4 enzyme”. *JAMA,* 290(11), 1500-1504.

455. Izzo, A. (2004) “Drug interactions with St. John's Wort (Hypericum perforatum): a review of the clinical evidence”. *International journal of clinical pharmacology and therapeutics,* 42(3), 139-148.

456. Taylor, B.J., Dempster, M. & Donnelly, M. (2007) “Grading gems: Appraising the quality of research for social work and social care”. *British Journal of Social Work,* 37(2), 335-354.

457. Greenfield, S.*, et al.* (2007) “Heterogeneity of treatment effects: implications for guidelines, payment, and quality assessment”. *Am J Med,* 120(4), S3-S9.

458. Dans, A.L., Dans, L.F. & Guyatt, G. (2008) “Advanced topics in applying the results of therapy trials: Applying results to individual patients”, In Guyatt, G.*, et al.* (Eds.), *Users' guides to the medical literature* (2nd ed., pp. 273-290). New York: McGraw Hill Medical.

459. Charlton, B.G. (1994) “Understanding randomized controlled trials: explanatory or pragmatic?”. *Fam Pract,* 11(3), 243-244.

460. Wilkinson, I.*, et al.* (2010) *Oxford Handbook of Clinical Medicine*: Oxford University Press.

461. Truog, R.D. & Arnold, J.H. (1992) “The "ethics of evidence" and randomized controlled trials”. *J Clin Ethics,* 3(1), 65-67.

462. Miller, F.G. & Brody, H. (2003) “A critique of clinical equipoise: therapeutic misconception in the ethics of clinical trials”. *Hastings Center Report,* 33(3), 19-28.

463. Freedman, B. (1987) “Equipoise and the ethics of clinical research”. *N Engl J Med*.

464. Pocock, S.J. (1992) “When to stop a clinical trial”. *BMJ: British Medical Journal,* 305(6847), 235.

465. Emery, P.*, et al.* (2008) “IL-6 receptor inhibition with tocilizumab improves treatment outcomes in patients with rheumatoid arthritis refractory to anti-tumour necrosis factor biologicals: results from a 24-week multicentre randomised placebo-controlled trial”. *Ann Rheum Dis,* 67(11), 1516-1523.

466. Oldfield, V., Dhillon, S. & Plosker, G.L. (2009) “Tocilizumab”. *Drugs,* 69(5), 609-632.

467. Yokota, S.*, et al.* (2008) “Efficacy and safety of tocilizumab in patients with systemic-onset juvenile idiopathic arthritis: a randomised, double-blind, placebo-controlled, withdrawal phase III trial”. *Lancet,* 371(9617), 998-1006.

468. Bongartz, T. (2008) “Tocilizumab for rheumatoid and juvenile idiopathic arthritis”. *Lancet,* 371(9617), 961-963.

469. Deighton, C.*, et al.* (2009) “Management of rheumatoid arthritis: summary of NICE guidance”. *BMJ,* 338.

470. Yoshida, K.*, et al.* (2011) “An observational study of tocilizumab and TNF-α inhibitor use in a Japanese community hospital: different remission rates, similar drug survival and safety”. *Rheumatology,* 50(11), 2093-2099.

471. Prevoo, M.*, et al.* (1995) “Modified disease activity scores that include twenty-eight-joint counts development and validation in a prospective longitudinal study of patients with rheumatoid arthritis”. *Arthritis & Rheumatism,* 38(1), 44-48.

472. Fransen, J. & Van Riel, P. (2005) “The Disease Activity Score and the EULAR response criteria”. *Clin Exp Rheumatol,* 23(5), S93.

473. Fransen, J., Creemers, M. & Van Riel, P. (2004) “Remission in rheumatoid arthritis: agreement of the disease activity score (DAS28) with the ARA preliminary remission criteria”. *Rheumatology,* 43(10), 1252-1255.

474. Thiele, K.*, et al.* (2013) “Performance of the 2011 ACR/EULAR preliminary remission criteria compared with DAS28 remission in unselected patients with rheumatoid arthritis”. *Ann Rheum Dis,* 72(7), 1194-1199.

475. Bykerk, V.P. & Massarotti, E.M. (2012) “The new ACR/EULAR remission criteria: rationale for developing new criteria for remission”. *Rheumatology,* 51(suppl 6), vi16-vi20.

476. Zarin, D.A.*, et al.* (2011) “The ClinicalTrials. gov results database—update and key issues”. *New England Journal of Medicine,* 364(9), 852-860.

477. Oxman, A.D. & Guyatt, G.H. (1992) “A consumer's guide to subgroup analyses”. *Annals of Internal Medicine,* 116(1), 78-84.

478. Guyatt, G.H., Wyer, P.C. & Ioannidis, J.P. (2002) “Advanced topics in systematic reviews: When to believe a subgroup analysis”, In Guyatt, G. & Rennie, D. (Eds.), *Users' Guides to the Medical Literature* (pp. 571-593). McGraw Hill Medical: New York.

479. Song, F.*, et al.* (2003) “Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses”. *BMJ,* 326(7387), 472.

480. Bucher, H.C.*, et al.* (1997) “The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials”. *Journal of Clinical Epidemiology,* 50(6), 683-691.

481. Kravitz, R.L., Duan, N. & Braslow, J. (2004) “Evidence-Based Medicine, Heterogeneity of Treatment Effects and the Trouble with Averages”. *The Milbank Quarterly,* 82(4), 661-687.

482. Gabler, N.B.*, et al.* (2009) “Dealing with heterogeneity of treatment effects: is the literature up to the challenge”. *Trials,* 10(1), 43-43.

483. Melander, H.*, et al.* (2003) “Evidence b (i) ased medicine—selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications”. *BMJ,* 326(7400), 1171-1173.

484. Dias, S.*, et al.* (2011) *NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment*. Sheffield: ScHARR, University of Sheffield: Decision Support Unit.

485. Sackett, D.L. (2004) “Turning a blind eye: why we don't test for blindness at the end of our trials”. *BMJ,* 328(7448), 1136.

486. Sackett, D.L. (2007) “Commentary: Measuring the success of blinding in RCTs: don't, must, can't or needn't?”. *Int J Epidemiol,* 36(3), 664-665.

487. Sackett, D.L. (2011) “Clinician-trialist rounds: 6. Testing for blindness at the end of your trial is a mug's game”. *Clin Trials,* 8(5), 674-676.

488. Lawton, B.A.*, et al.* (2008) “Exercise on prescription for women aged 40-74 recruited through primary care: two year randomised controlled trial”. *BMJ,* 337, a2509.

489. Spector, R. & Vesell, E.S. (2006) “Pharmacology and statistics: recommendations to strengthen a productive partnership”. *Pharmacology,* 78(3), 113-122.

490. Tonelli, M.R. & Callahan, T.C. (2001) “Why Alternative Medicine Cannot Be Evidence-based”. *Academic Medicine,* 76(12), 1213-1220.

491. Charlton, B.G. (2002) “Randomized trials in alternative/complementary medicine”. *QJM,* 95(10), 643-645.

492. Borgerson, K. (2005) “Evidence-based alternative medicine?”. *Perspectives in Biology and Medicine,* 48(4), 502-515.

493. Pigliucci, M. & Boudry, M. (2013) “The Dangers of Pseudoscience”. *The Stone: New York Times,* 10th October 2013 (available at: http://opinionator.blogs.nytimes.com/2013/10/10/the-dangers-of-pseudoscience/?smid=pl-share, accessed 07/08/15).

494. Harlan Jr, W.R. (2001) “New opportunities and proven approaches in complementary and alternative medicine research at the National Institutes of Health”. *Journal of Alternative & Complementary Medicine,* 7(1), 53-59.

495. Ernst, E., Pittler, M.H. & Wider, B. (2006) *The Desktop Guide to Complementary and Alternative Medicine: An Evidence-Based Approach*. Philadelphia: Mosby: Elsevier, Ltd.

496. Fontanarosa, P.B. & Lundberg, G.D. (1998) "Alternative medicine meets science". *JAMA,* 280(18), 1618-1619.

497. Cooley, K., *et al.* (2009) "Naturopathic care for anxiety: a randomized controlled trial ISRCTN78958974". *PLoS One,* 4(8), e6628.

498. Berman, B., *et al.* (1999) "A randomized trial of acupuncture as an adjunctive therapy in osteoarthritis of the knee". *Rheumatology,* 38(4), 346-354.

499. Shen, J., *et al.* (2000) "Electroacupuncture for control of myeloablative chemotherapy–Induced emesis: A randomized controlled trial". *JAMA,* 284(21), 2755-2761.

500. Cherkin, D.C., *et al.* (2003) "A review of the evidence for the effectiveness, safety, and cost of acupuncture, massage therapy, and spinal manipulation for back pain". *Annals of Internal Medicine,* 138(11), 898-906.

501. Ernst, E. (2010) "Winnowing the chaff of charlatanism from the wheat of science". *Evidence-Based Complementary and Alternative Medicine,* 7(4), 425-426.

502. American Psychiatric Association. (1974) *Diagnostic and statistical manual of mental disorders* (2nd ed., 7th printing ed.). Washington, D.C.: American Psychiatric Association.

503. American Psychiatric Association. (2000) *Therapies focused on attempts to change sexual orientation (Reparative or conversion therapies): Position statement*. Arlington, Va: American Psychiatric Association Board of Trustees Retrieved from http://web.archive.org/web/20110407082738/http://www.psych.org/Departments/EDU/Library/APAOfficialDocumentsandRelated/PositionStatements/200001.aspx, accessed 13/02/14.

504. American Psychiatric Association. (1987) *Diagnostic and statistical manual of mental disorders* (III-R ed.). Washington, D.C.: American Psychiatric Association.

505. Silverstein, C. (2009) "The implications of removing homosexuality from the DSM as a mental disorder". *Archives of sexual behavior,* 38(2), 161-163.

506. Piaggio, G., *et al.* (2006) "Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement". *JAMA,* 295(10), 1152-1160.

507. Schumi, J. & Wittes, J.T. (2011) "Through the looking glass: understanding non-inferiority". *Trials,* 12(1), 106.

508. Ioannidis, J.P. (1998) "Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials". *JAMA,* 279(4), 281-286.

509. Chan, A.-W. & Altman, D.G. (2005) "Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors". *BMJ,* 330(7494), 753.

510. Turner, E.H., *et al.* (2008) "Selective publication of antidepressant trials and its influence on apparent efficacy". *New England Journal of Medicine,* 358(3), 252-260.

511. Huston, P. & Moher, D. (1996) "Redundancy, disaggregation, and the integrity of medical research". *Lancet,* 347(9007), 1024-1026.

512. Ernst, E. & Resch, K.L. (1994) "Reviewer bias: a blinded experimental study". *Journal of Laboratory and Clinical Medicine,* 124(2), 178-182.

513. Chan, A.-W., *et al.* (2004) "Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles". *JAMA,* 291(20), 2457-2465.

514. Chan, A.-W., *et al.* (2004) "Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research". *Canadian Medical Association Journal,* 171(7), 735-740.

515. Kirkham, J.J., *et al.* (2010) "The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews". *BMJ,* 340.

516. Hart, B., Lundh, A. & Bero, L. (2012) "Effect of reporting bias on meta-analyses of drug trials: reanalysis of meta-analyses". *BMJ,* 344.

517. Hopewell, S., *et al.* (2007) "Grey literature in meta-analyses of randomized trials of health care interventions". *Cochrane Database Syst Rev,* 2(2).

518. McAuley, L., Tugwell, P. & Moher, D. (2000) "Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses?". *The Lancet,* 356(9237), 1228-1231.

519. Relton, C. (2009) *A new design for pragmatic RCTs: a "patient cohort" RCT of treatment by a homeopath for menopausal hot flushes.* University of Sheffield: [PhD Thesis] ISRCTN 0287542.

520. Relton, C., *et al.* (2010) "Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design". *BMJ,* 340, c1066.

521. Relton, C., O'Cathain, A. & Nicholl, J. (2012) "A pilot 'cohort multiple randomised controlled trial'of treatment by a homeopath for women with menopausal hot flushes". *Contemporary Clinical Trials,* 33(5), 853-859.

522. Viksveen, P. & Relton, C. (2014) "Depression treated by homeopaths: a study protocol for a pragmatic cohort multiple randomised controlled trial". *Homeopathy,* 103(2), 147-152.

523. Mitchell, N., *et al.* (2011) "A randomised evaluation of CollAborative care and active surveillance for Screen-Positive EldeRs with sub-threshold depression (CASPER): study protocol for a randomized controlled trial". *Trials,* 12(1), 225.

524. Kwakkenbos, L., *et al.* (2013) "The Scleroderma Patient-centered Intervention Network (SPIN) Cohort: protocol for a cohort multiple randomised controlled trial (cmRCT) design to support trials of psychosocial and rehabilitation interventions in a rare disease context". *BMJ Open,* 3(8).

525. Relton, C., *et al.* (2011) "South Yorkshire Cohort: a'cohort trials facility'study of health and weight-Protocol for the recruitment phase". *BMC Public Health,* 11(1), 640.

526. Adamson, J., *et al.* (2006) "Review of randomised trials using the post-randomised consent (Zelen's) design". *Contemporary Clinical Trials,* 27(4), 305-319.

527. Zelen, M. (1979) "A new design for randomized clinical trials". *N Engl J Med,* 300(22), 1242-1245.

528. Zelen, M. (1990) "Randomized consent designs for clinical trials: an update". *Statistics in medicine,* 9(6), 645-656.

529. Zelen, M. (1969) "Play the winner rule and the controlled clinical trial". *Journal of the American Statistical Association,* 64(325), 131-146.

530. Ioannidis, J.P. (2005) "Why most published research findings are false". *PLoS Med,* 2(8), e124.

531. Ioannidis, J.P. (2007) "Why most published research findings are false: author's reply to Goodman and Greenland". *PLoS Med,* 4(6), e215.

532. Lipton, M.A., *et al.* (1973) "Megavitamin and orthomolecular therapy in psychiatry". *American Psychiatric Association Task Force Report*(7), 54.

533. Spector, R. (2009) "Science and pseudoscience in adult nutrition research and practice". *Skeptical Inquirer,* 33(3), 35-41.

534. Sauve, R.S., *et al.* (1990) "Megavitamin and megamineral therapy in childhood". *Canadian Medical Association Journal,* 143(10), 1009-1013.

535. Stephens, N.G., *et al.* (1996) "Randomised controlled trial of vitamin E in patients with coronary disease: Cambridge Heart Antioxidant Study (CHAOS)". *The Lancet,* 347(9004), 781-786.

536. Tatsioni, A., Bonitsis, N.G. & Ioannidis, J.P. (2007) "Persistence of contradicted claims in the literature". *JAMA,* 298(21), 2517-2526.

537. Goldacre, B. (2007) "Media Watch: Tell us the truth about nutritionists". *BMJ,* 334(7588), 292.

538. Goldacre, B. (2010) *Bad science: quacks, hacks, and big pharma flacks*: McClelland & Stewart.

539. Lewis, R. (2002) "Dietary supplements", In Shermer, M. & Linse, P. (Eds.), *Encyclopedia of Pseudoscience* (Vol. 1, pp. 85-92). Santa Barbara, California: ABC Clio.

540. Spector, R. & Johanson, C.E. (2007) "Vitamin transport and homeostasis in mammalian brain: focus on Vitamins B and E". *J Neurochem,* 103(2), 425-438.

541. Guallar, E., *et al.* (2013) "Enough is enough: stop wasting money on vitamin and mineral supplements". *Annals of Internal Medicine,* 159(12), 850-851.

542. Fortmann, S.P., *et al.* (2013) "Vitamin and mineral supplements in the primary prevention of cardiovascular disease and cancer: an updated systematic evidence review for the US Preventive Services Task Force". *Annals of Internal Medicine,* 159(12), 824-834.

543. Shearer, M. (1992) "Vitamin K metabolism and nutriture". *Blood reviews,* 6(2), 92-104.

544. Hathcock, J.N., *et al.* (2007) "Risk assessment for vitamin D". *The American journal of clinical nutrition,* 85(1), 6-18.

545. Barratt, A., *et al.* (2004) "Tips for learners of Evidence-Based Medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat". *CMAJ,* 171(4).

546. Cook, R.J. & Sackett, D.L. (1995) "The number needed to treat: a clinically useful measure of treatment effect". *BMJ,* 310(6977), 452-454.

547. Laupacis, A., Sackett, D.L. & Roberts, R.S.,. (1988) "An assessment of clinically-useful measures of the consequences of treatment". *New England Journal of Medicine,* 318, 1728-1733.

548. Cartwright, N. (2012) "Presidential address: Will this policy work for you? Predicting effectiveness better: How philosophy helps". *Philosophy of Science,* 79(5), 973-989.

549. Ushakov, N.G. (2012) "Unimodal distribution". *Encyclopedia of Mathematics,* European Mathematicl Society: Springer, available at: https://www.encyclopediaofmath.org/index.php/Unimodal_distribution, accessed 01/09/15.

550. Gregg, B. (2015) "Frequency trails: modes and modality" *Frequency trails*. Available at: http://www.brendangregg.com/FrequencyTrails/modes.html, accessed 17/08/15: New Jersey: Prentice Hall.

551. Moore, D.S. & McCabe, G.P. (1989) *Introduction to the Practice of Statistics*. New York: W H Freeman.

552. Gravetter, F. & Wallnau, L. (2006) *Statistics for the behavioral sciences*. Belmont, CA: Cengage Learning.

553. von Hippel, P.T. (2005) "Mean, median, and skew: Correcting a textbook rule". *Journal of Statistics Education,* 13(2), n2.

554. Siyavula. (2015) "Symmetric and skewed distributions" *Statistics*. Available at: http://everythingmaths.co.za/maths/grade-11/11-statistics/11-statistics-05.cnxmlplus, accessed 17/08/15: Siyavula.

555. Roses, A.D. (2000) "Pharmacogenetics and Future Drug Development and Delivery". *Lancet,* 355, 1358-1361.

556. Arends, L.R., *et al.* (2000) "Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses". *Statistics in medicine,* 19(24), 3497-3518.

557. Cox, D.R. (1984) "Interaction". *International Statistical Review,* 52, 1-31.

558. Smith, G.D., Song, F. & Sheldon, T.A. (1993) "Cholesterol lowering and mortality: the importance of considering initial level of risk". *BMJ,* 306(6889), 1367-1373.

559. Glasziou, P.P. & Irwig, L.M. (1995) "An evidence based approach to individualising treatment". *BMJ,* 311(7016), 1356-1359.

560. Smith, G.D. & Egger, M. (1994) "Who benefits from medical interventions?". *BMJ,* 308(6921), 72.

561. Rose, G. (1992) "The strategy of preventive medicine". *The strategy of preventive medicine.*

562. Rose, G. (2001) "Sick individuals and sick populations". *Int J Epidemiol,* 30(3), 427-432.

563. Tallarida, R.J. (2001) "Drug synergism: its detection and applications". *Journal of Pharmacology and Experimental Therapeutics,* 298(3), 865-872.

564. Qato, D.M., *et al.* (2008) "Use of prescription and over-the-counter medications and dietary supplements among older adults in the United States". *JAMA,* 300(24), 2867-2878.

565. Bickel, W.K., Marion, I. & Lowinson, J.H. (1987) "The treatment of alcoholic methadone patients: a review". *Journal of Substance Abuse Treatment,* 4, 15-19.

566. Liebson, I.A., Tommasello, A. & Bigelow, G.E. (1978) "A Behavioural Treatment of Alcoholic Methadone Patients". *Ann Intern Med,* 89, 342-344.

567. McCormick, A.W., *et al.* (2003) "Geographic diversity and temporal trends of antimicrobial resistance in Streptococcus pneumoniae in the United States". *Nature medicine,* 9(4), 424-430.

568. Hauben, M. & Aronson, J.K. (2006) "Paradoxical reactions: under-recognized adverse effects of drugs". *Drug Safety,* 29, 270.

569. Hauben, M. & Aronson, J.K. (2007) "Classifying Paradoxical Adverse Drug Reactions". *British Medical Journal; Rapid Responses,* Available at: http://www.bmj.com/content/333/7581/1267/reply, accessed 07/08/15.

570. Rothwell, P.M., *et al.* (2005) "From subgroups to individuals: general principles and the example of carotid endarterectomy". *Lancet,* 365, 256-265.

571. Cebul, R.D., *et al.* (1998) "Indications, outcomes and provider volumes for carotid endarterectomy". *Journal of the American Medical Association,* 279, 1282-1287.

572. Inztari, D., *et al.* (2000) "The causes and risk of stroke in patients with internal-carotid-artery stenosis". *New England Journal of Medicine,* 342, 1693-1700.

573. Busuttil, R.W., et al.,. (1981) "Carotid Artery Stenosis--Hemodynamic Significance and Clinical Course". *Journal of the American Medical Association,* 245(14), 1438-1441.

574. Wilterdink, J.L., *et al.* (1996) "Performance of Carotid Ultrasound in Evaluating Candidates for Carotid Endarterectomy is Optimized by an Approach Based on Clinical Outcome rather than Accuracy". *Stroke,* 27, 1094-1098.

575. Barnett, H.J., *et al.* (1998) "Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators". *N Engl J Med,* 339(20), 1415-1425.

576. European Carotid Surgery Trialists' Collaborative Group. (1998) "Randomised trial of endarterectomy for recently symptomatic carotid stenosis: final results of the MRC European Carotid Surgery Trial". *Lancet,* 273, 1379.

577. ACAS Executive Committee. (1995) "Endarterectomy for asymptomatic carotid artery stenosis". *Journal of the American Medical Association,* 273, 1421-1428.

578. Gould, D.A. & Birkmeyer, J.D. (1999) "Efficacy versus Effectiveness of Carotid Endarterectomy". *Effective Clinical Practice,* 2, 30-36.

579. Rothwell, P.M. & Warlow, C.P. (1999) "Prediction of benefit from carotid endarterectomy in individual patients: a risk modelling study". *Lancet,* 353, 2105-2110.

580. Wong, J.H., Findlay, J.M. & Suarez-Almazor, M.E. (1997) "Regional performance of carotid endarterectomy appropriateness, outcomes, and risk factors for complications". *Stroke,* 28(5), 891-898.

581. Tu, J.V., *et al.* (1998) "The rise and fall of carotid endarterectomy in the United States and Canada". *New England Journal of Medicine,* 339, 1441-1447.

582. Karp, H.R., *et al.* (1998) "Carotid endarterectomy among Medicare beneficiaries: a statewide evaluation of appropriateness and outcome". *Stroke,* 29, 46-52.

583. Williams, B.A. (2010) "Perils of Evidence Based Medicine". *Perspectives in Biology and Medicine,* 53, 106-120.

584. Coplen, S.E., *et al.* (1991) "Efficacy and safety of Quinidine Therapy for Maintenance of Sinus Rhythm after Cardioversion". *Circulation,* 83(2), 714.

585. Kannel, W.B., *et al.* (1982) "Epidemiological Features of Chronic Atrial Fibrillation - the Framingham Study". *New England Journal of Medicine,* 306, 1018-1022.

586. Lafuente-Lafuente, C., *et al.* (2006) "Antiarrhymthmic drugs for maintaining sinus rhythm after cardioversion of atrial fibrillation: a systematic review of randomized controlled trials". *Archives of Internal Medicine,* 166(7), 719-728.

587. Flaker, G.C., *et al.* (1992) "Antiarrhymthmic drug therapy and cardiac mortality in atrial fibrillation". *Journal of the American College of Cardiology,* 20(3), 527-532.

588. Stupp, R., *et al.* (2005) "Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma". *New England Journal of Medicine,* 352(10), 987-996.

589. Hegi, M.E., *et al.* (2005) "MGMT gene silencing and benefit from temozolomide in glioblastoma". *New England Journal of Medicine,* 352(10), 997-1003.

590. Sathornsumetee, S., *et al.* (2007) "Molecularly targeted therapy for malignant glioma". *Cancer,* 110(1), 13-24.

591. Furnari, F.B., *et al.* (2007) "Malignant astrocytic glioma: genetics, biology, and paths to treatment". *Genes & development,* 21(21), 2683-2710.

592. Hamdy, R.C., *et al.* (1982) "The Pharmacokinetics of Benoxaprofen in Elderly Subjects". *European Journal of Rheumatology and Inflammation,* 5, 69-75.

593. Kamal, A. & Koch, I.M. (1982) "Pharmacokinetic Studies of Benoxaprofen in Geriatric Patients". *European Journal of Rheumatology and Inflammation,* 76-81, 5.

594. World Health Organisation (WHO). (1981) "Health Care in the Elderly: Report of the Technical Group on Use of Medicaments by the Elderly". *Drugs,* 22, 279-294.

595. Kent, D.M., Ruthazer, R. & Selker, H.P. (2003) "Are some patients likely to benefit from recombinant tissue-type Plasminogen Activator for Acute Ischemic Stroke even beyond 3 hours from symptom onset?". *Stroke,* 34, 464-467.

596. Clark, W.M., *et al.* (1999) "Recombinant tissue-type plasminogen activator (alteplase) for ischemic stroke 3 to 5 hours after symptom onset: the ATLANTIS study, a randomized controlled trial: Alteplase Thrombolysis for Acute Noninternventional Therapy in Ischemic Stroke". *Journal of the American Medical Association,* 282, 2019-2026.

597. GUSTO Investigators group. (1993) "An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction". *New England Journal of Medicine,* 329, 673-682.

598. Kent, D.M., *et al.* (2002) "An independently derived and validated predictive model for selecting patients with myocardial infarction who are likely to benefit from tissue plasminogen activator compared with streptokinase". *American Journal of Medicine,* 113(2), 104-111.

599. Assmann, S.F., *et al.* (2000) "Subgroup analysis and other (mis) uses of baseline data in clinical trials". *The Lancet,* 355(9209), 1064-1069.

600. Brookes, S.T., *et al.* (2001) "Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives". *Health Technology Assessment,* 5(33), 1-56.

601. Rothwell, P.M. (2005) "Subgroup analysis in randomised controlled trials: importance, indications, and interpretation". *Lancet,* 365(9454), 176-186.

602. Sun, X., *et al.* (2012) "Credibility of claims of subgroup effects in randomised controlled trials: systematic review". *BMJ,* 344, e1553.

603. Brookes, S.T., *et al.* (2004) "Subgroup analyses in randomized trials: risks of subgroup-specific analyses: power and sample size for the interaction test". *Journal of Clinical Epidemiology,* 57(3), 229-236.

604. Athey, S. (2015). *Machine Learning and Causal Inference for Policy Evaluation.* Paper presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

605. Athey, S. & Imbens, G. (2015) "Machine Learning Methods for Estimating Heterogeneous Causal Effects". *arXiv preprint arXiv:1504.01132*.

606. Garson, J. (2010) "Connectionism", In Zalta, E.N. (Ed.), *Stanford Encyclopedia of Philosophy*. available at: http://plato.stanford.edu/entries/connectionism/, accessed 12/09/14.

607. Rumelhart, D.E. (1989) "The Architecture of Mind: A Connectionist Approach", In Posner, M (Ed.), *Foundations of Cognitive Science* (pp. 133-159). Cambridge, Mass: MIT Press.

608. Klein, R.J., *et al.* (2005) "Complement factor H polymorphism in age-related macular degeneration". *Science,* 308(5720), 385-389.

609. Upshur, R.E. (2006) "The complex, the exhausted and the personal: reflections on the relationship between evidence-based medicine and casuistry. Commentary on Tonelli (2006), Integrating evidence into clinical practice: an alternative to evidence-based approaches. Journal of Evaluation in Clinical Practice 12, 248-256". *J Eval Clin Pract,* 12(3), 281-288.

610. Bachmann, L.M., *et al.* (2003) "Accuracy of Ottawa ankle rules to exclude fractures of the ankle and mid-foot: systematic review". *BMJ,* 326(7386), 417.

611. Dimond, E.G., Kittle, C.F. & Crockett, J.E. (1960) "Comparison of internal mammary artery ligation and sham operation for angina pectoris". *Am J Cardiol,* 5(4), 483-486.

612. Straus, S.E. & McAlister, F.A. (2000) "Evidence-based medicine: a commentary on common criticisms". *CMAJ,* 163(7), 837-841.

613. Ioannidis, J.P., Mulrow, C.D. & Goodman, S.N. (2006) "Adverse events: the more you search, the more you find". *Annals of Internal Medicine,* 144(4), 298-300.

614. Tobin, M.J. (2008) "Counterpoint: Evidence-based medicine lacks a sound scientific base". *CHEST Journal,* 133(5), 1071-1074.

615. Sacks, H., Chalmers, T.C. & Smith, H. (1982) "Randomized versus historical controls for clinical trials". *Am J Med,* 72(2), 233-240.

616. Cruse, H., *et al.* (2002) "Quality and methods of developing practice guidelines". *BMC Health Services Research,* 2(1), 1.

617. Saint, S., *et al.* (2000) "Journal reading habits of internists". *J Gen Intern Med,* 15(12), 881-884.

618. Coomarasamy, A., Taylor, R. & Khan, K. (2003) "A systematic review of postgraduate teaching in evidence-based medicine and critical appraisal". *Med Teach,* 25(1), 77-81.

619. Coomarasamy, A. & Khan, K.S. (2004) "What is the evidence that postgraduate teaching in evidence based medicine changes anything? A systematic review". *BMJ,* 329(7473), 1017.

620. McColl, A*., et al.* (1998) "General practitioners' perceptions of the route to evidence based medicine: a questionnaire survey". *BMJ,* 316(7128), 361-365.

621. Krumholz, H.M*., et al.* (1996) "Aspirin for secondary prevention after acute myocardial infarction in the elderly: prescribed use and outcomes". *Annals of Internal Medicine,* 124(3), 292-298.

622. Krumholz, H.M*., et al.* (1998) "National use and effectiveness of β-blockers for the treatment of elderly patients after acute myocardial infarction: National Cooperative Cardiovascular Project". *JAMA,* 280(7), 623-629.

623. U.S. Department of Education: Institute of Education Studies: What Works Clearinghouse. (2015) *What Works Clearinghouse: what we do*. Available at: http://www.w-w-c.org/whatwedo/overview.html, accessed 17/07/15.

624. U.S. Department of Health and Human Services: Substance Abuse and Mental Health Services Administration (SAMHSA). (2013) *About NREPP*. National Registry of Evidence-Based Programs and Policies, available at: http://www.nrepp.samhsa.gov/AboutNREPP.aspx, accessed 17/07/15.

625. Means, S.N*., et al.* (2015) "Comparing rating paradigms for evidence-based program registers in behavioral health: Evidentiary criteria and implications for assessing programs". *Evaluation and program planning,* 48, 100-116.

626. Hennessy, K.D. & Green-Hennessy, S. (2014) "A Review of Mental Health Interventions in SAMHSA's National Registry of Evidence-Based Programs and Practices". *Psychiatric Services*.

627. GRADE Working Group. (2015) *What's new? GRADE workshops*. Available at: http://www.gradeworkinggroup.org/news.htm#workshops, accessed 20/07/15.