THE ROLE OF RCT EVIDENCE IN DRUG REIMBURSEMENT DECISION-MAKING PROCESSES:

AN INTERDISCIPLINARY INVESTIGATION

By

Merav Kaplan

ID 203867106

A Dissertation Submitted to the School of Public Policy and Government

The Hebrew University of Jerusalem

In Partial Fulfillment of the Requirements of the Degree of Master of Arts

Written under the supervision of:

Prof. Raanan Solizeanu-Kenan,
Federmann's School of Public Policy
and Government

Dr. Ittay Nissan-Rozen
Department of Philosophy,
and the PPE program

★★★

31 December , 2019

Abstract:

In recent years, healthcare systems have been experiencing increasing budgetary distress. In an era of cost containment, the need to construct rational processes for allocating scarce health resources is becoming more acute. The assessment of estimated effectiveness has a pivotal role in avoiding inefficiencies and optimizing resource allocation processes. Randomized Controlled Trials (RCTs) are considered the "gold standard" in clinical research to evaluate the effectiveness of medical products. As such, findings obtained from RCTs are perceived as vital sources of evidence for informing policymakers in the context of drug regulation in general and reimbursement decision-making in particular.

However, recently some have questioned the traditional approach of evaluating evidence for drug regulatory processes and the role assigned to RCT evidence within it. To encourage the rethinking of current practices, however, it is important to understand both the epistemic qualities of RCTs and the actual use of them in the context of clinical effectiveness evaluation processes. With this objective in mind, the aim of this thesis is to investigate the role of RCT evidence in drug reimbursement decision-making, while ultimately arguing that policymakers may benefit from the incorporation of Bayesian thinking into clinical evidence assessment processes. Recognizing the interdisciplinary nature of the issue discussed in this thesis, this work brings together various perspectives and combines several methodological approaches, using both philosophical inquiry and empirical analysis tools.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CADTH	Canadian Agency for Drugs and Technologies in Health
EBM	Evidence-Based Medicine
EMA	The Federal Joint Committee
FDA	Food and Drugs Administration
G-BA	The Federal Joint Committee
HAS	Haute Autorité de Santé
HTA	Health Technology Assessment
ICER	Incremental Cost Effectiveness Ratio
IQWiG	Institute for Quality and Efficiency in Health Care
NICE	National Institution for Clinical Excellence
QALY	Quality-Adjusted Life Year
RCT	Randomized Controlled Trial
RWE	Real-World Evidence
SMC	Scottish Medical Cortosiom
WHO	World Health Organization

Introduction

"The practice of medicine is an art, based on science. Medicine is a science of uncertainty and an art of probability." ~ Sir William Osler

In recent years, healthcare systems have been experiencing increasing budgetary distress. This challenges the ability of public health systems to provide their citizens with high quality, equitable, and affordable services. In an era of cost containment, the need to construct rational processes for allocating scarce health resources is becoming more acute. The assessment of estimated effectiveness has a pivotal role in avoiding inefficiencies and optimizing resource allocation processes.

Randomized Controlled Trials (RCTs) are considered the "gold standard" in clinical research for evaluating the effectiveness of medical products. As such, findings obtained from RCTs are perceived as vital sources of evidence for informing policymakers in the context of drug regulation in general and reimbursement decision-making in particular.

However, recently some have questioned the traditional approach of evaluating evidence for drug regulatory processes and the role assigned to RCT evidence within it. First, in some cases, the conducting of RCTs may be impossible or unethical. Those cases are becoming more common as the development of personalized treatments is gaining influence in medical practice. At the same time, advances in information technology are expanding the range of non-RCT evidence available for informing and supplementing medical research.

The trends described above call for a reconsideration of the role of RCT in drug evaluation processes. However, to stimulate a rethinking of the current practices it is important to understand the epistemic qualities of RCTs as well as the configurations of the actual use of this evidence in the context of clinical effectiveness evaluation processes. With this objective in mind, the aim of this thesis is to investigate the role of RCT evidence in drug reimbursement decision-making while ultimately arguing that policymakers may benefit from the incorporation of Bayesian thinking into clinical evidence assessment processes.

The evaluation of clinical evidence for drug reimbursement decisions lies at the intersection of multiple areas of knowledge. As such, a thorough investigation of this issue requires the incorporation of different points of view. Recognizing the interdisciplinary nature of the issue discussed in this thesis, several methodological tools were combined to bring together various perspectives; these refer to both normative and descriptive considerations. The normative evaluation includes philosophical inquiry into the epistemic qualities of RCT in general and

their role in policy decisions in particular. This normative investigation builds upon the literature in the fields of decision theory and the philosophy of science. At the descriptive level, methods and analytical tools from the field of public policy and health economics were applied to investigate the existing practices and their significance in the broader context of drug regulation. In light of the above, I hope this study will serve as an example of how exploring policy problems may stimulate philosophical inquiry that is more connected to real-world problems on the one hand, and that it will highlight the potential contribution that may arise from incorporating the philosophical perspective into thinking about policy issues on the other.

Outline of the Thesis:

The structure of this thesis is as follows: In the first chapter, we provide background information and set the normative foundation of the discussion of RCTs. Within this context, we briefly introduce the principles of the Evidence-Based Medicine (EBM) approach and the role of RCT evidence within it. Based on this review, we turn to a normative exploration of the epistemic benefits attributed to the RCT method in supporting clinical effectiveness claims while discussing its limitations.

The second chapter is descriptive and offers an examination of the role of RCT evidence as reflected in actual drug reimbursement decisions; this is done using a mixed-logit model while assessing the relationship between actual policy and the stated policy.

The third chapter is an integration of the findings of the two previous sections, harnessing the insights formulated in the first chapter to critically evaluate the findings of the second. This chapter begins by presenting an investigation of the characteristics of the decision problem of clinical effectiveness from a decision-theoretical perspective. Equipped with a better understanding of the challenges emerging from the decision problem at hand, we turn to examine the use of the Bayesian approach as a possible pathway for addressing these challenges while discussing both the opportunities and shortcomings associated with it.

I deeply thank my advisors, prof. Raanan Solizeanu-Kenan and Dr. Ittay Nissan-Rozen for their helpful and insightful comments and guidance, and for endless support and patience.

CHAPTER I

"That all science is a description and not explanation… For science cause is meaningless, The aim of science ceased to be the discovery of 'cause' and 'effect'; in order to predict future experience it seeks out the phenomena which are most highly correlated " ~ Karl Pearson

"No causality in, no causality out" ~ Nancy cartwright

This chapter provides a general background and sets the conceptual ground for this thesis. The discussion presented in this section begins by reviewing the development of the Evidence-Based Medicine (EBM) movement, and by outlining the principles underlying it. From this, we will turn to a normative evaluation of the role of Randomized Controlled Trials (RCT) in EBM, considering both the epistemic advantages and limitations attributed to it. This discussion will set the foundations for the descriptive investigation concerning the role of RCT evidence in drugs' reimbursement decision-making processes, as reviewed in chapter 2, as well as to the problem of weighting of evidence of different types, that is discussed in chapter 3.

1.1 Evidence-Based Medicine

1.1.1 Brief History

Evidence-based medicine (EBM) is considered one of the most important movements in the domain of medicine in the past century. The rationale underlying it is defined in the literature as: *"the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients*" (Sackett, 1996). As a social and ideological movement, EBM has been evolved in light of continuous dissatisfaction from the inference mechanisms and reasoning tools used by the medical community in the period preceding its emergence. Against this, EBM proponents have strived to rationalize medical decision-making processes, so as to ensure a more effective, efficient and "scientific" clinical practice and care. Looking back, it seems that this attempt has succeeded beyond expectations. Since its emergence, EBM has remarkably reconfigured our understanding of clinical knowledge formation processes and clinical research, reshaped the delivery of medical practice and profoundly influenced health-related policymaking, all in accordance with rigorous empirical standards.

While the EBM movement has historically evolved in the late twentieth century, many claims that its roots can be traced much earlier. The figure most associated with the evolvement of the paradigm is the Scottish physician Archibald Cochrane (1909-1988), who is sometimes

referred to as the *"Father of EBM*".[1] In his seminal 1972 monograph *"Effectiveness and Efficacy: random reflection on health services",* Cochrane strongly criticized the medical establishment of his time, and the British NHS in particular, in light of three parameters ('yardsticks'): effectiveness, efficacy, and equality. Within this manifest's framework, the main critique presented focused on medical education and decision-making processes, in relation to evidence assessment. In his pioneer discussion on effectiveness, Cochrane demonstrated, by citing various studies, how the common then-contemporary medical practice is based on approximations, unfounded hypotheses, and inferior evidence. Hence, Cochrane concluded that the medical research and practice of his time are inherently flawed and called for a reform in both the type of evidence used by the medical world and in the decision-making protocols. In particular, Cochrane highlighted the superiority of RCT evidence over other types of evidence and stressed the importance of enhancing the production and accessibility to evidence of this type. (Masic et al., 2008; Hill, 2000)

While Cochrane's ideas had become increasingly popular in the years that followed, it was not until the mid-1990s that the official institutionalization of EBM took place, and the term *"Evidence-Based Medicine"* was first coined. The event that marks the evolution of EBM as a standardized and influential movement in the scientific medical community was the 1992 gathering of the Evidence-Based Medicine Working Group. Following this meeting, the participants published a manifest (later known as the "JAMA paper") opening with the following statement: *"A new paradigm for medical practice is emerging. Evidence-based medicine de-emphasizes intuition, unsystematic clinical experience and pathophysiological rationale as sufficient grounds for clinical decision making and stresses the examination of evidence from clinical research."* (Evidence-based medicine working group 1992, p. 2420).

Over time, various institutions and structures have evolved in light of EBM's vision, intending to spread and implement its fundamentals as an integral part of the "normal" medical science. The most prominent among these organizations is the *'Cochrane Collaboration'*, founded in 1993 as a not-for-profit international organization aiming at creating and promoting the accessibility of up-to-date, systematic, high-quality evidence of various medical intervention and practices. Another manifestation is the formulation and publication of guidelines based on EBM principles. Those guidelines have become an integral part of the medical practice. Some

[1] While many Others attribute the development of EBM to David Sackett, therefore referring to Cochrane as the "Grandfather of EBM".

of those are published by government agencies and used as an influental, binding regulatory tool (Masic et al., 2008).

1.1.2 The underlying principles of EBM

Notwithstanding, to fully grasp the significance and transformative force of the EBM paradigm in the medical field, it is essential to be more precise and detailed about the nature of this approach. In particular, the characteristics, assumptions, and principles governing this paradigm must be examined. Such an investigation will allow for a critical assessment as to the degree of alignment between the objectives of EBM, the tools that it is applying, as well as the extent of its translation into actual policy.

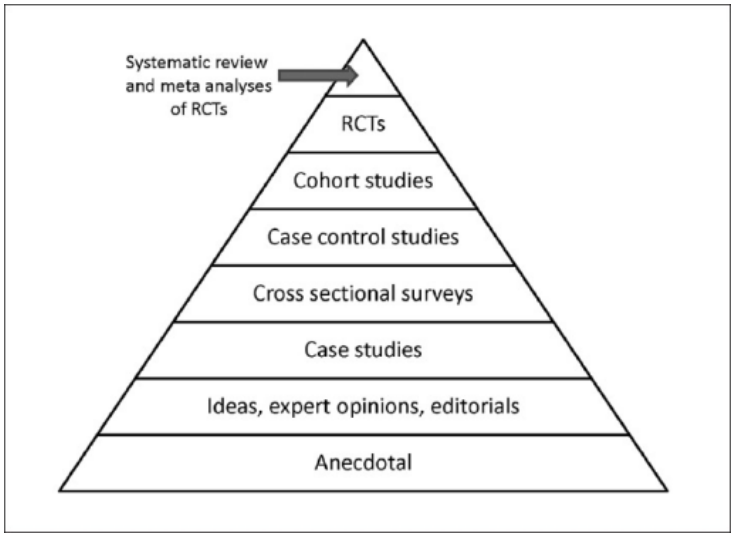
First, it is essential to note that the term *"Evidence-Based Medicine"* is used in the literature in two distinct meanings, at two different levels. The first usage pertains to the micro-level. That is, to the local, physician-patient relationship that addresses decision-making for the individual patient. In contrast, the second usage concerns the macro, policy-level regulatory decision-making processes. As will become evident below, it is only the second type of use that will be of interest in this thesis.

Second, it is crucial to understand in what sense the use of EBM has brought about a conceptual and practical change. On the face of it, the mere claim that medical decision making should be anchored in evidence sounds almost trivial and at the very least undisputable. In this regard, it is important to note that the medical establishment in the modern era, even preceding the development of the EBM approach, sought to rely on scientific evidence. However, the element that makes EBM informative and innovative - and at the same time controversial - related to the way that this approach understands the concept of evidence, as well as the relationship between different types of evidence in decision-making processes. As mentioned above, EBM refers to the use of the *'best evidence'* in clinical decision making. That is, its underlying assumption is that, by design, *"not all evidence is created equal",* and that some evidence is "better", in a strong sense, comparing to others.

This latter idea is manifested through the formulation of "evidence hierarchy", which ranks different types of evidence by their perceived "quality". This structure is sometimes referred to as "*The fundamental principle of EBM"* (Montori & Guyatt, 2008, p.10). The "evidence

hierarchy" is visually represented in the shape of a pyramid that places different types of evidence on a single-dimensional uniform scale, based on their epistemic merits. [2]

Figure 1.1: The pyramid of evidence (Yetley et al., 2016)



The pyramid of evidence is constructed and applied based on lexical reasoning, in the sense that "higher" evidence (evidence that is perceived as superior in terms of their quality) dominate evidence that is placed "lower" in the pyramid (that is, inferior quality evidence). It should be noted that the term "quality" of evidence in this regard refers mostly to the perceived resistance of the method of inference and study design to biases and confounders (Griffiths, 2011). The relationship between such biases and different types of evidence in the hierarchy of evidence is discussed below, in section 3.

At the very top of the traditional evidence-pyramid are filtered or synthesized evidence, extracted from meta-analysis and systematic reviews of RCTs.[3] Those are followed by evidence from unfiltered RCTs. At the next level there is non-RCT evidence from cohort studies, evidence from controlled observational trials, and evidence from uncontrolled observational trials, all of which conceived as inferior comparing to RCT evidence. At the bottom of the pyramid evidence from case reports, theoretical mechanistic studies, and expert-

[2] In recent years, few attempts were made to formulate new hierarchies of evidence, so that it would reflect the complexity and multi-level of consideration that needed to be considered once evaluating evidence in different contexts (e.g., Murad et al., 2016 ; Petrisor & M Bhandari, 2007, GRADE working group, 2013).

[3] Following Guyatt et al (1998) and Atkins et al., (2004) , meta-analyses and systematic reviews would be regarded in this work as a method for integrating evidence, that is – a method of using the hierarchy of evidence - rather than as evidence in themselves.

opinions are placed.[4] To avoid attacking a straw-man, it is important to note that the hierarchy of evidence does not serve as a sole determinant of decision-making guided by EBM principles, but nevertheless is "*an important component and may be the defining component*." (Vere, 2019) in evidence evaluation within it. It is also essential to highlight that there is no single agreed representation of the hierarchy of evidence in the literature (West el al., 2002 ; Goldenberg et al., 2009, Vere, 2018). Nevertheless, despite the wide variation observed, the ranking of RCT evidence at the top of the pyramid is a shared property of most representations. This makes RCT the conventional "*gold standard"* of medical scientific evidence (Vere, 2018).

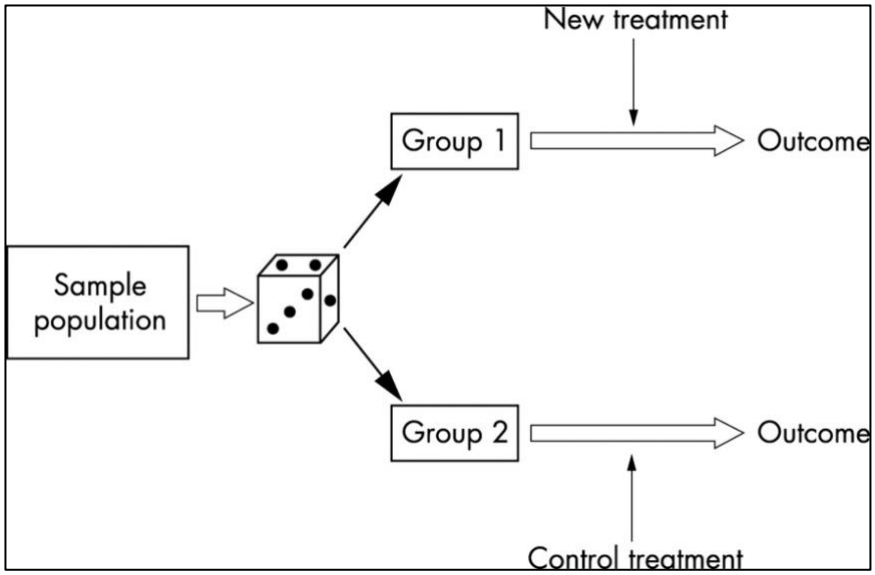
The remainder of this chapter is dedicated to the examination of the unique role of RCT evidence in the EBM framework. As a preliminary step, we provide a brief review of the basic element of RCT study design. Subsequently, we will turn to a philosophical analysis of the epistemic force attributed to this type of evidence. This normative discussion will, on the one hand, acknowledge the evident benefits of using RCTs as an inference tool for minimizing potential biases. However, at the same time, the limitations of RCT on various levels would be discussed. The ultimate conclusion to be derived from the discussion, therefore, will challenge the sharp dichotomy underlying the EBM's evidence hierarchy, and thereby will question the role assigned to RCT evidence vis-a-vis other evidence within it.

The process of conducting randomized controlled trials

In an RCT study design, a homogenous group of participants corresponding to the target population is being recruited, based on well-defined inclusion and exclusion criteria, which are set a-priori. Subsequently, the patients' population is randomly assigned into two mutually exclusive and exhaustive groups: the first group is the treatment group or the experiment group, receiving the treatment whose effect that is sought to be tested; the second is a control group, receiving an alternative conventional treatment, or placebo treatment (that is, a substance with no biochemical benefit). Alongside these, RCT study design often involves blinding - meaning that neither the study subjects nor the researchers (in Single-Blind study design) or both (in Double-Blind study design) are aware of the allocation.

4 The lexical use of the hierarchy of evidence, and the unique role of RCT in this framework can be exemplified by the following advice, appearing in a 2005 EBM handbook, discussing improving medical education thought EBM principles: *"If the study wasn't randomized, we'd suggest that you stop reading it and go on to the next article in your search. (Note: We can begin to rapidly critically appraise articles by scanning the abstract to determine if the study is randomized; if it isn't, we can bin it.) Only if you can't find any randomized trials should you go back to it". [Straus et al., 2005, p.118]*

Figure 1.2: Describing the process of randomized controlled trials (Kendall JM, 2003)



In the final part of the study, the analysis part, the effect of the medical treatment given to each group is compared. The averaged difference in outcomes between the two groups is considered the treatment effect. (Kabisch et al., 2011).

RCTs are usually conducted as phase III clinical trials. The traditional division of clinical trials consists of four phases (that in some instances may be combined): In phase I the tolerance, metabolism, and interaction of the drug under investigation are being assessed; phase II includes the examination of dose respondent and limited efficacy, based on biomarkers as outcomes and a small population of patients. The aim of phase III trials is to confirm clinical efficacy at a larger scale and to establish safety. Trials seeking to examine the drug in a broader (or alternatively special) populations, while detecting uncommon adverse events, are phase IV trials (Friedman et al., 2010).

Non-RCT study designs are titled '*observational studies'*. The set of observational studies is characterized by non-experimental study design, and include diverse inference methods, such as (but not limited to): *Historically controlled studies* (comparing present group receiving a treatment to a similar group receiving different treatment or no treatment); *Cohort studies* (longitudinal studies following group of patients receiving treatment over time. Cohort studies may be prospective, or retrospective), *Cross-sectional studies* (non-comparative trials studying a population at a given point in time). The data used in observational studies is extracted from various sources, ranging from data utilized from active observations, to registries or

administrative data. The latter sources of data sources are sometimes referred to in the literature as "Real-World Data" (RWD). [5]

After reviewing out the descriptive elements, we now turn to a more in-depth philosophical discussion of these biases, and accordingly of the benefits of using the RCT method as an inference tool for formulating clinical knowledge.

1.2 The Epistemic Power of RCT Evidence

1.2.1 Correlational and Mechanistic Knowledge

First, to set the foundations for the discussion in this section, we start by distinguishing two 'types' of medical knowledge: *mechanistic* knowledge on the one hand and *correlational* knowledge (or statistical knowledge) on the other hand. The first type of knowledge, the mechanistic knowledge, is based on a theoretical understanding of the mechanism that establishes the causal link between two variables. This type of knowledge, therefore, provides an answer to *"how?"* And *"Why?"* questions, and is usually derived from basic science research. Correlational knowledge, on the other hand, is based mostly on statistical reasoning. This model of knowledge is grounded in the use of a reliable mechanism, in the sense that it enables us to predict with a high probability the occurrence of some result in light of another phenomenon. While in the formation of correlational knowledge certain outcomes of an interaction between two variables can be justifiably expected, we remain ignorant as to the causal mechanism that establishes the relationship observed. Therefore, this kind of knowledge does not provide us with an *explanation*[6] (or at the very least, not with a *causal* explanation) as to the observed occurrences.

The variety of biological processes in the human body are interacting as part of a complex system. Acknowledging this complexity, clinicians and researchers traditionally refrain from claiming for comprehensive or even sufficient knowledge as to the mechanism underlying the operation of the human body. This perspective is a product of long historical experience, in both past and present, that indicates that often even when there was a strong sense of understanding of the causal mechanism underlying biological interaction, when this theoretical logic has been translated into actual medical applications, the observed results turned out to be inconsistent with our so-called well-established theory-based prediction.

5 for an elaborated discussion on the definition and use of RWD see Makady (2017).

6 For a detailed review of different accounts of causality and explanation in medicine see: Steel, 2011.

A recent example can be provided from the sphere of Alzheimer's Disease (AD) research. In recent years continuing progress in the field of neuroscience and biochemistry gave rise to the development of advanced potential technologies for treating AD, thereby giving patients around the world a cause for optimism. One of the most promising treatments that were developed was Aducanumab. This monoclonal antibody, presented by pharmaceutical company Biogen Inc., was based on an assumed understanding of the contribution of beta-Amyloid protein in forming plaques in the brain. This protein builds up at the early stages of the disease, so its activity was thought to compromises nerve cells, thus causing (or so it was thought) confusion and memory loss. However, after successful phase I and II trials, in March 2019, the company has announced halting two phase III trials testing Aducanumab, this due to insufficient observed effect (Selkoe, 2019). The case of Aducanumab serves as just one painful reminder of the unexpected gap between the theoretical understanding - that many times is considered well-established - and the disappointing results in its applications for actual treatment.[7]

In light of the skeptical attitude toward claims for a mechanistic understanding of biological processes, the knowledge that EBM seeks to establish is *correlational* rather than purely *mechanistic*. The epistemology on which it is grounded, therefore, is *reliabilist* rather than *evidential*. That is to say that it is anchored by the method used and not directly by the evidence in themselves. So, while the process of producing medical knowledge in the early 20th century was largely focused on basic science and mechanistic theoretical assumptions[8], the growth of the EBM movement can be seen as a return to the empiricist roots of formulating medical knowledge (Gifford, 2011). Using Newton's (2001) words: *"for all its rhetoric of novelty, Evidence-Based Medicine represents a counter-revolution of traditional empiricism, draped in modern clothes of statistics and multi-variate analysis"* [p. 314].

To avoid misunderstanding, however, we should stress that theoretical knowledge plays a vital role and is widely used in the clinical research process during the development of the

[7] This is not to say, however, that Amyloid is not causally connected to the development of AD. That is, those findings only indicates that the current mechanistic understanding of its role in the development in the complex structure of the development of the disease is inadequate. And indeed, in 22 October 2019, Biogen had announced that in light of findings from later sub-analysis of the data, it would seek FDA approval for Aducanumab.(Abbott, 2019). To this day, the question of whether the FDA should approve the drug is still hotly debated, and is considered by some *"the most important decision' the FDA will make in 2020".* (Lovelace, 2019).

[8] The most influential manifest contributing to the rise of the theoretic-based rationalist school in North America in the early 20th century was the Flexner report, written by Abraham Flexner and published in 1910 and calling for emphasizing subjects such as physiology, anatomy and microbiology in modern medical education.

intervention and the formulation of the hypothesis. In this sense, such knowledge may be confirmed or refuted in light of the experiment results. However, the inference process in itself, as it is perceived by EBM proponents, seeks to predict the effect of the treatment while minimizing the reliance on theoretical knowledge.

1.2.2 Correlational Knowledge and Confounding Effects

The acknowledgment of our lacking mechanistic understanding has not only established an empiricist shift in the medical realm, but also shaped the normative assessment as to the appropriateness of various tools for formulating medical correlational or statistical knowledge. Specifically, this attitude explains the salient dissatisfaction of the medical establishment with relying on observational studies as evidence for effectiveness evaluation. This results from the fact that, in the absence of an understanding of the mechanical process governing the relationship, a mere correlation between two variables - the treatment and the observable effect – cannot support the existence of a *causal* relation. This is because by merely observing an effect, the presence of a third, confounding variable invoking a similar outcome cannot be ruled out. That is, there may be an unknown variable that is unequally distributed between participants, and has an impact on the measured outcome but is nevertheless not part of the causal pathway between the independent and dependent variable (Jager et al., 2008).

Occasionally, the presence of an unknown confounding variable may mask an existing relationship between the two variables of interest. In those instances, the correlation between the two variables wouldn't be observed, even though they are indeed causally connected (thus producing false-negative results). Alternatively – in what is presumably a more common scenario - a third variable may bring about a false representation of the causal relationship between two variables, when such a relationship does not exist (thus producing false-positive results). In particular, there is a significant concern that the exposure to treatment is correlated in the first place with characteristics that affect the desired outcome. This worry is especially evident in health-related policy-level decisions.

When a theoretical background knowledge of the causal mechanism exists, confounding variables can be identified and controlled for (either in during the study design or in the statistical analysis stage). However, in the absence of a complete theoretical knowledge of a mechanism to identify the intervening variables that are otherwise unknown, it is not possible to justifiably draw a warrant causal link between the treatment and the effect.

1.2.3 The Epistemic Strength of RCTs

Considering the above distinction, we will now turn to characterize the epistemic advantage of using RCT as a tool for establishing reliable correlational knowledge. In particular, we would argue that RCT advocates perceive it as a tool that makes it possible to deduce a causal connection without requiring theoretical, mechanistic knowledge. That is, without turning to a type of knowledge that is not available to us. To substantiate this argument, we will use Nancy Cartwright's (2007a, 2011) distinction between two types of scientific experiments: experiments that are *"clinchers"* and experiments that are *"Vouchers".* In trials of the first type, the clinchers, once the experimental tool is correctly applied, and the assumptions underlying the experiment are met, then the causal relationship between the variables in question are deductively derived from the result obtained by the experiment. In contrast, in trials of the second type, the Vouchers, the result obtained supports the causal relationship between the variables probabilistically but is not guarantee it (Cartwright, 2007a).

Under this conceptualization, RCTs are identified as clinchers. However, Cartwright notes that RCT is not the only tool of this sort. In econometric analyses, for example, the causal relations can be derived deductively from the result obtained by the estimators as well, provided that the assumptions of the model are met. Nevertheless, RCTs still has a unique epistemic power vis-à-vis other clinchers, since the assumptions underlying its use are validated by the structure of the research design itself. That is, in contrast to econometric models, the assumptions underlying the use of RCT are satisfied *by construction*. To clarify this point, we shall turn to examine the function of the fundamental elements composing the structure of RCTs.

The most significant assumption for deriving a causal conclusion in RCT study design is given by the epistemic device of randomization. This assumption concerned the balanced distribution of confounders between the two groups – the control group and the treatment group. Given that the two groups come from the same distribution, the process of randomization renders that *in expectations* (we will come back to this feature later), both known and unknown confounders are equally distributed between the two groups.

As mentioned above, the failure to control for unknown confounders is the main barrier for tracing causation and establishing a prediction in accordance. Let t_1 be therapeutical treatement ans t_0 no treatment. A *correlational* relationship between treatment t_1 and an outcome O is obtained iff the probability of obtaining the outcome is greater given the treatment, compared to the probability of obtaining the outcome without treatment $p(O \mid t_1) > p(O \mid t_2)$.

Nevertheless, as long as potential confounding factors (Z) are not controlled for, a *causal* claim supporting a prediction cannot be derived from this mere correlation. While some of the intervening factors are known and observable and thus can be controlled for indeed, provided our incomplete mechanistic understanding other remain unknown.

Let $(z_j^k \in Z)$ be known confounders and $(z_i^{\sim k} \in Z)$ be unknown confonders. Randomization renders that *in expectation* the confounders are identically distributed between the treatment (t_1) and control group (t_0), so that $E\left(\sum_{i,j=1\dots n}\left(z_i^k, z_j^{\sim k}\right), t_o\right) = E\left(\sum_{j=1\dots n}\left(z_i^k, z_j^{\sim k}\right), t_1\right)$. Holding the confounding factors fixed in expectation between groups would allow for deriving the treatment effect, as no other explanation for the occurrence of the difference in outcome can be provided, apart from the intervention alone. That is, given that the treatment effect is a linear combination of the treatment effect (β_1) and the confounding effect (β_2), where treatment is a dummy variable:

$$\left(\overline{O_{t_1}} - \overline{O_{t_0}}\right) = [\beta_1 \cdot (1-0)] + \beta_2 \cdot \left(\sum_{i,j=1\dots n}\left(z_i^k, z_j^{\sim k}, t_1\right) - \sum_{i,j=1\dots n}\left(z_i^k, z_j^{\sim k}, t_0\right)\right)$$

Since the second argument equals zero we are left with an *unbiased* estimator of treatment effect alone : $\left(\overline{O_{t_1}} - \overline{O_{t_0}}\right) = \beta_1$.

The above structure suggests that randomization eliminates the effect of the intervening variables and thus allowing for the isolation and extraction of the causal relationship between the treatment and the observed outcome alone. The balancing assumption is crucial in this context since it allows us to assume that the dispersion of confounders is fixed in expectation across the groups being compared. This holds not only for the known confounders but also for the unknown confounders. Thus, the main advantage of the RCT method is that it does so without having to identify the unknown confounding variables. That is, without relying on the kind of knowledge that we do not have access to. [9]

All in all, the structure presented above establishes the distinctive epistemic power attributed to RCT. The application of this tool enables us to "bypass" (or so does it seems) the problem stemming from our insufficient mechanistic understanding while alleviating the concern of the influence of unknown confounding variables. This way, a causal conclusion can be established

[9] It should be noted that in a non-RCT study design different methods are used for dealing with confounder factors, among others: matching, restriction and stratification. However, these methods are usually considered inferior to randomization, as they all rely on substantial prior knowledge and are applicable only to known confounders. For further discussion see Jager et al., 2008.

independently of our (lacking) theoretical knowledge. That is, we can reliably predict the treatment effect without answering "how" and "why" questions.

So far, we referred to concern regarding bias stemming from lack of knowledge as to possible unknown confounders ("epistemic concerns"). However, RCT study design is designated to address another type of concerns. Those are associated with psychological effects that may distort the experimental results ("psychological concerns"). To counteract the effects of these biases, rigorous RCT study design involves, along with randomized allocation, some additional features: Placebo or active comparator is used to control for possible impact that awareness of treatment itself, regardless of its therapeutic effectiveness, may have on the outcome. In the same fashion, a masking strategy (blindness) is applied to reduce possible cognitive biases of either researchers or participants, resulting from knowledge about the allocation. While the two types of concern – the epistemic and the psychological – aimed at the same object of reducing bias, it is important to distinguish the two, since their roots, the type of tools used for addressing them, as well as the weight that is given to them in evaluating the quality of evidence, is distinctive.

To sum up this section, under relatively minimal assumptions RCT provides an unbiased estimator and therefore, it may be suggested, has rightfully taken its place as the "gold standard" of the evidence hierarchy. However, in recent years, a growing body of literature has been suggesting to widespread misconceptions as of the mechanism underlying RCT and hence pointing to an underestimation of its limitations (e.g., Cartwright & Deaton, 2016, 2018; Cartwright, 2011; Worrall, 2007; Worrall, 2011). The discussions in the literature can be classified into three categories: (1) Arguments concerning the use of RCT's as an inference mechanism establishing unbiased estimation of treatment effect; (2) Criticism of the interpretation of RCT's results and their generalizability for policy-purposes. (3) Descriptive discussions pointing to a discrepancy between the actual use of RCT in medical trials and the "ideal" RCT. In accordance with this classification, in the following section each of these discussions shall be reviewed and assessed.

1.3 The Limitations of the RCT Method

1.3.1 RCT as an Inference Mechanism:

Unbiasedness and precision

In discussing the benefits of RCT compared to other research methods, the unbiasedness of the estimate provided by RCT is often used as a central argument for its superiority. However, in

their seminal paper on RCT of 2016, Deaton and Cartwright emphasize the distinction between the unbiasedness of the estimator provided by the RCT design and the precision of such an estimate. While unbiasedness is obtained when the gap between the expected value and the true value (i.e., the bias) equals to zero, precision is related to the absence of random error. Quoting Deaton and Cartwright (2018): *"An archer who misses two feet to the right half of the time and two feet to the left half of the time is shooting unbiased arrows but* ***never*** *hits the target"* [p.5, emphasis in original]. Therefore, to be considered accurate, an estimator should be both unbiased and precise. As a matter of fact, the estimator provided by well-conducted RCT may meet the former condition by construction (at least for the sample population)[10], but in many cases, it may fail to meet the latter. [11]

The unbiasedness of the RCT method

On top of the above, some have questioned the extent to which RCT, as it is being applied in clinical trials, does indeed provide an unbiased estimate. In this discussion, the assumptions related to the allocation plays a crucial role.[12] As suggested in the previous section, randomization renders that the difference in other intervening factors equals zero *in expectation*. This means that in repeating randomization infinite times over the same sample, the estimated average treatment effect to be received from the trial would equal the "true" average treatment effect in the population (Deaton & Cartwright, 2016). In this sense, unbiasedness is a frequentist concept. In clinical trials, however, randomization is not repeated but conducted only once (more specifically, it cannot be repeated, at least not on the same

[10] The unbiasedness proof in RCT depends on the linearity of the operators. Thus, while sometimes we are interested in the variance or the median treatment effect, the only unbiased estimator that can be provided by this method is the mean (that is, other statistics can be assessed only under stronger assumptions).

[11] To stress the last point, a closer look at elementary statistics may be helpful. Indeed, unbiasedness is a desirable feature of an estimator, in the sense that when other things are equal, we would prefer an unbiased estimator to a biased one. Nevertheless, other features may as well be desirable, assuming that our ultimate goal is to provide a warrant statistical analysis in which the divergence from the 'true' parameter or interested is minimized. This divergence can be represented mathematically by the concept of mean squared error (MSE), calculated as the sum of the variance error and the squared bias: $MSE(\hat{\theta}) = var(\hat{\theta}) + \left(bias(\hat{\theta}, \theta)^2\right) + \sigma_e^2 = E_{\hat{\theta}}\left[(\hat{\theta} - E_{\hat{\theta}}[\hat{\theta}])^2\right] + \left(E_{\hat{\theta}}[\hat{\theta}] - \theta\right)^2 + \sigma_e^2$. This formula imply to a trade-off between the variance error (that is, error that is related to variability in the result of the model given a change in dataset) and error due to bias (i.e., expectation of error in the prediction of the model due to assumptions underlying it), in what is known in the literature as the *bias-variance dilemma*. Thus, under certain circumstances, we will be willing to trade some unbiasedness to reduce the variance. That is, we cannot prefer RCT to other methods just because the estimators provided by it is unbiased.

[12] In keeping with the most prominent literature in the field, the discussion here focuses on the element of randomization. However, attention was drawn also to other features of classical RCT. Howick (2008, 2011), for example, review the limitation of masking as well, but this element would not be discussed here.

population, or for a sufficient number of times). Unfortunately, in one-time randomization a balance between the two groups cannot be guaranteed.

As indicated by Worrall (2002, 2007, 2011), In many instances, after one-time randomization, the known covariates turn out to be unbalanced between the two groups once checked. For example, frequently, an unbalanced distribution between men and women in allocation is observed. In such instances, the researchers will initiate adjustments in the assignment, or re-randomized, so as to reach an appropriate balanced distribution of sex – a factor that is known to play a causal role in many cases - between the groups. An alternative possible solution would be to stratify the sample based on the known intervening variable proceeding the randomization process.[13] Indeed, many published RCT papers are summarizing the results of baselines tests of significance to determine covariates and review the method used for dealing with them.

Yet, if one-time randomization does not guarantee a balanced distribution of *known* observable confounders between the groups, it seems that we have no reason to believe that it produces an even distribution when it comes to *unknown* confounders. Unfortunately, when potential unknown confounders are involved it is not possible to amend the randomized allocation or to adjust the analysis for the sake of achieving a better balance. This implies that the assessment of the unbiasedness of the estimator provided by RCT depends on our background knowledge as to potential hidden covariates, and is not guaranteed by construction alone as was suggested above. Following this line of thought, Worrall (2007) suggest that the allocator should match the known confounders between the two groups in the first stage, and only then, on top, randomize the two groups into treatment and control. This highlights the fact that randomization is a second-best used only in light of insufficient knowledge. [14]

A famous example of the importance of prior background knowledge in assessing the precision and unbiasedness of the results extracted from RCT is a 2001 article that was published in the British Medical Journal. In this article, results from RCT consisting of 3,393 patients who hospitalized due to bloodstream infection several years before were presented. As described, the study population was randomly assigned into treatment and control groups, where the "treatment" received in this study was the prayer carried for the well-being of the participants in the treatment group by a stranger. As the research was conducted retrospectively – after 4-

13 The researcher may as well control for covariates in the statistical analysis stage. For an elaborate discussion of different methods used to adjust for covariates in the randomization process see: Dongsheng et al., 2000.

14 This point will be discussed more extensively in the discussion of Bayesian structures in Chapter 3.

10 years in time – the participants in both groups were obviously not aware of their placing, or of the treatment investigated in the study. Surprisingly enough, the results of the study indicated that the length of stay (p=0.01) and the recovery time (p=0.04) was significantly shorter in the treatment group comparing to the control group (Leibovici, 2001).

Although the experiment results were statistically significant and obtained by employing a "rigorous" study design, the scientific community, naturally, has not taken them seriously. The article has been published in the BMJ holiday issue and therefore has been dismissed as an innocent, amusing holiday-satire (which was indeed the case). However, given that the experiment was performed as described, it seems that despite its publication being satirical, a valuable lesson could still be learned from it.

In assessing the response to Leibovici's study, the observed effect has been attributed to a random error rather than to the treatment itself, and this even though the random assignment and control methods were adequately applied, and were mostly not questioned. In the absence of any available mechanistic scientific explanation that could have accounted for the interaction between remote retroactive intercessory prayer and physical indicators, the empirical observations alone were perceived insufficient for establishing the causal claim in question. This example may serve as an extreme case indeed, but it highlights nevertheless that not only that the interpretation of the results obtained from RCT is *in practice* highly dependent on our prior beliefs, but it is also that - epistemically speaking - it *should* be so. Those prior beliefs are grounded by *evidence* - though from a type and source other than RCT- that gives us justified reason to doubt the legitimacy of rejecting the null hypothesis in cases of this kind.

From the above, we can conclude that randomization is not a "bias-free" or "theory-free" mechanism. Evaluating its merit as a device providing an unbiased estimator depends on background knowledge as to potential confounding factors.

Nevertheless, to avoid misinterpretation of the course of the argument so far, it is important to stress that while claiming that RCT does not meet the expectations attributed to it by many, this is not to underestimate the evident benefits and importance of the RCT as an inference tool used for reducing *some* potential biases.

In particular, the primary benefit of RCT, and the of using randomized allocation process, is related to its resistance to *selection bias* within the sample: the randomized allocation does make it harder for the researcher (though not impossible) to manipulate the assignment for the sake of achieving favorable results. This aspect indeed justifies the preference of RCT over

non-randomized trials, especially when there is an established concern that the entity conducting the research has a strong incentive to manipulate the results, such as in the case of industrial pharmaceutical clinical research. It should be noted, however, that as such, randomization is providing a response to a concern anchored in behavioral aspect that is context-sensitive, rather than to the epistemic challenges rising from possible non-causal correlations .[15]

In addition, while RCT in some cases does not provide a biased estimate, other mechanisms probably will not provide better estimates in these terms. Considering the above, RCT is expected to be the least likely to yield biased results of the estimated sample treatment effect, even if this holds only under certain caveats.

1.3.2 The generalizability of RCT's findings

The elements concerning internal validity as discussed above may be necessary for the evaluation of the various inference methods, however, those do not appear to be sufficient for our purposes. First, as indicated previously, we may be interested not only in unbiasedness, but also in warrant and precision. Furthermore, ultimately, the objective of a clinical trial is to support decision-making with regard to the *entire* target population, not only the sample population. Hence, properly interpreting the results for extrapolation purposes is a component of particular importance. This concern is also known as the "Efficacy-Effectiveness" problem. The term "Efficacy" refers to the ascription of some specific effect in the *study population* - usually a sample of the target population - resulting from receiving the treatment. The term "Effectiveness", however, denotes the treatment effect in the *target population*.

Unfortunately, even if RCT is regarded as a reliable tool for supporting efficacy claims, when it comes to the justification of claims about effectiveness, it turns out to be inadequate. First, While the participants in RCTs are randomly partitioned into the treatment group and control groups, the study population is rarely selected randomly from the general target population. That is, RCT design grants *random allocation* of the sample but not *random sampling*. Indeed, natural experiments may allow for a random selection of the sample. In medical RCTs, however, participants are usually actively recruited and enrolled, and informed consent plays a key role. Thus, in most cases, it is unlikely to assume that the study population is similar to, or representative of, the target population. In this sense, RCT is also susceptible to selection bias. (Papineu, 1994 ; Cartwright & Deaton 2016 ; Worrall 2007).

[15] See more on sponsorship bias in section 4.3.

Another reason for assuming that the target population may differ from the sample population has to do with the study protocols and specifically with the exclusion criteria, which establishes selection bias. Setting stricter exclusion criteria may reduce "noise" and thus allow for better precision. Involving participants that suffer from comorbidities, for example, makes it harder to control for other intervening variables, and thus makes the tracing of causal relationships less likely. Therefore, in many cases, we can expect that those patients would not be recruited to trials.[16]

The existence of such recruitment and selection bias is supported by empirical literature, showing that participants in medical trials tend to be younger, healthier and are characterized by a stronger socio-economic background comparing to the general target population (Susukida et al., 2016). Ethnic and racial minorities tend to be consistently underrepresented in RCTs as well (FDA Snapshot, 2017 ; 2018). [17]

A possible solution for those potential bises would be to conduct a variety of experiments, each time on a different population, and apply the conclusions to the population on which the trial was conducted only. However, the use of this tool may be very costly. Moreover, it allows control only for the observable characteristics, so the concerns mentioned above as to the selection of participants remain intact. In other words, once again prior background knowledge, of the kind that is not provided from within the experiment itself, is needed.

Another concern is related to the short follow-up period of most RCT, which tends to be relatively short due to funding considerations, and therefore cannot determine the long-term effect of the treatment. Finally, we can expect a potential difference in the behavior of patients when they are closely observed and monitored during the experiment, compared to the messy

[16] For example, recent analysis RCTs of immunotherapy registration for metastatic melanoma patients shows that 59% of the target population did not meet the exclusion criteria (Donia et al., 2017).

[17] In 1977 the FDA banned women from participating in most clinical trials. However, in 1993 inclusion of women became obligatory. The current regulation require that the exclusion and inclusion criteria, as well as the distribution of population characteristics in the experiment would be presented transparently. Alongside this, there are various initiatives that seek to emphasize this problem in order to encourage better practice. See, for example, the FDA's project on the diversity of participants in drug trials: https://www.fda.gov/drugs/drug-approvals-and-databases/drug-trials-snapshots. Nevertheless, unfortunately, while many legal and regulative tools were applied to increase representation of underrepresented populations in the past years, those are not always effective. For example, despite the binding regulation as to the representation of women in clinical trials, recent analysis shows that while women makes 51% of cardiovascular patients, they still amount only to 39% of participants in clinical trials in this field (Feldman et al., 2019).

real-world setting which involves behavioral and environmental considerations such as lack of responsiveness, drug interactions, etc. [18]

Before proceeding, it should be noted that the problem of generalizability highlights a broader and more substantial lesson regarding the relationship between RCT and other types of evidence. On the face of it, the discussion over the superiority of RCT over different types of evidence seems to be a matter of *empirical* dispute. All that is required in order for it to be settled, so one might think, is to simply reject the null hypnosis suggesting an equal degree of compatibility between the expected results as obtained from RCT and non-RCT studies and actually observed "real world" outcomes. This, prima facie, can be done by applying prospective analysis tools. However, at a closer look, it turns out not to be as simple as it seems, and this is due to several reasons. The first point to be mentioned is that in fact, while there are some famous examples of disparities between the results obtain from RCT and non-RCT data, [19]the findings from most meta-analysis comparing results from (well-conducted) observational studies and RCTs indicated no detectable systematic difference between the two (Vandenbroucke, 2011). Second, studies found that replications of RCTs investigating the same population and treatment produce many times inconsistent results (Zeilstra et al., 2017).

Another essential issue related to the interpretation of discrepancies, once observed. If the observed "real-world" results are incompatible with those of the controlled experiment, it can always be suggested that there is an effect of intervening confounding variables. That is, one can propose that the expected result was not obtained in practice not because the results were misleading, but because the causal treatment effect - which was successfully inferred from the RCT - is concealed or disrupted by confounding variables that cannot be controlled for in real-world setting. After all, this is the reason for utilizing the RCT study design in the first place. Still, this reasoning, unfortunately, establishes a circular argument that does not meet Popper's refutation test, as the results are judged by assuming the correctness of the experiment.

[18] One way of addressing this issue is by conducting "Pragmatic trials". That is, investigating the effectiveness of treatment under a broader "everyday" setting. However, these kinds of trials usually rely (at least in part) on non-RCT study design. See: Patsopoulos, 2011.

[19] The most famous case is the case of hormonal repalacement therapy. In the early 1990's, observational studies have suggested that hormonal replacement therapy is associated with the reduction of 30%-50% in the development of coronary artery disease in post-menstural women. However, the finding from a first large RCT investigating hormonal replacement therapy, presented in 1998, indicated showed no effect. The debate over the "real" effect of hormonal replacement therapy continues to this day.

From the above, it turns out that the problem of generalizability cannot be settled empirically, and therefore such reasoning cannot be used to resolve the disagreement as to the relationship between the types of evidence.

The issue of generalizability poses a significant challenge to the use of the results obtains by RCT for policy-making. This raises primary concern as - using Cartwright's words - the question we are interested in is not only "does it work (under specific circumstances)?", but mainly "will it work *for us*?". While RCTs (under some assumptions) can provide a reliable answer to the former question, nothing in the construction of the experiment itself allows for the answering of the latter. Therefore, the judgment as to the extrapolation of RCT's findings to the target population necessarily depends (at least partially) on the knowledge extracted from other, non-controlled, evidence, for providing an indication as to the expected reproducibility in real-world settings.

1.3.3 Ideal RCTs vs. Actual RCTs

The last group of argumentations questioning the attributing a superiority to RCT vis-à-vis other types of evidence is more contingent in nature. This criticism does not address the essential characteristics of RCT, but the way it is actually performed in practice. In particular, it pertains to problems that arise from technical constraints, or from manipulations shaped by the interests of stakeholders involved in the conduction of the clinical trials. The result is that many of the actual RCTs are not the same as the ideal RCT outlined in the literature, and hence their preference over different types of evidence (which may be biased as well) is at least questionable.

Today, the pharmaceutical industry is the primary funder of clinical trials. Over the years, many studies have shown that trials sponsored by pharmaceutical companies systematically yield favorable results comparing to trials that are financed by non-commercial bodies (Ahn et al., 2017; Als-Nielson, 2003; Bero et al., 2007; Wareham, 2017). Those results hold both in terms of the likelihood of achieving results that are statistically significant (Lundth et al. 2017), in terms of presenting positive results (Bhandari et al., 2004), as well as in the magnitude of the effects estimated (Lundth et al. 2017).

Moreover, studies have shown that RCTs are also susceptible to "sponsorship bias" (Delgado & Delgado, 2017). Recent meta-analysis calculated Odd-ratios for favorable results in 509 RCTs across all medical fields found higher OR for favorable outcomes in trials funded by for-profit entities. By and large, trials conducted by the pharmaceutical industry tend to be of a

higher "quality" in terms of the hierarchy of evidence comparing to trials sponsored by non-commercial entities (Delgado & Delgado, 2017). This, in part, is a result of the high cost involved in conducting an RCT.

It should be noted that while some of the problems mentioned in the previous section, such as issues of internal validity, can be addressed by enlarging sample size, this may come at the price of increased risk of potential sponsorship bias. Many times, a trade-off exists between regulatory requirements regarding minimal sample size and power, leading to higher costs, on the one hand, and the independence of funding on the other. The stricter the requirement becomes, the harder it is to fund a trial without an industrial sponsorship (Doucet & Sismondo, 2008).[20]

Moreover, RCTs in practice are many times open-labeled, due to unavoidable technical issues. Even in cases were the blindness method is applied, maintaining an effective ignorance can turn out to be very challenging in the actual administration of RCTs. In many instances, the occurrence of common, even minor, side effects can give researchers or participants reason to believe that one group has received active treatment, thus causing an effect that enhances their response (Berna et al., 2017). Studies have shown that correct guessing of the allocation, and in some cases merely beliefs and expectations, can have a substantial influence on the results and overstate the treatment effect (Bang, 2016).

Before concluding, two important remarks on the feasibility of RCT should be noted. First, in some cases, the conduction of RCT may violate ethical standards or norms. Worrall (2007) uses the case of ECMO to illustrate the problematic ethical consideration involved in the conduction of RCT in some cases. In the late 1970s, a new technology for treating a condition of persistent pulmonary hypertension (PPHS), a congenital pulmonary disease whose mortality rate in infants was over 80% at the time. Early uses of this technology, called *extracorporeal membraneous Oxygenation* (ECMO), resulting in mortality rates of less than 20%. These results, obtained from an observational historical control analysis, while not serving as conclusive evidence, gave the researchers a strong reason to believe that ECMO is probably effective, with a high potential of preventing unnecessary mortality.

20 In response to the finding presented above since 2008, a deliberate attempt was taken for minimizing sponsorship bias. This was done, among other things, through to the advancement of transparency to the public, and pre-specification of end-points. However, Meta-analysis that was conducted by Delgado & Delgado in recent years (2017) has yielded similar results as those of previous studies. This observation indicates that unfortunately, those steps are not as effective as one would expect.

However, in light of the conventional standards of the medical community, the researcher felt "compelled" to conduct RCT to confirm the results. The use of classical "pure" randomization would have required them to knowingly assign infants to the control group, anticipating that 80% of them would die, while at the same time believing that a treatment that may prevent most of the deaths is available. Therefore, eventually a "randomized plays the winner" method was applied. All in all, from a total of 12 patients, 11 patients were assigned to the ECMO and survived. One patient (the first patient) was assigned to the conventional treatment and died. (Worrall, 2007).

This case indicates that RCTs are not produced in a vacuum. Prior knowledge is not only essential for the conduction of RCT but it also sometimes set deontological ethical restrictions on the conduction of scientific trials. In such cases, when there is a strong indication of the effectiveness of alternative treatment, the placement of patients into the control is exposing them to unnecessary risk thus coming in conflict with the moral duty of treating the interest of the patient as paramount.

The problem presented here gave rise to what is known in the medical and bioethical literature as the 'equipoise' principle. According to this principle, conducting an RCT would be considered unethical, unless the two treatments are regarded as *"equal bet in prospect"* (Fried, 1974), or at least in case of "*honest disagreement*" between professionals (Pimple, 2017). While the requirement for equipoise has was criticized by many (e.g., Miller & Brody, 2003; Veatch, 2007), a weaker version of this principle is still accepted by most as valid.

The second issue to be discussed here is related to the technical feasibility of conducting RCTs. In many instances, the conduction of RCT may not be feasible at all due to the small size of the target population. This problem is particularity evident in personalized genetic-based treatments. In such cases, enlarging the sample size by the expansion of the inclusion criteria may allow for RCT design, but at the same time may result in ethical problems, of the sort mentioned above (that is, including patients that are unlikely to response or patient that may be subjected to considerable risk if participated).[21]

[21] This worry is especially evident in the case of personalized molecular-based treatment, where the likely to benefit only small fraction of participants. Selecting only the patient with the relevant bio-markers is not always possible in such cases, and even when it is, it will result in small sample size mostly that would not meet the minimal evidential standard for the conduction of an RCT (Nardini, 2014).

Conclusion:

In conclusion, while the vital contribution of the EBM movement to the development of clinical research and the improvement of medical practice in the past century cannot be denied, some of the assumptions underlying it are controversial, in particular with regard to the use of the hierarchy of evidence and the unique status that is granted to RCT within it.

Considering all the above, we can conclude that RCT does have unique epistemic power. However, its merits are many times overstated by EBM proponents. RCT's experiment design may contribute to reducing potential biases, but only in a context-dependent manner that is not granted by construction. Both the judgment regarding the internal validity of RCT as well as the evaluation of the appropriateness of its translation into practical uses requires additional knowledge. Such knowledge cannot be derived from the RCT itself and entail reliance on non-experimental sources of evidence that are considered "inferior" in terms of the classical EBM's hierarchy of evidence. For example, theoretical knowledge, experts-opinion and "real world" observational data. Furthermore, RCT is not always available due to technical and ethical issues. In these instances, effectiveness assessment can rely only on non-RCT data.

These insights suggest that different types of evidence cannot be perceived in a discrete manner as they are represented in the evidence hierarchy. As a result, it is hard to justify and maintain the lexical use of evidence in decision-making processes as dictated by the logic of this hierarchy.

Once the boundaries between different types of evidence are blurred, new issues and challenges that ought to be addressed rises. In particular, methods for the weighting and synthesizing of different types of evidence need to be established, on both theoretical and technical level. In this sense, it seems that the above discussion calls for shifting the course of thinking from a traditional hierarchy of evidence into a *web of evidence* of various sorts, based on degrees of warrant and coherence.[22] This challenge has practical implications in the context of policy, especially with regard to decision-making processes that are based on the principles of EBM. Thus, before turning to formulate the decision problem regarding evidence assessment more explicitly and clearly, we will review the existing policy regarding the incorporation of non-RCT evidence into drug reimbursement decision-making processes.

---

[22] This line of thought is compatible with the idea of "epistemic" theory of causality rather than "difference making" or "mechanistic" accounts of cuaslity. See: Russo & Williamson, 2011.

CHAPTER II

"There is an eternal dispute between those who imagine the world to suit their policy, and those who correct their policy to suit the realities of the world" ~ Sorel

Introduction

In most public healthcare systems, drug reimbursement decision-making processes are supported by the health technology assessment (HTA). In this chapter, we briefly review the fundamentals of the HTA process and discuss the existing literature on the utilization of non-RCT evidence. Then, we explore the role of RCT evidence in drug reimbursement decisions in five different countries using qualitative and quantitative analysis tools. The main objective of this chapter is to investigate the extent to which the type of available evidence is predictive of the evaluation of a drug in the HTA process and, subsequently, the formulation of reimbursement recommendations of the drugs being appraisal. Another objective is to assess the compatibility between actual and stated policies regarding this topic. The findings of the descriptive analysis described in this chapter, combined with the theoretical discussion from the first chapter, will improve the understanding of the role of RCT evidence in drug reimbursement decision-making processes and set the foundations for exploration of possible ways to address evidential uncertainty during assessment of drugs' effectiveness, which is the subject of the following chapter.

1.2 Policy Review – Health Technology Assessment

The healthcare market is characterized by various market failures, the most notable of which are imperfect information and information asymmetry between providers and consumers (Arrow, 1963). Endeavors to minimize the inefficiencies resulting from these market failures, along with distributional justice considerations, have led to the extensive involvement of the government in the healthcare market. In many developed countries, the government not only regulates the healthcare market but also is directly or indirectly involved in financing and providing healthcare services and health insurance. Such systems are usually referred to as public medical care systems, and they generally aim to provide all individuals and populations within their jurisdiction with adequate access to health resources and medical services.

However, in recent years, the task of providing adequate, accessible healthcare has become more challenging. In the past decade, healthcare systems in developed countries have faced an increase in the cost of health inputs, resulting in part from the continual rise in drug prices (WHO, 2018). In addition, demographic changes associated with aging populations and

unhealthy behavioral trends due to contemporary lifestyles are leading to substantially increased demand for healthcare services (Dall et al., 2013).

As the cost of healthcare inputs and the extent to which health needs remain unmet increase, so do the resource limitations facing public healthcare systems. In this context, the need to prioritize resource allocation and optimize the delivery of care is of utmost importance. Faced with these challenges, various healthcare systems are seeking to formulate sophisticated models for prioritization and decision-making processes to enable more efficient and equitable distribution of scarce health resources (Nagel & Lauerer, 2015).

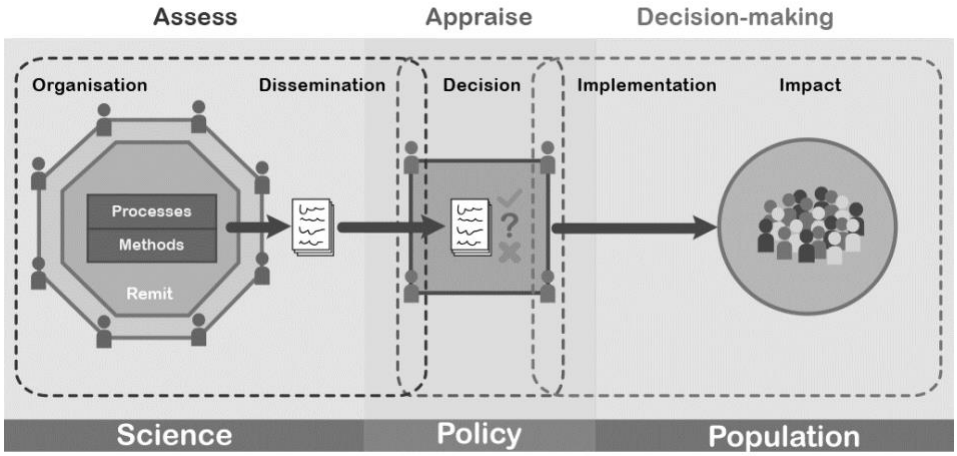
Following the discussion in the previous chapter, the measures taken to rationalize prioritization processes in advanced public health systems involve systematic collection and analysis of scientific evidence, as dictated by the principles of EBM. In most developed public healthcare systems, the setting of priorities regarding public coverage of medical products and services is supported by the "Health Technology Appraisal" (HTA) mechanism, an interdisciplinary method of policy analysis defined as “the bridge between evidence and policymaking” (Battista & Hodge, 1999). Consistent with this definition, the ultimate aim of the HTA process is to inform policymakers involved in decision-making processes concerning public funding by assessing the added value of a candidate technology compared to existing technologies. It should be noted that the term “value” is used in health-related HTA in a broad sense, incorporating a variety of clinical, economic, social, and ethical considerations (Detiček et al., 2018).

While the HTA is comprehensive and covers a wide range of attributes, the discussion in this thesis will be limited to the segment of the HTA process that concerns evaluation of the evidence supporting the clinical effectiveness of drugs for reimbursement recommendation purposes. The assessment of relative clinical effectiveness is a fundamental and central component of the HTA process, serving as a significant predictor of the nature of recommendation to be provided by the agency performing the HTA. A drug that failed to convincingly demonstrate an adequate level of effectiveness is not likely to be recommended for public funding (Sorenson & Chalkidou, 2012). The formal clinical effectiveness assessment procedure features less variability across countries than other components of the HTA process, and therefore it is a more appropriate subject for the comparative analysis in this thesis. In addition, while the term “health technology” covers diverse technologies ranging from prevention programs to diagnostic tests, devices, and procedures, we will focus only on therapeutic drugs.

Fundamentals of the HTA Process

In public health systems, the HTA process is performed by one or more national governmental agencies. The reimbursement decision-making process consists of three main phases: (1) assessment, (2) appraisal, and (3) final decision-making. The first two elements are usually conducted by the HTA agency (or agencies), whereas the third is usually performed by another body based on the recommendations provided by the HTA agency as well as a broader set of social and normative considerations (HUPATI, 2015).

Figure 2.1: HTA process, three-phase model (EUPATI, 2015)



The HTA process starts with identification of the topic to be assessed. Some countries, such as France and Poland, carry out the HTA process for *all* new drugs that have EMA authorization, while others, such as England and Canada, only assess drugs that meet pre-defined criteria. For example, for a technology to be evaluated by the English HTA agency, NICE, it should meets some criteria as a precondition. For instance, it should already demonstrate a significant benefit and be supported by appropriate evidence (NICE, 2014).[23]

The assessment phase involves collection and retrieval of relevant evidence. In most cases, the evidence supporting the evaluation is submitted in the reimbursement application dossier by the market authorization holder. In some (uncommon) instances, the HTA agency may initiate an independent primary data collection process (EUPATI, 2015).

The body of evidence collected for the HTA process is appraised by an independent expert committee. In most agencies, the evaluation process is comparative, taking the form of a relative effectiveness assessment (REA). The aim of this type of assessment is to evaluate the

[23] In light of criticism regarding the lack of transparency in the pre-selection process, NICE has attempted to make the prioritization process and methods more explicit.

additional clinical benefits of the pharmaceutical under investigation in comparison to the prevailing standard of care. In addition, most agencies perform an economic evaluation based on budget impact or an incremental cost-effectiveness analysis of various forms (ICER).[24] In the final phase, the findings are summarized and a reimbursement recommendation is formulated for submission to the authorized decision-making body.

The essential framework of the HTA process is outlined above. Nevertheless, it must be emphasized that the process is not uniform across countries; in practice, the scope of the process and method for assessment differ in different contexts. Some agencies perform the full HTA process, which covers the technical features, safety, clinical efficiency and effectiveness, budget impact or ICER, and social, legal, and ethical impact of technologies. Others default to rapid, partial assessment focusing on clinical effectiveness and safety and evaluate other value components only in exceptional cases when the need arises (Anglelis et al., 2018; Oortwijn, 2018). The weight given to economic and social variables, as well as the methods used for estimating them, also vary between different health systems (Kristensen, 2017). This variation is a result of the legal, institutional, and regulatory environments within which the process is being performed, reimbursement arrangements, dominant values and norms, and policy legacy (Anglelis et al., 2018; Akehurst et al., 2017; Allen et al., 2017). For a detailed comparison of different HTA processes in the five countries under examination in this thesis, see *Appendix A.*[25]

The considerable heterogeneity that characterizes HTA processes in different countries makes international comparison of HTA processes and resolutions especially challenging (Fischer, 2012). However, it is agreed that some key challenges and opportunities are shared by all HTA agencies, regardless of the specific regulatory context within which they operate. One such challenge is establishing mechanisms for incorporating and integrating different types of clinical evidence. Overcoming this challenge is essential to provide more rational and justified recommendations regarding drug reimbursement to inform decision-making.

[24] However, as mentioned above, evaluation of economic factors is beyond the scope of this work.

[25] Due to the high heterogeneity in HTA processes across countries, recognition of the value of collaboration between different medical regulatory agencies has grown over the past decade. In addition, Article 15 of the EU's directive for cross-border healthcare, which states that "the Union shall support and facilitate cooperation between national authorities or bodies responsible for HTA"[25] (2013/329/EU), gave rise to various institutionalized initiatives for standardizing the HTA process. The most prominent of these is the EUnetHTA project, which aimed to improve the transference of knowledge and information between HTA agencies across Europe (Allen, 2017). However, while efforts have been made to standardize processes, the ultimate goal is still far out of reach and HTA processes are still characterized by a high level of fragmentation and variation (Allen, 2017).

The need for an evidence integration mechanism has become especially relevant recently, as treatments have become more personalized and the technological development of effective treatments is occurring more rapidly. In this context, carrying out extensive RCTs to inform decision making in a timely manner is a significant challenge. , advancement in information technologies makes previously unattainable "real-world" observational data available to utilize. These trends raise the need for rethinking the existing evidence assessment mechanisms.

1.3 Literature Review

Recognizing the opportunities and challenges involved in expanding the scope of evidence used to support clinical efficacy assessments, recent studies have expressed interest in analyzing HTA recommendations in general and the role of different types of evidence in this process in particular. This literature review includes both qualitative analyses of experts' attitudes toward non-RCT evidence and quantitative estimations of the extent to which these evidence are utilized.

First, various qualitative studies have identified increasing recognition of the value of non-RCT data for informing drug coverage decisions among policymakers in Europe. However, while there is more receptiveness to the idea of incorporating such evidence into evaluation processes, decision-makers have expressed concerns regarding possible biases and stressed the need for new methodological frameworks and skills to overcome this problem (Angelis et al., 2017; Gill et al., 2016; Kamphuis et al., 2018). The increased interest in the use of non-RCT data has also resulted in a growing number of studies that formulate methodological guidelines and initiatives for investigating the best practices regarding the use of non-RCT data (sometimes referred to as real-world evidence) by various regulatory and NGO bodies. such as the US Food and Drug Administration (FDA, 2018, 2019), European Medical Agency (EMA, 2018), ISPOR special task force (2017), and HTAi Global Policy Forum (2019).

While the acknowledgment of non-RCT data as valuable is gradually increasing, and so is the recognition of the limitations associated with the traditional role of RCT, recent quantitative studies have indicated that the use of non-RCT data for informing actual policy recommendations remains limited nevertheless.

Griffiths et al. (2017) assessed the role of noncomparative evidence (i.e., that used in studies that do not present the results of another treatment) in HTA decision making. The findings, based on an analysis of 549 reports published from 2010–2015, suggested that reliance on

noncomparative evidence alone is limited; only 4–6% of recommendations were based solely upon this data. Noncomparative data were considered in analyses more often, but with higher variance; 12%, 13% and 38% of IQWiG (Germany), CDATH (Canada) and NICE (England) appraisals, respectively, referenced non-comparative data in some part of the HTA process (i.e., either clinical or economic assessment). Finally, in total, 13% of CADTH recommendations (n = 2), 40% of NICE recommendations (n = 2), and no IQWIG recommendations based on noncomparative evidence alone were positive.

Similar results were presented by Vreman et al. (2018) and Makady et al. (2018). The former study investigated reimbursement recommendations for conditionally approved drugs with non-comparative evidence, finding that only 12 of the 62 investigated drugs (13%) were given positive recommendations when no RCT data was available (Vreman et al., 2018). The latter study investigated melanoma treatment appraisals in five European countries that incorporated non-RCT data in effectiveness evaluations, finding such data in 22% of NICE (England) evaluations, 9% of HAS (France) evaluations, and no IQWiG (Germany) or ZIN (the Netherland) evaluations. Although there is a growing interest in the incorporation of non-RCT data among stakeholders, the authors did not find a substantial change in the use of this data for melanoma drug appraisals over the years (Makady et al., 2018).

While the extent of use of non-RCT data for reimbursement decisions has been investigated in the literature, previous studies based their analyses on investigations of final HTA reports. This methodology is problematic for two main reasons. First, the HTA drug evaluation process is confined to drugs that have received market approval. A drug that is not approved by either the EMA or FDA cannot be a candidate for public coverage. Therefore, to assess the attitude towards RCT evidence within the context of public reimbursement decisions, one should first consider the proportion of drugs with non-RCT evidence that received market approval. If the rate of approved drugs for which there is no available RCT data is particularly low, it is expected that the rate of public reimbursement of such drugs would be low as well, even in an extreme scenario in which all drugs without RCT evidence would be given a positive recommendation. Second, the existing research did not consider the selection processes that precede the HTA. Some agencies, such as NICE, conduct a pre-screening process to select the technologies to be evaluated based on general prioritization criteria (e.g., availability of evidence, population size, disease severity). Consequently, it is possible that market-approved drugs with no supporting RCT at the time of market approval will not be evaluated in the first place, or will be assessed under exceptional circumstances , of the type that is correlated with

favorable (or unfavorable) recommendations. In such a case, studying only the final HTA reports would produce biased results.

In the following section, we investigate the role of RCT evidence in drug reimbursement decision-making processes. To address the above concerns, we considered the effect of evidence type on the probability that a drug will undergo the HTA process and the impact of the type of recommendation provided after assessment. Our research questions are as follows:

(1) Does evidence type predictive of the evaluation status of approved drugs undergoing the HTA process?
(2) Does evidence type predictive of the final reimbursement recommendation for a drug after the HTA?
(3) Do the stated attitudes of various HTA agencies regarding which evidence types can be used to support effectiveness claims align with their actual policies?

Building upon the theoretical discussion of the benefits and limitations of the RCT method in the previous chapter and upon the findings from the existing literature, which are reviewed above, we hypothesize that drugs with RCT evidence are more likely to be evaluated with the HTA processes and granted more favorable recommendations after the HTA assessment compared to drugs whose effectiveness is not supported by RCT evidence. We also hypothesize that regulatory bodies will manifest concern about the use of non-RCT evidence in clinical effectiveness assessment processes and that this concern is manifested in the actual policies of the investigated bodies.

1.4 General Method

A mixed-methods approach was used to answer the research questions. In the preliminary stage, qualitative document analysis was performed to examine HTA agencies' stated positions regarding the role of various types of evidence in clinical effectiveness assessment process. In the second stage, using quantitative analysis tools, we retrospectively analyzed the reimbursement recommendations of HTA agencies for all drugs that received market approval by either the FDA or EMA from 2015–2018. The use of mixed methods is essential for tracing possible disparities between the stated positions of the agencies and their policies in practice. In the following sections, we will review the qualitative part of the analysis and then provide a detailed review of the quantitative part.

Both the qualitative and quantitative analyses were based on investigation of five HTA agencies, including four from Europe—IQWiG (Germany), NICE (England), SMC (Scotland),

and HAS (France)—and one non-European agency—CADTH (Canada). These bodies were selected based on the following criteria: (1) the agency is a governmental entity operating within a public medical care system; (2) the jurisdiction of the institution is authorized by law and the recommendations it provides have a direct influence on the public coverage status of the drug under investigation; and (3) both the recommendations and reports formulated by the agency are publicly available in English and provide an overview of the studies evaluated during the assessment process.

Table 2.1. Agencies selected for investigation

Abbreviation	Full Name	Country
HAS	Haute Autorité de Santé	France
NICE	National Institute for Health and Care Excellence	England
IQWiG	Institute for Quality and Efficiency in Health Care	Germany
CADTH	Canadian Agency for Drugs and Technologies in Health	Canada
SMC	Scottish Medicines Consortium	Scotland

2.3.1 Qualitative Document Analysis

2.3.1.1 Data Collection

The documents used for this part of the study are the latest official methodological guidelines published byvarious the five HTA agencies. In this context, we referred to only reimbursement decisions and did not include guidelines regarding managed access or conditional programs. Relevant information was extracted from the relevant sections dedicated to effectiveness assessment.

2.3.1.2 Results

Table 2.2. summarizes the types of effectiveness evidence that are considered admissible by each agency. The document analysis indicated that almost all HTA agencies apply some form of hierarchy of evidence, either explicitly or implicitly. Moreover, all agencies emphasize the unique epistemic strength of RCT evidence. However, the willingness to accept non-RCT evidence varies between different agencies. In particular, NICE and CADTH express the most pluralistic position with regard to the types of evidence that are acceptable, while IQWIG has the most conservative position.[26] The full document analysis can be found in *Appendix B.*

[26] We must note that, unlike the other agencies, SMC seems to feature a hierarchy based on the sub-characteristics of the RCT study design (i.e., one that uses an active comparator). Interestingly, in the Scottish Guide, there is no

Table 2.2. Admissible Evidence for Effectiveness Assessment by Agency

Issue	NICE (England)	IQWiG (Germany)	CDATH (Canada)	SMC (Scotland)	HAS (France)
RCT	Yes**	Yes**	Yes	Active-controlled randomized trials**	High-powered RCT **
				Placebo-controlled randomized trials	Low-powered RCT
Observational Studies	Yes	In exceptional circumstances only	Yes	In exceptional circumstances only	Comparative observational studies
Animal Studies	N/A	N/A	N/A	N/A	No
Expert's Opinion	Yes	N/A	Yes	In exceptional circumstances only	In exceptional circumstances only

** Explicit priority granted.

Sources: *NICE, 2013 ; IQWiG 2017 ; SMC 2019 ; CADTH 2017, HAS 2007*

An in-depth consideration of the arguments provided in the agencies' methodological documents indicates two different perspectives on the role of non-RCT evidence in the effectiveness assessment process. The first type reflects hierarchical reasoning; from this standpoint, for non-RCT data to be considered in an effectiveness assessment, special justification should be provided. As indicated in the previous chapter, such justification is usually given when conducting RCT would be impracticable or unethical. That is, in line with the classical EBM approach, when the golden standard (i.e., RCT evidence) is unavailable, there is no choice but to rely on observational studies as the second-best option. Such reasoning is reflected most significantly in the methodological guidelines published by the German HTA agency, IQWiG.

The second perspective on using non-RCT evidence, which is particularly apparent in the publications of NICE and CADTH, stems from recognition of the limitations of RCT data in the establishing effectiveness claims. These limitations mainly pertain to external validity issues such as those reviewed in the previous chapter. The role of non-RCT data in this context is based on non-hierarchical reasoning, as each type of evidence provides different kinds of information and has its own merits and shortcoming. That is, from this point of view, different

explicit distinction between the use of placebo-based RCT, which is considered inferior to active RCT, and observational evidence. Other guides do not make a distinction based on RCT characteristics in only comparative experiments.

types of evidence are not substitutional, but complementary. While there remains uncertainty relating to the vulnerability of non-RCT data to various biases, this approach evaluates evidence in terms of multiple dimensions, (a lack of) bias being only one of them. Under such a view, non-RCT data may be informative despite its lower level of evidential certainty.[27]

The following section examines how these attitudes regarding the utilization of various evidence types are translated into actual policy decisions regarding public reimbursement of drugs.

2.3.2 Quantitative Analysis

2.3.2.1 Data Collection

To address the methodological issues identified in previous studies, data collection was performed in two phases. In the first phase, we formed a list of all drugs that received market approval by either the EMA or FDA from 2015–2018, specifying the study design of the pivotal efficacy study (or studies) used in their approval assessment. In the second phase, we matched each approved drug to their reimbursement recommendation status in each of the five HTA agencies.

The data collection process was based on a systematic review of publicly available data extracted from the official websites of the two market authorization agencies and the five selected HTA agencies. The following sections describe the process that was used to form the list of approved drugs and their corresponding recommendations.

- ***Market Authorization Dataset***

To avoid selection bias, the market authorization agencies' online databases were searched to identify *all* new active substances granted market authorization between January 1, 2015, and December 30, 2018. We did not include data on drugs approved in 2019, as data for this year is partial and there is a possible time gap between market authorization and HTA submissions, usually due to the capacity of the institution performing the HTA.[28]

The inclusion criteria for analysis were submission of human medication for market authorization for therapeutic proposes, reference to new molecular entities (NMEs), recombination of existing formulations, and new biological substances evaluated by either the

[27] However, it should be noted that it is still unclear whether non-RCT data alone can support effectiveness claims.
[28] Early assessment is conducted about 25–36 weeks after market authorization is granted, depending on the regulatory context.

FDA or EMA.[29] The exclusion criteria were the following: veterinary products, technologies used for diagnostic purposes, generics and biosimilar drugs, and drugs that were reexamined for dose-response assessment purposes. Moreover, we excluded medicines that were refused market authorization and those that were excluded or withdrawn after being approved.[30]

The data on EMA approval were extracted from a list of European public assessment reports (EPARs) generated on July 13, 2019.[31] Information about pivotal evidence in efficacy assessment was collected from data provided by the reviewer, as presented in the efficacy assessment section of the European Public Assessment report on the drug. The list of drugs approved by the FDA was extracted from the CDER's (Center for Drug Evaluation and Research) Novel Drug Approvals reports, which were compatible with our inclusion criteria. Data about the properties of the drugs as well as the types of primary evidence used in appraisals were collected from the drugs@FDA database (https://www.accessdata.fda.gov) based on inspection of the statistical review reports. Missing data regarding trial characteristics were supplemented by data provided by the US National Library of Medicine and EU Clinical Trial Register database.

The data included the specification of clinical indication(s), the type of pivotal efficacy study[32] (RCT or non-RCT), date of approval, and selected properties of the drug (e.g., approval through accelerated assessment pathway, orphan status, and oncological designation).

- ***HTA Agency Dataset***

Data on HTA recommendations were collected from reports published on the agencies' official websites by searching for the approved drugs by name using the websites' search engines. We focused on initial assessment of new pharmaceutical entities rather than full assessment. While full assessment is conducted with a time lag after the drug has already been on the market for some time (approximately a year), the initial assessment is conducted about 25–36 weeks after market authorization is granted (Ivandic, 2014).[33] We chose to focus on early evaluations because the levels of uncertainty at this stage are more substantial and the data available for

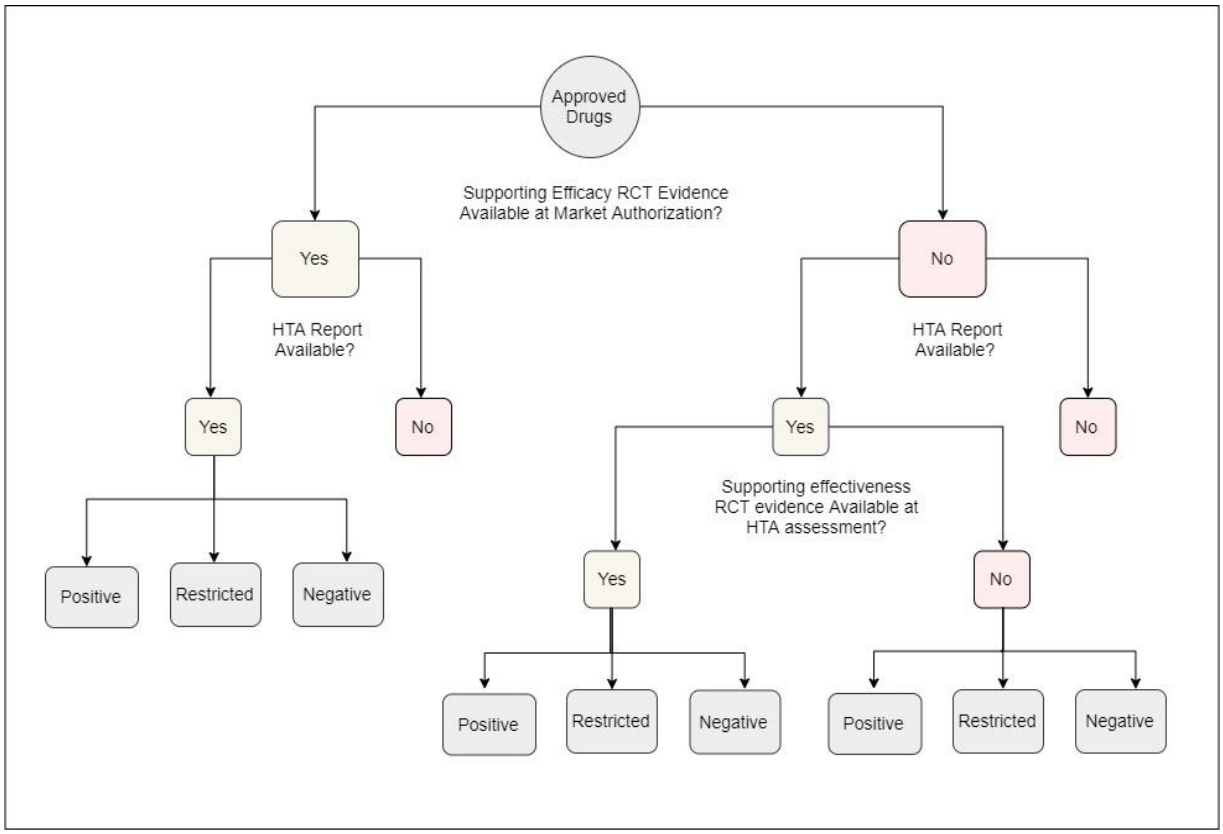
[29] Historically, each country in Europe was independently responsible for pharmaceutical licensing. However, in 1995, the EMA was founded to serve as a centralized licensing and authorization agency. The decisions of the EMA are binding and apply to all member states.
[30] Data regarding medicines that were denied market authorization are available for the EMA, but not the FDA. As we used the list of authorized drugs as a benchmark for investigating reimbursement decisions in the context of HTA processes, which are conducted only on authorized drugs, this does not result in a selection problem.
[31] As mentioned above, drugs approved in 2019 were not included in the analysis.
[32] In cases with more than one pivotal study (≥ 2), we included the study with the most rigorous design.
[33] The time frame changes from country to country.

the assessment are more likely to be similar to the data available during the market authorization review.

The HTA data included the following variables: evaluation status (a binary variable referring to whether the drug has been investigated with the HTA process), recommendation status when applicable (positive, negative, restricted), year of assessment, and type of primary evidence used for the effectiveness evaluation (RCT or non-RCT).[34] Figure 2.2 presents a flow chart of the data collection process.

For further discussion on classification considerations and detailed information about the data sources see *Appendix C.*

Figure 2.2. Data collection flowchart



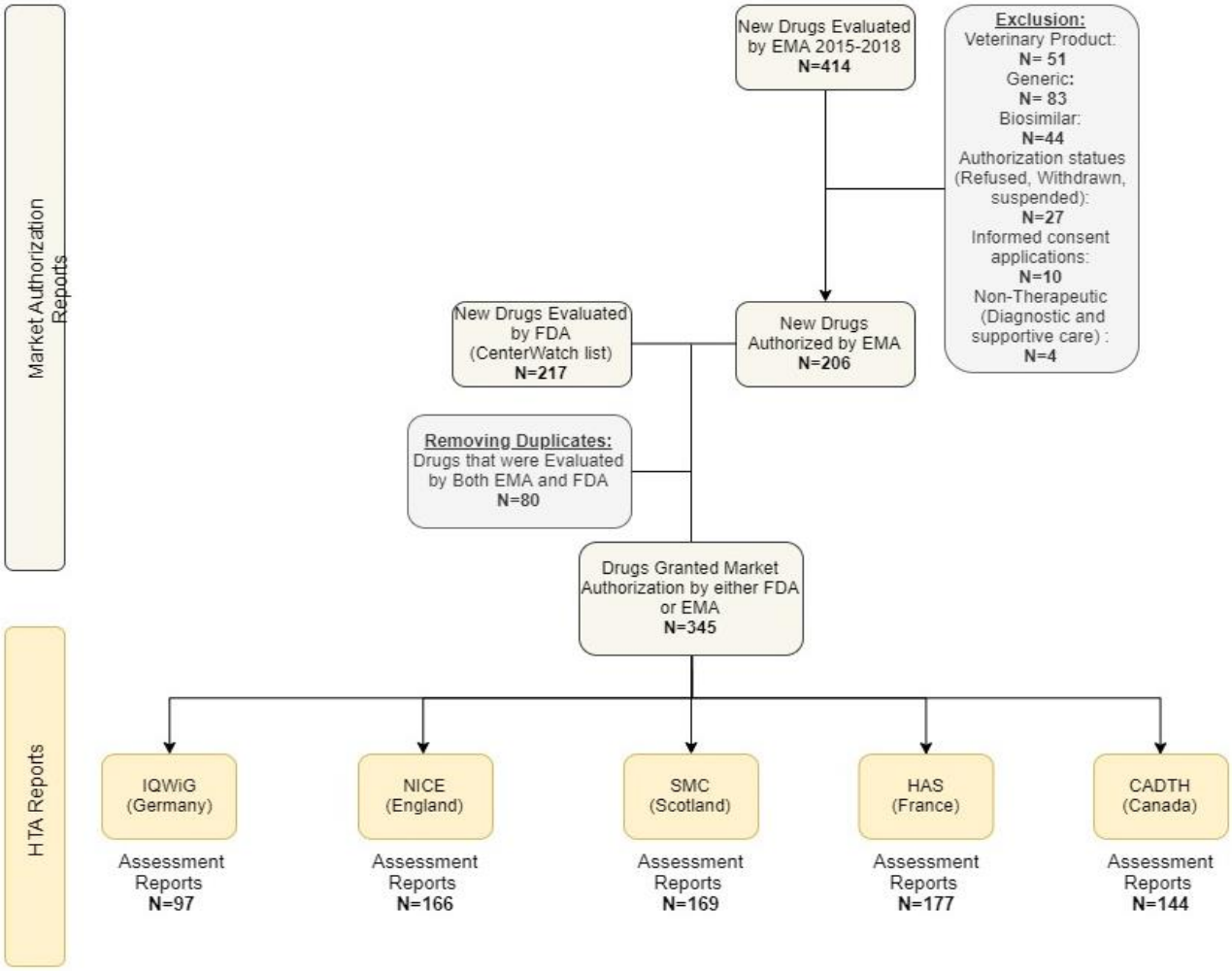
Data Summary

A total of 1,407 reports were identified. Of these, 1,188 reports that met the inclusion criteria were included in the analysis: 435 Market authorization reports and 753 HTA reports. The market authorization reports include 208 EMA appraisals and 217 FDA appraisals. Eighty of the reports referred to the same active substance and received market approval for the same

[34] In some cases, the main study used for evaluation is not explicitly presented as such, but nevertheless it could usually be inferred from the discussion.

indications by both EMA and FDA.[35] In these cases, we used the evidence from the more recent report. Overall, following deduplication, the market authorization dataset contained information on 345 pharmaceutical products that were approved from 2015–2018.

Figure 2.3. Flow chart of data structure



The HTA report database includes 114 appraisals by CADTH, 166 appraisals by NICE, 169 appraisals by SMC, 177 appraisals by HAS, and 97 appraisals by IQWiG. It should be highlighted that, according to German law (social code book V, § 35a, par. 1 sentence 11), the added benefit of authorized orphan drugs is regarded as proven by the fact that they have

[35] We cannot rule out the possibility that some of the medications approved by the EMA were refused by the FDA, as data on refused drugs is unavailable for the FDA. Nevertheless, as we are interested in drugs approved by any of the agencies, this does not distort the data. Moreover, the decisions of the two agencies have a high rate of concordance (91–98%), and rare divergences are usually the result of different interpretation of the same efficacy data by each agency (see Kashoki, 2019).

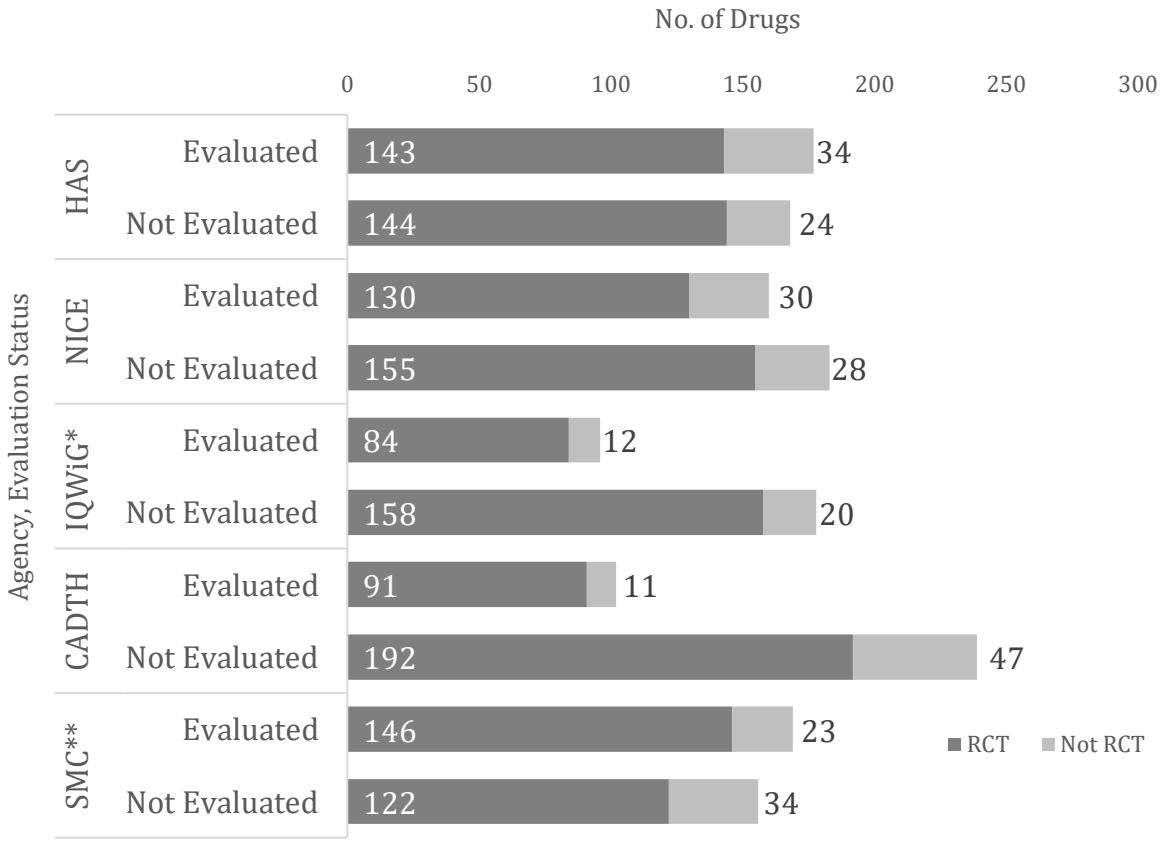
received market authorization. Therefore, the analysis for IQWiG includes only non-orphan drugs.

Descriptive Statistics

In the entire investigated period, 58 approved drugs (16.8%) received market authorization based upon a pivotal study (or studies) with a non-randomized design (the majority, 77%, were prospective single-group assignment clinical studies). The list of drugs without pivotal RCT evidence at the market approval phase can be found in *Appendix D*.

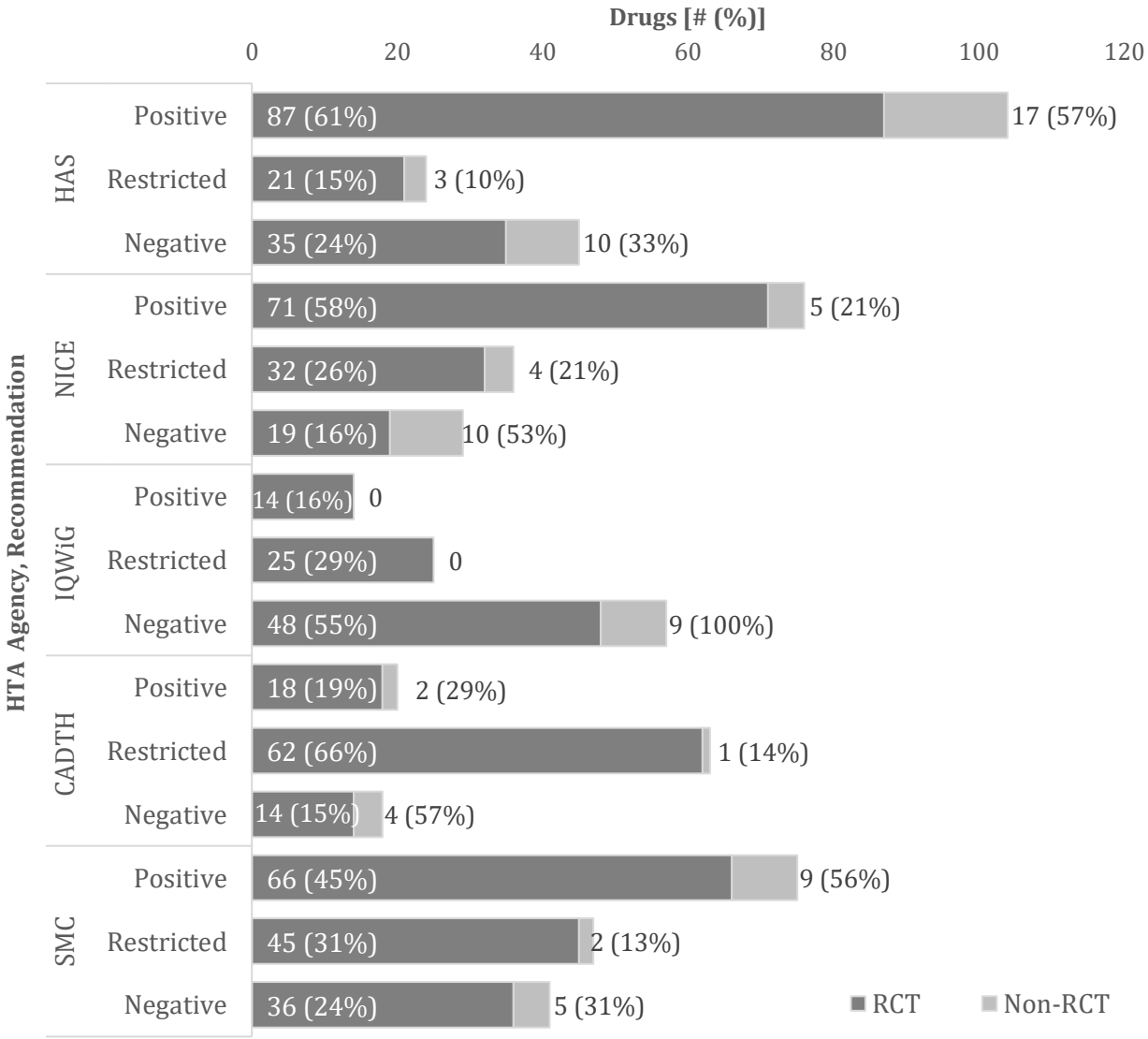
Figures 2.4 and 2.5 presents the distribution of the two main variables: evidence type, evaluation status, and recommendation status by type of pivotal evidence. The distribution roughly aligns with the findings of previous studies and is compatible with NICE's official statistics regarding the overall rate of favorable recommendations (Griffiths et al., 2017).

Figure 2.4 Evaluation Status by Evidence Type



* 71 medicine granted "Orphan" status were omitted
** 20 drugs specified "non-submission" were omitted

Figure 2.5. Evaluation Status by Evidence Type



2.3.2.2 Econometric Model

To test the relationship between the type of evidence available for a drug and its reimbursement recommendation, we used a mixed-effects logistic model. This type of model is well-suited for our analyses because it allows for incorporation of both fixed and random effects. As the observations for the same drug are correlated, when constructing the model, we clustered drug as an identifier variable and applied a random effect to allow for drug-specific random intercepts. The analysis presented below was conducted in two parts, each focused on a different point in time.

As mentioned above, for a drug to be a candidate for public coverage, it must be evaluated via the HTA and be given a favorable recommendation. In some countries, the HTA agency

engages in a preliminary process to select drugs for the HTA. Therefore, we begin by investigating the effect of having RCT evidence during market authorization on the probability of being assessed through the HTA pathway (i.e., evaluation status). Then, we turn to study the effect of RCT evidence during the HTA process on the type of recommendation provided for the drugs going through HTA appraisals in two sub-analysis: One with a binary dependent variable concerning the probability of obtaining a favorable recommendation, and the other concerning the probability of obtaining a specific type of recommendation.

In each part of the analysis, we used three similar models. The basic model included the type of evidence available (RCT or non-RCT) and the HTA agency as predictor variables. The second model included drug-specific properties (orphan status, oncological designation, and market approval in an accelerated pathway) as control variables. In the final model, we included the interaction terms between each HTA agency and the type of evidence.
The analysis was performed using Stata version 13.1. For the full formal specifications of the models, see *Appendix E.*

2.3.2.3 Results

Evaluation within HTA

Table 2.4. summarizes the estimated impact of RCT evidence at the time of market approval on the probability of being assessed through the HTA pathway.
As can be seen from the table, overall, evidence type at market authorization had no significant influence on the probability of being assessed. Similar results were obtained from the second model, which controlled for drugs' features.

When considering the pre-selection procedures, however, it is essential to be aware of variation in the regulatory contexts of different agencies. For example, the German HTA agency, IQWiG, does not initiate selection of the topics for assessment but performs HTA assessment at the request of the Federal Joint Committee (G-BA), which serves as the highest decision-making body in Germany. In contrast, according to the French regulation, HTA evaluation is obligatory for all EMA-approved medicines, and thus the initial filtering process does not apply in this regulatory context for most drugs. Of the remaining three agencies, CADTH (Canada) and NICE (England) have a formal process with predefined criteria for topic selection. Finally, although it is not obligated to do so by law, SMC (Scotland) strives to evaluate all approved drugs. However, it does not actually assess all approved drugs, nor does it present prioritization criteria or establish an orderly pre-evaluation process.

Table 2.4. The effect of RCT at market approval on evaluation status

Dependent Variable: *Evaluated=1*
Model: Mixed-effects logistic regression

		(1) Odd-Ratio	(2) Odd-Ratio	(3) Odd-Ratio
RCT at market authorization $ref. = No$	Yes	1.45 (0.49) [0.49 − 2.68]	1.45 (0.49) [0.49 − 2.68]	0.78 (0.43) [0.26 − 2.33]
HTA agency $Ref. = HAS\ (France)$	NICE (England)	0.69^{**} (0.14) [0.46 − 1.05]	0.70^{**} (0.15) [0.46 − 1.05]	0.60 (0.30) [0.22 − 1.62]
	IQWiG (Germany)	0.34^{**} (0.08) [0.21 − 0.54]	0.34^{**} (0.08) [0.21 − 0.54]	$0.21^{\dagger}$ (0.14) [0.06 − 0.77]
	CADTH (Canada)	0.17^{**} (0.04) [0.11 − 0.28]	0.17^{**} (0.04) [0.11 − 0.28]	0.03^{*} (0.02) [0.10 − 0.12]
	SMC (Scotland)	1.08 (0.23) [0.71 − 1.66]	1.09 (0.24) [0.71 − 1.67]	$0.26^{\dagger}$ (0.14) [0.10 − 0.74]
Drugs' Attributes $ref. = No$	Oncological Drug	−	8.88^{**} (3.45) [4.15 − 19.0]	9.3^{**} (3.69) [4.28 − 20.22]
	Orphan Status	−	0.76 (0.29) [0.43 − 0.37]	0.74 (0.27) [0.44 − 6.61]
	Accelerated Approval	−	0.29^{**} (0.11) [0.14 − 0.61]	0.28^{**} (0.11) [0.13 − 0.60]
Interaction terms Ref. = HAS (France) $\times\ RCT_{t_1}$	NICE (England) $\times\ RCT$	−	−	1.19 (0.67) [0.39 − 3.60]
	IQWiG (Germany) $\times\ RCT$	−	−	1.71 (1.18) [0.44 − 6.61]
	CADTH (Canada) $\times\ RCT$	−	−	6.78^{**} (4.56) [1.81 − 25.3]
	SMC (Scottland) $\times\ RCT$	−	−	5.65^{**} (3.27) [1.82 − 17.56]
	Obs (#)	1628	1628	1628
	Groups (#)	345	345	345
	Wlad Chi^2	79.8^{**}	107.18^{**}	114.64^{**}

Note: Each column reports the estimated effects of a regressions in which the dependent variable is the evaluation status.
Values: Odd-Ratio (Standard deviation) [95% confidence intervals]
** Significant at the 1% level , * Significant at the 5% level, † Significant at the 10% level

Recognizing that differences in the regulatory context may mask the effect of evidence type on the probability of being assessed in different countries, using the third model, we studied the marginal effect of RCT in different contexts by analyzing interactions between the HTA agency and evidence type. This allowed each HTA agency to have a different RCT effect. The results of this model suggest that the effect of the availability of pivotal RCT evidence at market approval on the probability of being evaluated is almost seven times greater for CADTH and six times greater for SMC compared to HAS, in which regulatory obligations allow for only

minimal filtering. However, no difference was detected for NICE, which also applies a pre-selection process.

Reimbursement Recommendations

As mentioned above, in the second part of the analysis, we investigated the reimbursement recommendations provided by HTA agencies for drugs for which an HTA report is available. There are three categories of recommendations classified into: (1) *positive recommendations*, in which the drug is recommended with market authorization as an option for treatment; (2) *restricted recommendations*, in which the drug is recommended for a narrower population or indication than that outlined in the market authorization[36]; and (3) *negative recommendations*, in which the drug is not recommended.

Table 2.5. presents the estimated impact of evidence type during HTA evaluation on the type of recommendation provided by the agency. As this analysis refers to a later point in time than the previous analysis, it should be noted that for some of the drugs that were granted market approval based on non-RCT evidence, RCT evidence (or ongoing partial RCT evidence) was available by the time of the HTA appraisal (NICE: n = 9, CADTH: n = 4, HAS: n = 2, IQWiG: n = 3, SMC: n = 6). Therefore, in this analysis, the RCT predictor used to estimate the models in the table below refers to the availability of RCT data at the later point in time (i.e., at the time of the HTA appraisal). As evaluations of the same drug may be conducted at different points in time, this variable does not include identical values across all five agencies.

Panel (A) in the table focuses on the probability of obtaining a favorable recommendation. In this analysis, the independent variable is binary: a favorable recommendation (i.e., a non-negative recommendation, whether positive or restricted) and an unfavorable recommendation (i.e., a negative recommendation). The findings suggest that the overall likelihood of a drug receiving a favorable HTA recommendation is three times higher when RCT evidence regarding effectiveness is available at the time of assessment. Similar results were obtained when controlling for the drugs' characteristics. The results regarding the interactions between evidence type and agency indicate variation in the effect of evidence type between different HTA agencies. In particular, the effect of RCT evidence on receiving favorable outcome is 4.29 times stronger for NICE and 6.77 times stronger for CDATH compared to HAS. It should also be noted that all drugs with non-RCT evidence evaluated by IQWiG were given negative recommendations. Therefore, the German agency was omitted from this model.

[36] In NICE, restricted recommendations are referred to as "optimized."

Table 2.5. The effect of RCT at HTA evaluation on reimbursement recommendation status

Variable		(A) Dependent Variable: *Favorable Recommendation* **Mixed-effects logistic regression**			(B) Dependent Variable: *Recommendation* **Mixed-effects ologit regression**		
		(4) Odd-ratios	(5) Odd-ratios	(6) Odd-ratios	(7) Odd-ratios	(8) Odd-ratios	(9) Odd-ratios
RCT at HTA evaluation $ref. = No$	Yes	3.07** (0.89) [1.77 − 5.31]	3.43** (0.99) [1.9 − 6.07]	1.77 (0.43) [0.26 − 2.33]	2.32** (0.62) [1.36 − 3.92]	2.77** (0.99) [1.9 − 6.07]	1.69 (0.75) [0.71 − 4.03]
HTA agency $Ref. = HAS\ (France)$	NICE (England)	1.33 (0.38) [0.76 − 2.32]	1.29 (0.37) [0.74 − 2.26]	0.42 (0.28) [0.11 − 1.51]	0.87 (0.20) [.55 − 1.32]	0.83 (0.19) [0.52 − 1.31]	0.25** (0.15) [0.07 − 0.84]
	IQWiG (Germany)	0.20** (0.06) [0.11 − 0.36]	0.21** (0.06) [0.16 − 0.38]	–	0.12**(0.03) [0.07 − 0.21]	0.13** (0.03) [0.07 − 0.22]	–
	CADTH (Canada)	1.40 (0.46) [0.73 − 2.65]	1.35 (0.44) [0.71 − 2.57]	0.26 (0.25) [0.04 − 1.72]	0.34** (0.08) [0.21 − 0.57]	0.35** (0.08) [0.21 − 0.56]	0.17** (0.16) [0.03 − 1.02]
	SMC (Scotland)	0.94 (0.25) [0.71 − 2.14]	0.92 (0.24) [0.55 − 1.55]	1.0 (0.73) [0.24 − 4.18]	0.64 (0.14) [0.39 − 0.94]	0.60 (0.14) [0.39 − 0.94]	0.93 (0.62) [0.25 − 3.42]
Drug Attributes $ref. = No$	Oncological Drug	–	1.11 (0.24) [0.72 − 1.71]	1.02 (3.14) [0.62 − 1.70]	–	1.34 (0.25) [0.92 − 1.94]	1.29 (0.27) [0.86 − 1.96]
	Orphan Status	–	1.25 (0.31) [0.77 − 2.02]	1.31 (6.89) [0.77 − 2.23]	–	1.44† (0.29) [0.97 − 2.14]	1.49† (0.32) [0.98 − 2.29]
	Accelerated Approval	–	1.31 (0.365) [0.77 − 2.21]	1.20 (0.77) [0.65 − 2.23]	–	1.19 (0.26) [0.77 − 1.82]	1.13 (0.28) [0.71 − 1.83]
Interaction Terms Ref. = HAS (France) × RCT_{t_1}	NICE (England) × RCT_{t_1}	–	–	4.29* (3.34) [1.02 − 18.0]	–	–	4.12* (2.73) [1.12 − 15.11]
	IQWiG (Germany) × RCT_{t_1}	–	–	–	–	–	–
	CADTH (Canada) × RCT_{t_1}	–	–	6.77† (6.89) [0.92 − 49.78]	–	–	2.18 (2.04) [0.35 − 13.65]
	SMC (Scottland) × RCT_{t_1}	–	–	0.98 (0.77) [0.21 − 4.55]	–	–	0.65 (0.46) [0.16 − 2.62]
	Obs (#)	675	675	578	675	675	578
	Groups (#)	238	238	237	238	238	237
	Wlad Chi^2	57.53***	58.14***	20.6***	74.37***	79.27***	35.7***

Values: Odd-Ratio (STDEV) [95% confidence intervals] | **Significance:** ** Significant at the 1% level , * Significant at the 5% level, † Significant at the 10% level

To investigate the effect of evidence type on the *specific* recommendation provided by each HTA agency, in the models presented in panel (B) of the table, we treated each type of recommendation separately, placing them on an ordinal scale in which positive recommendations are at the top, restricted recommendations are in the middle, and negative recommendations are at the bottom. Subsequently, we applied a mixed ordered logit model to investigate the effect of evidence type on the probability of a drug receiving a reimbursement recommendation that is one "level higher" than the recommendation that would have been received in the absence of such evidence. That is, we estimated the marginal effect of evidence type in relation to the three recommendation statuses. The results show that in total RCT evidence makes a drug twice more likely to receive a recommendation that is one level higher. This effect is four times stronger for NICE than for HAS. For the other agencies, however, no significant difference was detected.

2.4 Discussion

Confirming our hypothesis, the results of this research indicate that the type of available supporting evidence is predictive of both the evaluation status and reimbursement recommendation provided by the HTA agency. In most cases in which the HTA agency can select the drugs to be evaluated through the HTA pathway, it is likely that it will select drugs for which there was RCT evidence supporting their efficacy at the time of market approval. Therefore, the preliminary assessment can be regarded as a latent mechanism for screening out drugs with no RCT evidence. This finding supports the concern raised above: looking at HTA decisions alone, as previous studies have done, does not provide a complete understanding of HTA agencies' position on RCT evidence.

Moreover, as we expected, the results show not only that the probability of being evaluated is higher for drugs with RCT evidence at market approval but also drugs with RCT evidence are more likely to receive favorable HTA recommendations after appraisal than drugs with other types of evidence. However, the magnitude of this effect is not consistent across countries and is strongest in IQWiG, CADTH, and NICE. The results in this context also indicated that evidence type is a better predictor of the probability of receiving a non-negative recommendation than of the specific type of recommendation. In light of the above, assessment of the relationship between the findings of the two parts of the analysis it is possible that drugs with no RCT evidence are filtered out because they are perceived as having less potential to receive a positive assessment if they were evaluated.

Comparison of the quantitative results regarding actual policy and the qualitative findings regarding the stated policy in official documents reveals a low degree of compatibility. For example, CADTH and NICE expressed the most pluralistic views concerning admissible types of evidence for evaluating effectiveness. Still, in investigating actual decisions made by those agencies, the magnitude of the effect of RCT evidence on the probability of receiving favorable results those agencies them was one of the strongest. In addition, the filtering of non-RCT evidence during the pre-screening phase was particularly significant for CADTH.

Inconsistency in the other direction may be identified for IQWiG. Investigation of the agency's official documents reveals that IQWiG has firmly opposed the idea of using non-RCT evidence to support claims of effectiveness. However, this agency operates in a regulatory environment that applies bypass mechanisms to allow orphan drugs market access, even when no relevant evidence is available. In the next chapter, we will discuss the possible origins of these discrepancies in detail.

Before concluding, it is necessary to mention some limitations of the analysis presented above. First, drugs without RCT evidence may have been filtered out as early as at the market approval stage. Because we do not have information on drugs that were refused market approval, this cannot be ruled out. However, it should be noted that if this is indeed the case, the effect observed in this analysis will only be strengthened. That is, if approved drugs with no RCT have unique characteristics, those characteristics are likely to correlate, at least to some extent, with receiving a positive assessment during the HTA process.

Second, HTA recommendations are based on various considerations, not only effectiveness assessments. Other types of uncertainties, such as those regarding incremental cost-effectiveness ratios, long-term safety, durability, and social, legal and ethical values, may affect the final recommendations. Therefore, we want to stress that in the above analysis, we are only interested in investigating whether evidence type is a *predictor* of reimbursement recommendation, thus serving as indicator of HTA agencies' attitudes toward different types of evidence. That is, we do not seek to make an argument regarding causation.

In light of the investigation of the role of RCT evidence in the HTA process as presented in this chapter, in the final chapter, we perform a normative assessment of existing policies by modeling the decision problem faced by decision-makers when formulating public reimbursement recommendations, while highlighting the challenges that emerge from the structure of uncertainties. Based on this characterization, we investigate the use of the Bayesian approach as a possible pathway for addressing this challenge at both the theoretical and practical levels.

Chapter III

As we know, there are known knowns – these are things we know we know. We also know there are known unknowns – that is to say we know there are some things we do not know; but there are also unknown unknowns – the ones we don't know we don't know…. It is the latter category that tends to be the difficult one" ~ Rumsfeld

In the first chapter, we discussed the benefits and limitations of the RCT method, concluding that the sharp distinction between different types of evidence in the hierarchy of evidence and the role of RCT evidence within it cannot be fully justified on normative epistemic grounds. Subsequently, we empirically investigated the role of RCT evidence in the context of formulating drug reimbursement recommendations, as described in the second chapter. The findings from the second chapter suggest that while RCT evidence is recognized as valuable by policymakers, the extent of their acceptance as a legitimate source of evidence in actual decisions is still limited and it is inconsistent across countries. In this chapter, we aims to critically assess the current decision-making practices, while promoting a better understanding of the challenges involved in effectiveness assessments. Equipped with this understanding of those challenges we turn to discuss a possible framework for addressing them.

When looking at drug reimbursement decisions, coverage recommendations are susceptible to two types of errors with a trade-off between them: Providing a positive recommendation for an ineffective drug gives rise to a type I error; refusing coverage to a beneficial treatment results in a type II error. Facing the task of formulating reimbursement recommendations, decision makers are required to balance the concerns associated with each type of potential error. Over-conservative approaches regarding the type of admissible evidence may result in a large amount of type II errors, while over-incorporative practices will contribute to an increased incidence of type I errors.

As indicated by the findings of the previous chapter, medications without supporting RCT evidence are less likely to be evaluated under the HTA process or to be granted a favorable coverage recommendation once appraised. This suggests that when facing substantial evidential uncertainty, HTA reviewers, who are sensitive to type I errors, prefer to suspend judgment until RCT evidence becomes available.

However, reflecting on the current decision-making practices, it is important to note that postponing the decision is not cost-free and that suspending judgment is a value-laden decision in and of itself. Such a decision involves sacrificing the welfare of current patients for the sake of reduced uncertainty in the treatment of future patients. This point was highlighted by

Claxton (2005) who has been serving as a member of the NICE appraisal committee for two decades:

"…demanding the use of RCT evidence ignores the usefulness of other sources of information. For example, suppose a well-conducted observational (nonrandomized) study suggests that a treatment for a terminal illness is effective. ***Does it make sense to withhold this treatment from the population until a 'proper' RCT can be conducted****?"* (page 94; bold added, MK*).*

Moreover, as noted in the first chapter, in some cases, conducting an RCT is not feasible due to ethical reasons or because of a small target population size. In those instances, it is not likely that the results of an RCT would be available at a matter of time, so waiting for better evidence is not much help. As we have seen, some countries like Germany try to alleviate this difficulty by establishing exceptional coverage routes designated for such cases. Nevertheless, those semi-automatic mechanisms may lead to an inconsistent policy and may lack proper consideration of important hazards, therefore making them especially vulnerable to type II errors.

The evaluation of actual policy, thus, suggests that policymakers are struggling to balance the risks regarding the two types of errors. What is needed to better address these competing concerns, so it seems, is an integrated, holistic structure of the entire body of evidence which incorporates various elements and includes uncertainties at different levels, the degrees of relevance of the evidence, and the magnitude of potential benefits and losses.

Unfortunately, the lexical method of evaluation characterizing many EBM practices does not allow for such a weighing. First, it treats evidence in a discrete manner rather than as existing on a continuum. It therefore does not allow for trade-offs between different types of evidence or between evidence of the same type but of different strength. An RCT may be performed poorly (e.g., conducted on a small study population or not involving a masking strategy) or it may not be relevant for answering the question of interest due to the type of comparator, the specification of endpoints, or issues related to generalizability and external validity.[37] On the other hand, a well-conducted observational study can provide valuable data. While non-RCT evidence is more susceptible to bias, assigning that type with zero (or close to zero) weight when those are the only evidence available seems unreasonable.

[37] From the analysis of the data listed in the previous chapter, 32.7% of the RCTs used for efficacy assessment for market authorization were open label and 30.2% did not involve an active comparator.

Moreover, not only does the decision process following the classical hierarchy of evidence not allow for the weighing of different types of evidence, it also does not capture possible trade-offs between the level of potential benefits and the level of uncertainty. When the potential benefit is substantial, the risk we are willing to take might be more significant in comparison to cases in which the potential benefit is minor. Ruling out certain options based on the types of evidence before their potential benefits have been considered, in this sense, may result in sub-optimal policy decisions.

In light of the above, one may wonder why the use of the hierarchical method, despite its apparent shortcomings, is still dominant—either explicitly or implicitly—in evidence assessment processes for supporting reimbursement decisions. From the investigation of regulatory agencies' documents noted in Chapter II, we can conclude that in most cases, the reason for current practices is not due to a lack of awareness of the limitations of RCTs nor does it is a result of denial of the potential value of non-RCT evidence in providing relevant information. This gap between stated and actual policies implies that the failure to establish appropriate mechanisms for integrating and weighing evidence is rooted in a deeper issue. As we observed in the previous chapter, the difficulty in establishing mechanisms for weighing evidence seems to be particularly evident in cases in which adequate RCT evidence to support effectiveness claims is unavailable, namely when the body of evidence is limited and the level of uncertainty is high.

To promote a better understanding of the origins of the difficulty in establishing appropriate mechanisms for weighing evidence of different types, in the following section we take closer look at the types of uncertainties involved in reviewing evidence of clinical effectiveness.

3.1 Two Types of Uncertainty

Uncertainty is a key property of many policy problems and the object of investigation in many fields including statistics, psychology, law, economics, and philosophy. However, the concept of uncertainty is elusive and complex. Its variety of uses and meanings makes the task of providing a clear and coherent definition for it especially challenging (see Cynfin, 2015). In this context, we will distinguish between two types.

The first type can be referred to as *stochastic* uncertainty (also known as "*aleatory uncertainty"* or "*first-order"* uncertainty). This is associated with random variability in the outcomes within the sample size, such as the one observed when flipping a coin. The second type of uncertainty

is known as "*Knightian"* or *"epistemic"* uncertainty.[38] This kind, first formulated by Knight (1921), does not result from randomness but rather from the lack of information or knowledge.

In reviewing clinical evidence, both types of uncertainties can be traced: As scientists who use statistical analysis aim at inferring parameters from data samples, a random error may result in a gap between the estimated parameter and the true parameter, thus giving rise to first-order stochastic uncertainty. Provided that this type of uncertainty is tightly connected to the idea of variance in the sample mean, it is reduced as the sample size increases to infinity (Marchau et al., 2019).

However, along with this type of "chancy" uncertainty, the evaluation of clinical evidence involves uncertainty of the second type as well. This uncertainty stems from the lack of knowledge that is associated with the construction of the model itself and with the causal ties underlying the relationships among variables in particular. In the context of clinical trials, the epistemic uncertainty is related to our confidence in the quality of evidence or to the estimation of the extent of potential bias. As highlighted in the first chapter, given the skeptical attitude of the medical community toward the establishment of proper theoretical understanding of the causal mechanism governing biological processes, we cannot rule out the influence of unknown intervening variables by merely observing health effects following treatment (Manski, 2007). Therefore, when supporting RCT data are unavailable, the scope of epistemic uncertainty in the quality of evidence may be substantial. As this kind of uncertainty is the result of insufficient knowledge, it *may* be reduced by acquiring additional clinical information of "good" quality. However, it cannot be eliminated by simply enlarging the sample size within an experiment (O'Hagan, 2004).[39]

When the dominant type of uncertainty surrounding policy-issues is the result of stochastic variation, the probabilities are known from observations and policymakers may guide their decisions by referring to alternatives that are optimal *in expectations*. However, optimizing stochastic outcomes under second-order, epistemic uncertainty is much more challenging since

[38] This is also known as *"second-order"* uncertainty, *"structural"* uncertainty, and *"deep"* uncertainty. These terms are not completely similar and in different contexts there may be subtle distinctions among them. However, these distinctions are not significant in assessing the issue we are opening. Therefore, for our discussion, these concepts will be perceived as substitutes.

[39] However, having a robust body of evidence (that is, enlarging the number of experiments) may mitigate uncertainty. As the number of relevant studies and replications increase, so do our degrees of confidence in the parameter obtained. For this reason, meta-analyses and systematic reviews are considered valuable tools for effectiveness assessments.

at least some of the *objective* probabilities are unknown; therefore, measures of uncertainty cannot be easily quantified using classical statistical tools.

A famous example offered by Ellsberg can help to better understand the distinction between two types of uncertainties and demonstrate the challenges it poses for decision-making processes: Ellsberg (1961) presented an experiment involving a decision-maker that is gambling over prospective prize under partial probabilistic knowledge. The experiment involves an urn with 90 balls of which 30 are red. The remaining 60 are either black or yellow. The experimenters are presented with the following four lotteries:

Table 3.1 The Ellsberg Paradox

	30	**60**	
	Red	*Black*	*Yellow*
$\boldsymbol{L_1}$	100$	0$	0$
$\boldsymbol{L_2}$	0$	100$	0$
$\boldsymbol{L_3}$	100$	0$	100$
$\boldsymbol{L_4}$	0$	100$	100$

Note that in this decision problem two types of uncertainties are involved: on the one hand there is first-order uncertainty regarding the color of the ball that will be randomly drawn from the urn. At the same time, as the objective distribution of the yellow and black balls is only partially known, the experimenters are experiencing additional, second-order uncertainty.

When analyzing the decision problem using classical decision theory, L_1 and L_2 are expected to provide similar payoffs, except for the left column, and so does L_3 and L_4 Because the result in the left column is the same for each *pair* of options, minimal requirement of consistency dictates that the experimenters should express the same preference with respect to each pair. That is, if the experimenter prefers $L_1 \succcurlyeq L_2$, she must also prefer $L_3 \succcurlyeq L_4$ (See below the discussion of the expected utility model).

However, presented with the above lotteries, participants typically prefer $\boldsymbol{L_1 \succcurlyeq L_2}$ and $\boldsymbol{L_4 \succcurlyeq L_3}$**.** This seemingly irrational behavior was described by Ellsberg as *"uncertainty aversion"* or "*ambiguity aversion*".

While some have suggested psychological interpretation of this so-called violation of rationality, Ellsberg sought to adhere to the rational interpretation of the behavior that emerges from the experiment[40], suggesting that when such second-order uncertainty, or "ambiguity", is involved, different decision situation takes place, and therefore the existing decision rules are inappropriate for solving the optimization problem.[41]

Analogously, in the context of reviewing clinical evidence, decision-makers are faced with first-order uncertainty pertaining to the observed variability in the degree of treatment effect between individuals (such variability may be observed in the results obtained from an RCT as well as from those obtained from other sources). Unfortunately, when there is a second-order uncertainty stemming from a lack of knowledge regarding the extent of possible bias, the classical statistical methods are insufficient to guide policymaking for resolving first-order uncertainty.

In view of the inadequacy of classical statistical tools for managing cases in which substantial epistemic uncertainty is involved, using the hierarchical method of evidence ranking within drug regulatory processes can be regarded as a heuristic attempt to bypass this challenge. Endorsing a hierarchy of evidence in the context of public decision-making gives rise to a two-stage decision process; According to the lexical logic underlying the hierarchical use of evidence, as a first step one chooses the “best” evidence, where the concept of “best evidence” refers to the type of evidence that would allow as to eliminate (or at least minimize) the level of second-order uncertainty involved in the assessment. As RCT evidence is characterized by a lower risk for the influence of possible unknown confounders, the extent of epistemic uncertainty involved in their evaluation is lower, and thus they are regarded as “better” in terms of their quality compared to others.

In the second stage, being left with a body of “best” evidence characterized by a minimal level of epistemic uncertainty, one can relatively straightforwardly tackle the first-order measurable uncertainty using standard statistical analysis tools. The distinction between the two types of uncertainties presented above, therefore, can shed some light on the motivation behind the

40 Using Ellsberg’s own words: *“…None of the familiar criteria for predicting or prescribing decision-making under uncertainty corresponds to this pattern of choices. Yet the choices themselves do not appear to be careless or random. They are persistent, reportedly deliberate, and they seem to predominate empirically; many of the people who take them are eminently reasonable, and they insist that they want to behave this way.” (page 251)*

41 We should note that Ellsberg originally used his "paradox" as an argument against the Bayesian approach. For the purpose of the discussion, we used the example to clarify the distinction between the different types of uncertainty. Later in the discussion, we will also discuss the criticism that is reflected in Ellsberg's example with regard to the Bayesian decision-making models.

common practice and provide a possible explanation of the appeal of using the hierarchical method in formulating reimbursement recommendations, despite its evident drawbacks.

However, it should be evident by now that the use of such a tool for alleviating the difficulty of managing epistemic uncertainty comes at a sizable cost. Using Nancy Cartwright's words, *"Grading schemes [such as the evidence hierarchy] don't combine evidence at all—they go with what's on top. But it seems to me to be daft to throw away evidence. Why are we urged to do it? Because we don't have a good theory of exactly why and how different types of evidence are evidence and we don't have a good account of how to make an assessment on the basis of a total body of evidence. Since we don't have a prescription for how to do it properly, we are urged not to do it at all. That seems daft too. But I think it is the chief reason that operates. That is why the philosophical task is so important"* (Cartwright et al., np; cited in Blume & Borgerson, 2011, page 230).

The concept of epistemic uncertainty is *"intimately linked to the relationship between theory, knowledge, and evidence"* (Djulbegovic, 2011; page 301). Therefore, providing a coherent theoretical account of this type of uncertainty within a decision-analytical framework plays a vital role in establishing more justified mechanisms for evidence assessment. On the technical level, such mechanisms should allow for the representation of uncertainties on different levels while providing tools for weighing and integrating benefits and evidence from various sources. In the following section, we will explore the use of Bayesian methods as a potential pathway for meeting this challenge on both the theoretical and practical levels.

3.2 Bayesian Analysis–Introduction

The Bayesian method is a branch of statistical analysis that provides an elegant framework for dealing with uncertainty at various levels. Traditionally, statistical methodology has consisted of two major, competing, probabilistic theoretical approaches. The first is the frequentist approach, which interprets probabilities as relative frequencies of empirical events. The second is the Bayesian theory according to which probabilities are rational degrees of belief.[42] Assuming a parameter θ that is to be estimated, the Bayesian method consists of three main elements.

The first element is that of prior beliefs, referring to the subjective probability distribution denoted by $\pi(\theta)$, assigned to the parameter by the agent *before* considering the data. The

[42] Rational degrees of belief $c(\cdot)$ satisfy the Kolmogorov probability axioms: Given set of events $\{A_1 \dots A_n\}$ and sample space Ω, for any event A_I in Ω : **(1)** $c(A) \geq 0$ **(2)** $c(\Omega) = 1$ **(3)** $P(A_1 \cup A_2 \dots) = \sum_{i=1}^{\infty} P(A_i)$.

second is that of posterior beliefs, denoted by $\pi(\theta|x_1 \dots x_n)$, which represents the agent's subjective degrees of confidence in the parameter distribution *after* incorporating the evidence. The third element is a discrete likelihood function, $f(x_1 \dots x_n \mid \theta)$, which measures the support provided by the observed data for possible values of the parameter. In essence, Bayesian analysis is a method of linking the prior probability to the posterior probability. In particular, under this framework, as evidence is accumulated, the degrees of beliefs are updated in accordance with Bayes rule: $\pi(\theta|x_1 \dots x_n) \propto f(x_1 \dots x_n \mid \theta)\pi(\theta)$.

The Bayesian interpretation of probabilities has some appealing theoretical and technical properties for modeling policy problems and for supporting policy decisions in the field of medicine. First, Bayesian thinking allows for the representation of both stochastic uncertainty and epistemic uncertainty under a unified probabilistic theory. In the Bayesian approach, all unknown parameters are treated as random variables following a certain probability distribution. Therefore, probabilities can be assigned to *all* types of uncertainties including those associated with non-empirical events for which relative frequencies cannot be calculated. Treating probabilities as degrees of beliefs that are treated in themselves as random variables allows for the reduction of all kinds of uncertainties to a single quantifiable type. Considering the case of epistemic uncertainty surrounding clinical evidence due to possible bias, in the absence of sufficient knowledge as to the objective probability distribution of the parameter, a Bayesian is expected to use his or her own subjective belief to guide the decision. Such judgment is formed using the (partial) ex-ante knowledge available by the time of the decision. This way, the agent's beliefs about possible bias are integrated with the individual's beliefs on stochastic outcomes, ultimately providing a quantifiable measure of uncertainty under a single scale (O'Hagan, 2004).

Moreover, the Bayesian inference provides a more direct answer to questions relevant to medical research. With the frequentist approach, the parameter θ is treated as a fixed unknown and the data are considered random; in the Bayesian approach, all parameters are random variables and the data are considered fixed. Thus, in contrast to the measures of the frequentist approach, such as p-value and CI which provide the probability of obtaining data as extreme as the observed data when the hypothesis is true, the Bayesian inference involves directly calculating the probability of the parameter given the data $p(\theta|X)$ (Lee & Chu, 2012).

Another important advantage of the Bayesian approach, in the context of evidence assessment, is its integrative nature. The Bayesian approach requires that *all* relevant knowledge be

incorporated in the formation of posterior beliefs. Hence, it can can serve as a better theoretical platform for synthesizing evidence from multifarious sources, including theoretical background knowledge, real-world data, observational clinical trials, and RCTs. The picture arising from the Bayesian method is, therefore, that of a coherent "web" of subjective beliefs mutually supporting each other, therefore allowing for holistic evaluation of the *entire* body of evidence.

Finally, the Bayesian notion of probability is integrated into classical expected utility decision theory models, establishing a framework for the consideration and weighing of both uncertainties and benefits to inform decision makers and to support decision-making processes.[43]

From all of the above, we can conclude that the Bayesian approach constitutes a "richer" language for decision-making processes. The theoretical framework provided by it allows for a single, quantitative representation of various types of uncertainties and it is ideally suited for evaluation of the entire body of evidence on the technical level. This makes the Bayesian framework an especially promising candidate for addressing the problem of clinical effectiveness assessments in the context of formulating reimbursement recommendations. In the following section, we review some applications of the Bayesian method in clinical research and policymaking in practice.

3.3 Application of Bayesian Tools in Clinical Medical Research

3.3.1 Bayesian Clinical Trials

Over the past 50 years, frequentist statistics has been the dominant method in conducting clinical trials, based on the pioneering work of Neyman and Pearson (1933). Recently, however, the growing acknowledgment of the potential advantages associated with the Bayesian method along with computational advancements have contributed to the increasing utilization of Bayesian tools for medical research purposes.

The first application of Bayesian tools to be discussed refers to the generation of clinical evidence. In this context, lately the conduction of Bayesian clinical trials have become common. The results of those trials are presented in the form of the posterior probability of the parameter which is usually defined as the probability of observing a certain endpoint given the

[43] The frequentist theory of significance cannot be incorporated into a theoretical decision framework as it treats probabilities as relative frequencies of empirical events. When those are unknown, expected utility cannot be calculated. See Fisher's critique of Wald: "*The attempt to reinterpret the common tests of significance used in scientific research as though they constituted some kind of acceptance procedure and led to 'decisions' in Wald's sense, originated in several misapprehensions and has led, apparently, to several more"* (Fisher, 1955).

data $\pi(\theta|x_1 \dots x_n)$. As clinical trials are rarely created on their own, Bayesian clinical trials incorporate background knowledge in the formation of the prior distribution $\pi(\theta)$. This knowledge is obtained from various sources including information garnered from previous research phases, basic theoretical science, and observational data, among others. One of the main advantages of utilizing Bayesian analysis methods in performing clinical studies, therefore, is related to the explicit incorporation of background knowledge that renders a more transparent and reproducible use of all available evidence in deriving and interpreting study results (Lee & Chu, 2012).Moreover, while frequentist measures, such as CI and p-value are tied to a particular study design and cannot be compared across experiments, Bayesian analysis methods are more flexible in their evaluation of clinical data, and are more suitable for comparing the results of different studies and their designs may be adapted as data accumulate.

On the regulatory level, the attitude of market authorization bodies toward the use of Bayesian clinical trials is cautious. However, policymakers have recently been more receptive to the idea of Bayesian clinical studies as an admissible source for efficacy assessment within the context of regulatory decision-making. In 2010, the FDA released guidelines for using Bayesian methods in assessing medical devices (FDA, 2010). As of today, these guidelines refer to medical devices but they do not apply to medicines. As the use of Bayesian methods in the effectiveness of evaluating drugs is expanding, we can expect the construction of more comprehensive regulatory structures to be developed accordingly.

Bayesian analysis tools can be technically utilized to analyze the data obtained from experiments with an RCT study design as well as others. It should be noted, however, that from the philosophical perspective, the Bayesian logic is not compatible with the idea of randomized allocation. Bayesianism dictates that the experimenter will allocate participants according to her prior knowledge rather than doing so randomly. As Savage stated 60 years ago, an ideal Bayesian subjectivist *"would not need randomization at all. He would simply choose the specific layout that promises to tell him the most"* (Savage, 1962; page 34). From this perspective, randomization has no epistemic benefit, so assigning participants to a control group is perceived as a waste of resources. In light of the above, if we recall the example of ECMO presented in the first chapter, assigning patients randomly to either a treatment group or a control group when a substantial body of knowledge as to the effectiveness of the procedure already exists is not only unethical, it is also regarded as irrational from a Bayesian point of view.

As Savage himself admitted, though, a randomized allocation may be accepted as admissible by Bayesians on pragmatic grounds when concerns regarding selection bias arise (Savage, 1962). In this context, the distinction presented in the first chapter, between epistemic bias stemming from unknown confounders on the one hand, and psychological attributes leading to bias on the other, can be useful. While randomized allocation cannot be perceived by a Bayesian as a satisfactory way of addressing the concerns of the first type—that is, for avoiding causal fallacies stemming from potential confounders—it can nevertheless be accepted as a useful tool for addressing selection bias, which is related to the second type of concern. In Bayesian terms, RCTs are perceived as a "mixed" strategy over possible experimental allocation; this is never strictly optimal but may, nevertheless, be applied under certain circumstances as a second-best option.[44]

3.3.2 Bayesian Meta-Analysis and Evidence Synthesis

The second application of Bayesian tools in drug regulation is not related to the process of generating new clinical evidence but rather it pertains to the integration of various types of *existing* evidence to support decision-making processes when facing a considerable body of evidence. Bayesian tools are well-suited for evaluating the degrees of coherence and consistency of the entire body of evidence (Higgins & Green, 2011; Woertman et al., 2014; Sutton & Abrams, 2001). As a result, in the past decade, Bayesian methods have become more prominent in conducting meta-analyses and syntheses mainly through the use of hierarchical Bayesian models (e.g., Prevost et al., 2000) and the Markov Chain Monte-Carlo method (e.g., Chen, 2009). Bayesian meta-analyses include explicit specification of the prior distribution for the "between studies" mean effect and standard deviation, as well as the estimation of the extent of possible bias. In light of this specification, studies that are perceived as vulnerable to potential bias can be "discounted." Thus, observational studies may be assigned lesser weight than RCTs depending on their perceived quality and the type of research question associated with each. This specification process, we should note, is inevitably guided, at least to some extent, by subjective judgment.

[44] In recent work, Banerjee et al. (2017) modeled a Bayesian experimenter facing adversarial evidence (for example, a reviewer for market approval agency or an HTA agency). Assuming that the decision-maker is maximizing a mixture of her subjective expected utility and the welfare of an adversarial audience with non-common prior, and that he or she is placing non-zero weight on satisfying the audience, an RCT can turn can be an optimal solution if the experimenter is assigning non-zero weight on satisfying the audience.

3.3.3 Decision-Theoretic Analysis Applications

Expected Utility Theory

Using Bayesian tools to perform meta-analyses allows for trade-offs between degrees of confidence in different types of evidence. However, in formulating reimbursement decisions at the broader level, one may want to allow for trade-offs between the level of benefit and level of uncertainty in the quality of evidence as well. Here, too, the Bayesian method may offer a better framework for weighing different elements of the decision problem against each other. Such tools can be especially useful when facing a limited body of evidence and when RCT data are unavailable.

In order to understand how Bayesian tools can support decision-making in such cases, we will first briefly review the classical Bayesian expected-utility model. This model, as was formulated axiomatically by Savage[45] (1954), consists of three elements:

1. **States of the world** $S(w_1, w_2, w_3 \dots w_n)$. This indicates the object of uncertainty which represents an exhaustive list of possible scenarios. Each state is a full description of a possible world so that under situations guided by perfect information, if true, the consequence of each action is known. In standard decision theory, the state of the world can be represented by a subjective probability distribution concerning the agent's degree of belief that the description of each given possible world will turn out to be the "actual" state of affairs.
2. **Consequences** $[C(c_1, c_2 \dots c_n)]$. Consequences are defined as, "*anything that might happen to a person"* (Savage, 1954; page 13). The consequences embody all elements relevant to the agent's welfare for each choice of action in any given world.
3. **Acts** $[A(a_1, a_2, a_3)]$. Acts are courses of action available to the agent. The set of actions is a function from set S to set C; i.e., it attaches a consequence to each state of the world: $A(\cdot): S \rightarrow C$.

Under the classical decision theory model, as formulated in Savage's representation theorem, if the agent's preferences satisfy minimal axioms of rationality,[46] that person's behavior can be represented as if it maximizes expected utility relative to a unique subjective probability function over the possible states and a utility function (which is unique up to linear transformation) that assigns a numerical value to each possible consequence in each state:

[45] Savage's work was built on the foundations set by the subjective theory of probability developed by Ramsey (1926) and the expected utility theory of von-Neumann-Morgenstern (1944). However, the framework provided by Savage is more convenient to work with and it is widely accepted by economists due to its suitability for the purposes of economic analysis.

[46] The axioms are: Transitivity $(if\ a_1 \succ a_2\ and\ a_2 \succ a_3\ then\ a_1 \succ a_3)$; Completeness $[\forall, a_j \in A\ (i \neq j), either\ a_i \succ a_j, a_i \prec a_j, or\ a_i = a_j]$; Sure-Thing Principle (if a person prefers $a_1\ to\ a_2$, knowing whether a consequence c in another world is obtained, then he or she should prefer $a_1\ to\ a_2$ even if that person knows nothing about the consequence c); and the Rectangular Field Assumption.

$Max\big(E(a_i)\big) = \sum_{i=1}^{n}\big(p(w_i) \cdot u(c_i)\big)$.

In line with the expected utility model, a simplified version of the decision problem of providing a reimbursement recommendation can be formulated as follows:

The decision-maker is facing one group of patients and two treatments $T = \{t_0, t_1\}$ where t_0 is the status quo treatment (alternative, possibly inferior, treatment or no treatment) and t_1 is a novel treatment. The decision maker can provide a positive recommendation for the reimbursement of t_1 or a negative recommendation (thus, staying with the possibly inferior treatment t_0). The set of acts, therefore, would be defined as $A = \{Positive, Negative\}$. For simplicity, we will assume that there is not a competing group of patients and that the status quo treatment t_0 is "no treatment," therefore providing no health benefits and involving no additional, marginal costs. Furthermore, we shall assume perfect knowledge as to costs and potential utilities associated with t_1 as well as the magnitude of its long-term effect.

The decision maker is presented with data obtained from a *single observational trial*, denoted $\mathcal{D}_n = \{x_i \dots x_n\}$ and assigned probability distribution for the parameter θ_{t_i}, measuring the average treatment effect of treatment t_i given the data $\mathcal{D}_n$.[47] The degrees of beliefs assigned by the decision maker given the data is represented, therefore, by the probability function $p\big(\theta_{t_i} > 0 \big| \mathcal{D}_n\big)$ for the drug being effective and $p\big(\theta_{t_i} = 0 \big| \mathcal{D}_n\big)$ for the drug being ineffective.[48] These incorporate both beliefs concerning random variability and beliefs concerning the extent of possible bias.

Each alternative is linked to a health outcome in a possible world. Assuming a utilitarian framework, for each individual $j \in J$, outcomes are measured in units of utilities and costs, $u_j\big(\theta_{t_i}|t_i\big), c_{t_i}$, which are assumed to be known and agreed-upon[49]. Finally, as we are concerned with decision-making at the population level, we will assume a welfare function that aggregates individual utilities $W_{t_i,j} = \Sigma_{\mathrm{i,j}}(u_j\big(\theta_{t_i}|t_i\big), c_{t_i})$. For convenience, the notation for incremental social net benefit will be marked as $INB_{i=\{0,1\}} = u_j\big(\theta_{t_i}|t_i\big)$. In this model, we will consider opportunity loss (that is, health benefit foregone) as an actual loss. That is, refusing treatment

[47] As we assumed that t_0 was "no treatment" and therefore had no health benefit, any positive health benefit ascribed to receiving t_1 would make t_1 superior to t_0.

[48] That is, $\theta_{t_1} > \theta_{t_0}$. As the comparator t_0 was set to be "no intervention" and therefore provided no health benefit, an additional health benefit (i.e., relative effectiveness) is obtained when $\theta_{t_1} > 0$. Moreover, the complementary probability is actually $p(\theta_{t_1} \leq \theta_{t_0})$ but as we assumed the drug was safe (as indicated by its market approval), the inequality was omitted.

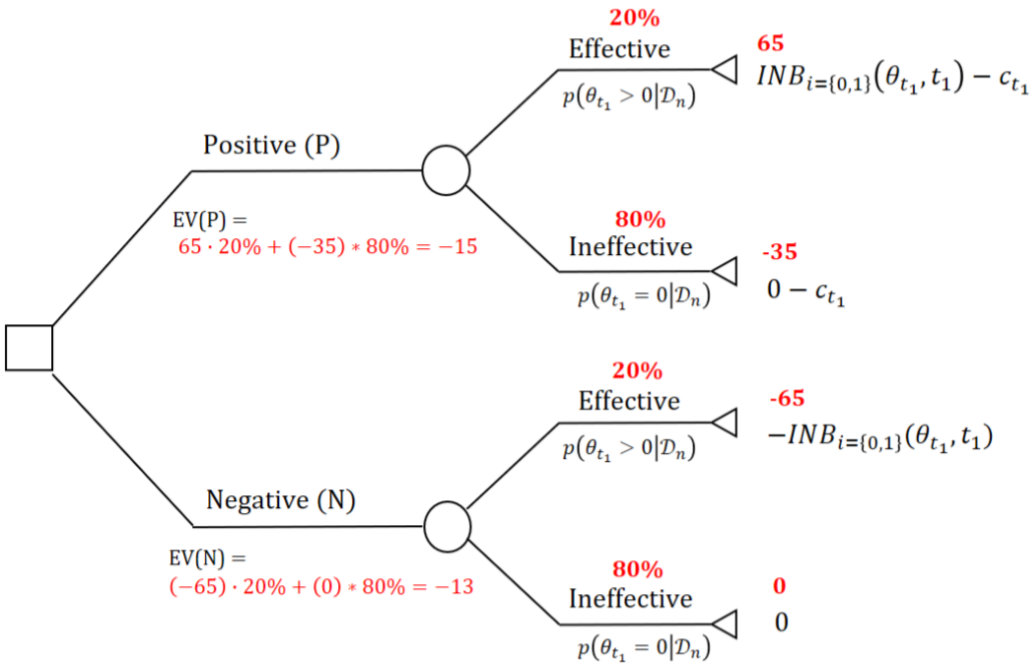
[49] In real-world settings, this is, of course, many times not the case.

in a state where it is effective will result in a negative incremental net benefit in a similar magnitude. In line with the expected utility model, when facing the above decision problem, a rational decision maker is expected to maximize expected utility; that is:

$$\max_{\substack{i=\{0,1\}\\ j=1\ldots n}} \left(\sum p\left(\theta_{t_i}|\mathcal{D}_n\right) \cdot \left(INB\left(\theta_{t_i}\right) - C_{t_i}\right)\right)$$

For illustration, let us consider the following numerical example: The subjective probability assigned by the agent of treatment t_1 being effective (that is, the average treatment effect is higher than zero) given the data obtained from the trial is 20% and the probability assigned for it being ineffective (that is, the average treatment effect equals zero) is 80%. If the drug is effective, 65 units of marginal social welfare would be obtained if the decision maker provides a positive recommendation and 35 INB units would be lost if a negative recommendation is provided. Alternatively, if the drug turned out to be ineffective, 65 INB units would be *lost* when a positive recommendation is granted and no social welfare would be added or lost when the decision maker formulated a negative recommendation. As the expected value of a negative recommendation [EV(N) = -13] is greater than the expected value of a positive recommendation [EV(P) = -15], a rational decision maker would choose to provide a negative recommendation.

Figure 3.2 Numerical Example using Decision Tree



Bayesian Value of Information Analysis

As illustrated by the numeric example above, in its essence, the expected utility model is based on the idea of weighing the level of uncertainty against the level of potential benefit, therefore providing a more holistic approach for the decision-making processes. The expected utility model has another potential advantage: Its theoretical framework may allow us to estimate the value of additional information when facing epistemic uncertainty by using the Bayesian value-of-information (VOI) analysis tool.

VOI analysis, first formulated by Raiffa and Schlaifer (1961), is a useful analytic framework defined as, *"a mean of valuing the expected gain from reducing uncertainty through some form of data collection"* (Wilson, 2015). Regulatory medical policy decisions under uncertainty may result in errors potentially imposing substantial costs on the health system. Acquiring more information may reduce uncertainty but many times also involve considerable costs. VOI analysis allows us to estimate the upper-bound of value that a rational decision maker would be willing to "pay" to obtain additional information. Therefore, under this framework, information is valuable only if it will cause the decision maker to make a different choice compared to the one he or she would have made had the information been unavailable. As such, the value of information has the potential to be a vital tool for better addressing epistemic uncertainty.

The Expected Value of Perfect Information (EVPI) is measured as the difference between the expected benefit obtained when the decision is made under perfect information and the expected benefit of a decision made without perfect information.

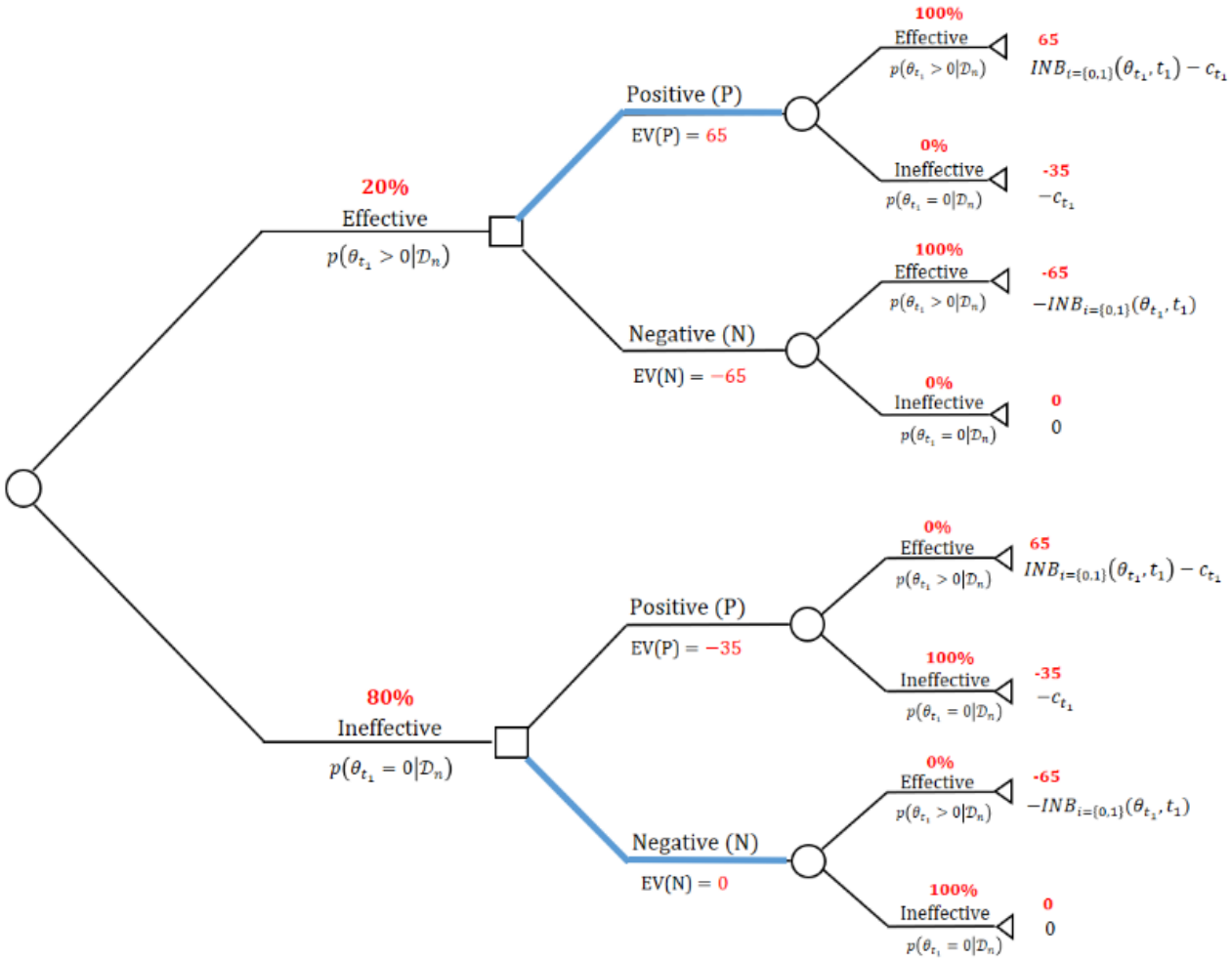
To illustrate the use of VOI in the context of clinical trials, let us consider again the simplified decision problem formulated above. This time let us assume that the decision-maker is facing stochastic uncertainty due to observed variation in the results obtained from an ideal RCT[50]. Solving the decision tree backward (see Figure 3.3), the expected value, without complete information of a positive recommendation as calculated above, is $\mathrm{EV}(P) = (-15)$. In the perfect information case, 20% of the time,[51] the additional benefit is 65; 80% of the time, the additional benefit is 0. Therefore, the value of perfect information is expressed as $\mathrm{EVPI} =$

[50] For the sake of discussion, we will ignore the concerns regarding uncertainty stemming from generalizability issues.

[51] Despite knowing with certainty, the quality of evidence, we considered the EV over a set of patients. The percentage, therefore, does not represent uncertainty but the fraction of patients.

$|13 - (-15)| = 28$. That is, we will be willing to "pay" a maximum of 28 units of expected value to reduce uncertainty.

Figure 3.3 Value of Information Analysis Using Decision Tree



Real-World cases are, of course, more complex than the simplified example presented above. The common method applied for performing VOI analysis for these cases involves the assignment of prior probabilities representing second-order probabilities of the estimate of each model's inputs, followed by incorporating them in a Monte Carlo simulation providing as output the distribution of the estimated INB(θ). The EVPI is finally calculated by measuring the difference between the estimated standardized posterior mean INB(θ) under full knowledge and the point of indifference between the two alternatives of choice under imperfect information. Such a method is utilized for various applications in the field of medical research and economic evaluation, for, among other things, assessing the benefits of sampling (EVSI)[52]

[52] Expected Value of Sample Information.

and for estimating rational costs of collecting the data needed for extrapolation inferences (e.g., Claxton, 2001; Heath et al., 2018).

Alongside these uses, VOI analyses can potentially be utilized to estimate the need for additional evidentiary support of existing data in making reimbursement decisions in the face of a limited body of evidence. In this context, one may assess the added benefit from collecting non-RCT evidence as complementary to existing RCT evidence to reduce uncertainty surrounding specific parameters within the model for which RCT provides partial information (i.e., parameters that are related to long-term effects or to generalizability of the results).

At the same time, one may estimate the "value" of RCTs by estimating the additional, expected benefit obtained from reducing epistemic uncertainty stemming from possible bias. That is, the VOI method can potentially be used for estimating how much one would be willing to "pay" in terms of expected benefit to reduce uncertainty as to the quality of evidence. As such, Bayesian VOI tools may support the management of epistemic uncertainty in clinical decision-making processes and mitigate the difficulty in addressing observational studies when the body of evidence is limited. In this context, it should be noted that according to the rationale underlying the value of information analysis, when the available information is already sufficient for making an optimal decision, collecting more information for the sake of reduced uncertainty will be perceived as irrational. Therefore, when we are sufficiently confident in the quality of non-RCT data available to us, it would be hard to justify the requirement for providing costly RCTs, as it is unlikely that additional information would shift our decision. In this sense, the use of VOI highlights that the value of RCTs in the context of decision-making is not absolute but rather context-dependent.

3.4 Challenges and Objections

In evaluating the potential use of VOI analysis for estimating the "value" of RCT, one may point to a fundamental difficulty arising from the use of this method as the calculation described above requires the formulation of a prior opinion regarding the potential extent of bias. However, as was highlighted in the first chapter, we do not have access to this kind of knowledge. As the risk of bias stems from the inability to account for unknown confounders (that is, it involves factors of "unknown unknowns"), one may argue that turning to the subjective beliefs of the agent as to the influence of such unknown factors yields arbitrary judgments. Such unwarranted formation of the prior opinion, it can be argued, undermines the

credibility of the posterior derived from it and therefore cannot legitimately guide our decisions.

This concern echoes a more general criticism of the Bayesian approach that is related to its subjective nature. In particular, some argue that while the Bayesian theory allows for the assignment of probability to any possible outcome, it does not provide us with tools to judge which probabilities are reasonable to adopt. That is, the Bayesian concept of scientific evidence does not provide an account of *evidential support* (i.e., an indication of the truth of the hypothesis) and therefore cannot establish *warrant* (i.e., a reason to infer the hypothesis) (Reiss, 2014). Under the subjective Bayesian framework, as long as the axioms of rationality are satisfied, a decision made based on arbitrary beliefs would be considered rational. For example, a radical Bayesian subjectivist may come up with any probability between $0 < p < 1$ for the event of a coin toss yielding tails when the coin is actually balanced. In scientific practice, however, we are interested in the actual state of affairs. That is, our beliefs are expected to be indicative of the "truth," at least to a certain extent. As the coin is balanced, such degrees of belief will not be accepted as legitimate in the context of scientific inquiry. Therefore, a strictly subjective Bayesian analysis may be perceived as ill-suited for scientific-practical applications of the type we are discussing.

The problem associated with the subjective nature of the Bayesian approach is especially critical when considering the credibility of the prior probability distribution. Under subjective interpretation, coherence is the only criterion that can impose constraints on the selection of the prior probability. One response in the literature as an attempt to address this challenge is to set some rational constraints on degrees of belief. This approach is known as *"Objective Bayesianism.*" Under this framework, while prior distribution cannot be uniquely determined, there are some prior distributions that would be referred to as "unreasonable," or at the very least, less reasonable than others. Objective Bayesianism usually incorporates empirical constraints, according to which the agent *knowledge* constrains his or her degrees of belief. That is, if the agent *has empirical data* about the nature of the coin, their degrees of belief that it would yield heads on the next toss should be around 50% (Williamson, 2005).[53]

[53] While setting constraints on degrees of belief may alleviate the discomfort associated with the subjective nature of the classical Bayesian inference, some argue that the addition of an external criterion to the theory is an attempt to *"make the Bayesian omelet without breaking the Bayesian egg."* That is, objective Bayesianism involves a substantial deviation from Bayesian logic and its theoretical foundations and therefore can no longer be regarded as a Bayesian method at all.

However, setting empirical constraints on the determination of the prior, however, may raise difficulties in cases where the agent does not possess sufficient objective knowledge to formulate "appropriate" degrees of belief. On the face of it, the case of estimating the extent of potential bias of clinical evidence due to possible *unknown* confounders appears to be a case of this kind. Objective Bayesians would usually respond to this problem by striving to minimize the influence of the prior on the posterior when no data are available to support the process of eliciting the prior distribution. Such attempts are known as the formulations of "*non-informative priors*." A central approach[54] for setting non-informative priors is by following Laplace's principle of insufficient reason [or its generalization under Jaynes' (1957) "maximum entropy" principle]. According to Laplace's rule, when the data are insufficient for justifying the assignment of a specific prior probability, one should assume uniform distribution. That is, given n mutually exclusive and exhaustive events, the prior probability of each event is $\frac{1}{n}$.

While this approach may seem promising at first sight, upon closer examination it turns out to be problematic. First, "non-informative" priors based on uniform distribution rarely remain uninformative under transformation. Consider, for example, the case of two consecutive coin tosses—the distribution of the second toss, $\phi = \theta^2$, is no longer uninformative. Moreover, using uniform distribution would not allow us to distinguish between cases where uniform distribution is assumed based on existing knowledge (for example, assuming a uniform distribution of 50% after observing the results obtained from repeated coin tosses) and cases in which such distribution is assumed due to ignorance. The two cases are fundamentally different on the epistemic level, but under the use of a non-informative prior, it would be impossible to tell them apart.[55] It turns out, therefore, that the translation of all types of uncertainties into a single scale, which at first glance was thought of as an advantage of the Bayesian method, come at the cost of losing an important distinction between different epistemic states.

From the above, it seems that the objective Bayesian approach does not provides as with algorithmic prescription as to the process of prior elicitation in the face of lack of knowledge.

[54] In the literature there are various other suggestions of principles for imposing constraint on degrees of beliefs. See, for example, *Reflection Principle (*Van Frassan, 1984) and *Calibration (*Williamson, 2010).

[55] This implies that we cannot convert causal knowledge to statistical knowledge as the Bayesian method aspires to do in its representation of uncertainty. As was argued by Pearl (2009), *"The vocabulary of probability calculus, with its powerful operators of conditionalization and marginalization, is simply insufficient for expressing causal information"* (page 40).

Unfortunately, those seem to be exactly the cases we are interested in when addressing the issue of the evaluation of clinical evidence.[56]

While we are sympathetic with the concern raised above and recognize its force on the theoretical level, it is important to be cautious in characterizing its scope with regard to actual decision-making and to avoid making too strong a claim. First, it is essential to emphasize that in evaluating the criticisms presented above, the theoretical and pragmatic aspects are interrelated, but they are not necessarily congruent. For the purposes of this discussion, therefore, it would be more convenient to examine the significance of the criticism of the Bayesian approach for each level separately.

On the pragmatic level, one should note that formulating a judgment about potential biases in assessing clinical evidence is normally not a case of complete ignorance: With cases in which the body of evidence (consisting of either RCT or non-RCT data) is considerable, there are yardsticks that can guide our judgment of potential bias—and therefore of specifying the prior—based on the degree of consistency and coherence with previous findings. For example, the reviewer may look for similarities in point estimates and in overlap of confidence intervals provided by results from various studies to evaluate the quality of evidence to form such a judgment.[57]

When the body of evidence is limited, however, the problem of the elicitation of the prior is more significant. Nevertheless, considering actual practice, cases in which the degree of belief in the estimated treatment effect grow in a vacuum are exceptional. In most instances, relevant information can be obtained from theoretical knowledge, expert opinion based on clinical experience, and an assessment of effect magnitude. Design features also serve as relevant knowledge that can support the determination of the prior regarding potential biases in those

[56] We shall highlight that the difficulty in the specification of the prior in the absence of sufficient information is not unique to the objective Bayesian approach since the orthodox subjective Bayesian method appears to be susceptible to a similar problem as well. When ignorance is involved, many would claim that it would be *unreasonable* to require that the agent would guide his or her decision by setting precise degrees of belief as the Bayesian approach dictates. Such an objection has been raised, inter alia, by Gilboa & Marinacci (2016) in their discussion of the Bayesian approach, arguing that, *"Being able to admit ignorance is not a mistake. It is, we claim, more rational than to pretend that one knows what cannot be known"* (page 13). It should also be stressed that the problem above is not a matter of cognitive incompetence in formulating precise or accurate probabilities under these circumstances (which may be the case just as well), but that the normative requirement to do so is in and of itself illogical.

[57] The considerations specified by the GRADE framework (Grading of Recommendations Assessment, Development and Evaluation rating system), used for presenting a summary of evidence for clinical practice recommendations, may be useful in such cases. See the GRADE handbook (2013).

cases. All the above can be utilized to inform the formulation of the prior despite it being limited and based on partial information.[58]

Finally, the use of complementary supportive tools to analyze Bayesian decision-making may alleviate the problem concerning the selection of the prior to some extent. Applying sensitivity analysis tools, for example, may turn out to be particularly valuable in this context. Such analysis includes setting different prior distributions under various assumptions, followed by an exploration of their effect on the output of the model. Therefore, the use of such tools may contribute to a better estimation of the uncertainty involved in formulating the posterior distribution and encourage the calibration of the model's inputs when necessary (e.g., Hendriek, 2009).[59]

As a final remark, we will say a few words about the evaluation of the problem regarding the specification of the prior in light of insufficient knowledge at the theoretical level. It should be recognized that even if a practical concern does not arise within the specific context of our discussion, the difficulties mentioned above still pose a threat more generally at the theoretical level. We shall not discuss the various theoretical considerations that ought to be considered when investigating this issue. Notwithstanding, we will briefly note a conceptual framework discussed in the literature which has evolved to address the intrinsic problems with the Bayesian Orthodox approach as discussed above, while potentially providing an additional advantage at the technical level as well.

This suggested approach, sometimes known as Robust Bayesianism, is regarded by some as a generalization of the Bayesian method (Dempster, 1968). Motivated by the problem with the specification of the prior in cases in which insufficient knowledge is involved, this method gives up the requirement of precise probabilities and turns instead to the use of imprecise probabilities. Such a model[60] may alleviate the general theoretical concern discussed above, as it avoids

[58] It should be noted that the problem of prior characterization may be especially significant at the pragmatic level when it comes to performing Bayesian clinical studies. In those situations, the researcher usually has an interest in characterizing the prior in a manner that would maximize the likelihood of yielding favorable results (this concern holds in general, but it can be especially problematic when it comes to commercial bodies). Explicit characterization of the prior and the rationale underlying it to be critically assessed is crucial in this context as well (see Teira, 2011). However, as the reviewers of regulatory bodies are perceived as "impartial," the concern for manipulation is not dominant in the context of using Bayesian tools for evidence assessment within regulatory processes.

[59] It is also important to note that with the Bayesian approach, the influence of prior distribution is becoming less significant as evidence accumulates.

[60] As for today there is no agreed-upon theory of imprecise probabilities and it is suffering from substantive problems with regard to constructing a plausible confirmation theory. For an elaborate discussion of imprecise probabilities epistemology and the difficulty in constructing a Bayesian Confirmation Theory, see Elkin (2017).

arbitrary judgment in formulating the prior distribution in the face of severe uncertainty. This framework also captures the distinctive notion of epistemic uncertainty through the use of *sets of probability measures* (Elkin, 2017). It therefore maintains the difference between cases in which uniform prior distribution is assigned due to (dis)beliefs and instances in which uniform prior distribution is assumed due to ignorance. Besides the better representation of our doxastic states at the theoretical level, an additional appealing feature of such a method is at the technical level; incorporating imprecise probabilities may support the application of innovative sensitivity tools. In particular, it can be useful in conducting more sophisticated forms of sensitivity analysis using probability bounds analysis (PBA) which allows the investigator to better communicate his or her uncertainty by providing results in bounds on probability distribution (Aughenbaugh & Paredis, 2007; Ali et al., 2012). Future research is needed to thoroughly examine the theoretical and practical implications of applying such tools in evidence assessment processes.

To sum up the discussion in this chapter, we have suggested that the use of the hierarchical method in clinical effectiveness assessment processes may be understood as the result of a failure in accounting for the relationships among clinical evidence, uncertainties, and knowledge of different types on the theoretical level and in establishing weighing mechanisms for supporting decision-making at the technical level accordingly.

Given this difficulty, despite its apparent drawbacks, the use of an evidence hierarchy is still dominant in public reimbursement decision-making processes. The Bayesian framework has been proposed as a more suitable theoretical framework for representing uncertainties at various levels while providing better tools for weighing evidential uncertainty in clinical effectiveness assessment processes.

However, the discussion in the end of this chapter implies that there is no perfect tool for dealing with the complicated problem of evaluating evidence for the formulation of reimbursement recommendations. Bayesian tools suffer from some theoretical and technical issues and are not free of limitations. Nonetheless, the fact that those are imperfect cannot, in and of itself, justify the adherence to another problematic practice just because it is the current practice. Limiting attention to simple problems in which the degree of uncertainty is low may be an easier path to take. Still, it comes at the price of considerable epistemic and distributional costs.

Recognizing that subjective judgment is an integral part of any inference and that scientific evidence alone is never sufficient for establishing an inference should guide us in the normative evaluation of potential mechanisms for addressing evidentiary uncertainty in drug

reimbursement decisions. As should be evident from the discussion in the first chapter, pieces of scientific evidence alone are *never* sufficient to inform policy, as they are underdetermined by our background knowledge. Both the interpretation and the degrees of confidence in the results are necessarily anchored in background prior-knowledge. This is true for RCTs as it is with other types of evidence. The choice, as Claxton (2001) formulated it, *"is not between speculation or evidence but between methods that expose the lack of evidence and make judgments and speculation explicit or those that leave the judgments and speculation for individuals to make implicitly and possibly inconsistently"* (page 51; bold added MK).

Acknowledging the clear advantages of the Bayesian analysis tools, we conclude that the Bayesian approach should be utilized to a further extent in formulating clinical effectiveness appraisals. However, the disadvantages associated with those tools should not be ignored and efforts should be made to minimize their impact and account for them in interpreting the findings of the analysis. In this context, clear and transparent presentation of the rationale underlying the model assumptions as well as evaluation of their influence on the outcomes using sensitivity analysis tools is essential.

The incorporation of Bayesian decision-theoretic tools as an integral part of health technology assessment processes is expected to promote transparency and consistency in decision-making and encourage debate and mutual deliberation among various regulatory agencies. Consequently, the use of these tools may contribute to the standardization of HTA processes across different countries, thereby increasing the coherence of drug regulation policies and recommendations for reimbursement in different contexts. In light of the above, successful implementation of the Bayesian tools will, hopefully, help healthcare systems better meet patient needs and contribute to addressing the significant challenges facing public health systems in the current age, for the benefit of the entire population.

Conclusion

The discussion in this thesis has gone some way toward enhancing our understanding of the role of RCT evidence in decision-making processes on the public coverage of drugs and intricacies associated with it. Recognizing the shortcomings of the current regulatory methods, we used the framework of normative decision theory to argue for the incorporation of Bayesian thinking into drug reimbursement decision-making processes.

The discussion in this work implied that the RCT method is characterized by a special epistemic power, granting it a unique status in the medical community. However, despite its advantages, the RCT method is not a panacea. While its limitations are often overlooked, a more nuanced understanding of the epistemic contribution of the RCT method suggests that it cannot be ascribed an exclusive, overriding weight; RCT evidence alone is often insufficient to substantiate claims about drug efficacy, on the one hand, and evidence from non-RCT sources may provide valuable information that cannot be dismissed, on the other. This suggests that a more pluralistic approach should be adopted with regard to the constituting elements of clinical evidence, by recognizing a broader set of attributes endowing clinical data its evidential force.

However, this pluralistic notion is not adequately manifested in actual policy processes, as indicated by the findings of the retrospective quantitative analysis of reimbursement recommendations. Approved drugs with no supporting RCT data are found to be less likely to be evaluated within HTA processes. Moreover, in the absence of RCT evidence, non-RCT data are seldom perceived as sufficient for supporting effectiveness claims and utilization of this type of data is inconsistent from country to country.

A more in-depth look into the character of the decision problem at stake, using normative decision theory, highlights the multidimensional nature of drug reimbursement decisions. Recognizing that standard tools are insufficient for addressing the challenges emerging from this complex structure, the Bayesian method, while not being free of limitations, has been suggested as a more suitable framework for supporting drug reimbursement decision-making processes. Such incorporation may promote a more legitimate, transparent, and consistent method of decision-making.

As a final remark, it should be noted that examination of the normative basis for evidence assessment processes, as discussed within this thesis with regard to the local context of RCTs, is expected to be of increasing importance at a broader level in the coming years. In the past decade, medical research has rapidly transformed in two directions. As treatment becomes

more personalized, molecular theoretical research is becoming more dominant. The aim of this research is to establish medical progression based on *mechanistic knowledge* at the cellular level. At the same time, computational advancement is contributing to the increasing utilization of Big Data in the conducting of "non-hypothetical" studies. Those studies present results in the form of strong *correlations* arising from the data set itself, while minimizing the use of theoretical knowledge, even at the hypothesis formulation stage.

These opposite trends are expected to pose serious challenges to medical research and may entail modification of the orthodox medical epistemology used for policy formation. Formulation of such epistemology would require a better understanding of the functions of different types of clinical knowledge in providing evidential support for effectiveness claims and the establishment of an adequate framework for integrating them within appraisal processes. Such an endeavor would require collaboration among scholars of different disciplines, including policymakers, health economists, statisticians, and philosophers, as well as others. I hope that our work will further motivate a study of this kind.

Bibliography

Ahn, R., Woodbridge, A., Abraham, A., Saba, S., Korenstein, D., Madden, E., ... & Keyhani, S. (2017). Financial ties of principal investigators and randomized controlled trial outcomes: cross sectional study. *bmj*, *356*, i6770.

Akehurst, R. L. (2017). HTA and reimbursement processes differ across Europe. *PharmacoEconomics & Outcomes News*, *772*, 22-25.

Ali, T., Boruah, H., & Dutta, P. (2012). Sensitivity analysis in radiological risk assessment using probability bounds analysis. International Journal of Computer Applications, 44(17), 1-5.

Allen, N., Walker, S. R., Liberti, L., & Salek, S. (2017). Health Technology Assessment (HTA) case studies: factors influencing divergent HTA reimbursement recommendations in Australia, Canada, England, and Scotland. *Value in Health*, *20*(3), 320-328.

Als-Nielsen, B., Chen, W., Gluud, C., & Kjaergard, L. L. (2003). Association of funding and conclusions in randomized drug trials: a reflection of treatment effect or adverse events?. *Jama*, *290*(7), 921-928.

Andreoletti, M., & Oldofredi, A. (2019). We are All Bayesian, Everyone is Not a Bayesian. *Topoi*, *38*(2), 477-485.

Angelis, A. N. (2017). *Multiple criteria decision analysis for assessing the value of new medical technologies: researching, developing and applying a new value framework for the purpose of health technology assessment* (Doctoral dissertation, London School of Economics and Political Science (LSE)).

Angelis, A., Lange, A., & Kanavos, P. (2018). Using health technology assessment to assess the value of new medicines: results of a systematic review and expert consultation across eight European countries. *The European Journal of Health Economics*, *19*(1), 123-152.

Arrow K.J. (1963). Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review*, 53(5), 941-973.

Atkins, D., Eccles, M., Flottorp, S., Guyatt, G. H., Henry, D., Hill, S., ... & Schünemann, H. (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 services research*, *4*(1), 38.

Aughenbaugh, J. M., & Paredis, C. J. (2007). Probability bounds analysis as a general approach to sensitivity analysis in decision making under uncertainty (No. 2007-01-1480). SAE Technical Paper.

Aute Autorité de Santé. *General Method For Assessing Health Technologies* published December 2007..

Bang, H. (2016). Random guess and wishful thinking are the best blinding scenarios. *Contemporary clinical trials communications*, *3*, 117-121.

Aubbott, D., (2019). “Fresh push for ‘failed’ Alzheimer’s drug”. Published in Nature’s website at 25 October 2019, Available at: https://www.nature.com/articles/d41586-019-03261-5 (Retrieved December 21, 2019)

Battista, R. N., & Hodge, M. J. (1999). The evolving paradigm of health technology assessment: reflections for the millennium. *CMAJ: Canadian Medical Association Journal*, *160*(10), 1464.

Berger, M. L., Sox, H., Willke, R. J., Brixner, D. L., Eichler, H. G., Goettsch, W., ... & Wang, S. V. (2017). Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiology and drug safety*, *26*(9), 1033-1039.

Berna, C., Kirsch, I., Zion, S. R., Lee, Y. C., Jensen, K. B., Sadler, P., ... & Edwards, R. R. (2017). Side effects can enhance treatment response through expectancy effects: an experimental analgesic randomized controlled trial. *Pain*, *158*(6), 1014.

Bero, L., Oostvogel, F., Bacchetti, P., & Lee, K. (2007). Factors associated with findings of published trials of drug–drug comparisons: why some statins appear more efficacious than others. *PLoS Medicine*, *4*(6), e184.

Berry, S. M., Carlin, B. P., Lee, J. J., & Muller, P. (2010). *Bayesian adaptive methods for clinical trials*. CRC press.

Bhandari, M., Busse, J. W., Jackowski, D., Montori, V. M., Schünemann, H., Sprague, S., ... & Devereaux, P. J. (2004). Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials. *Cmaj*, *170*(4), 477-480.

Boshuizen, H. C., & Van Baal, P. H. (2009). Probabilistic sensitivity analysis: be a Bayesian. Value in Health, 12(8), 1210-1214.

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

Cartwright, N. (2011). A philosopher's view of the long road from RCTs to effectiveness. *The Lancet*, *377*(9775), 1400-1401.

CDATH. Guidelines for the Economic Evaluation of Health Technologies*:* Canada. 4th edition, published March 2017. (Retrieved July 2, 2019)

Chen, F. (2009). Bayesian modeling using the MCMC procedure. In *Proceedings of the SAS Global Forum 2008 Conference, Cary NC: SAS Institute Inc*.

Claxton, K., Cohen, J. T., & Neumann, P. J. (2005). When is evidence sufficient?. *Health Affairs*, *24*(1), 93-101.

Claxton, K., Neumann, P. J., Araki, S., & Weinstein, M. C. (2001). Bayesian value-of-information analysis: an application to a policy model of Alzheimer's disease. *International Journal of Technology Assessment in Health Care*, *17*(1), 38-55.

Dall, T. M., Gallo, P. D., Chakrabarti, R., West, T., Semilla, A. P., & Storm, M. V. (2013). An aging population and growing disease burden will require alarge and specialized health care workforce by 2025. *Health affairs*, *32*(11), 2013-2020.

Deaton, A., & Cartwright, N. (2016). The limitations of randomised controlled trials. *VOX, CEPR policy portal.*

Deaton, A., & Cartwright, N. (2018). Reflections on Randomized Control Trials. *Social science and medicine.*, *210*, 86-90.

Delgado, A. F., & Delgado, A. F. (2017). The association of funding source on effect size in randomized controlled trials: 2013–2015–a cross-sectional survey and meta-analysis. *Trials*, *18*(1), 125.

Dempster, A. P. (1968). A generalization of Bayesian inference. Journal of the Royal Statistical Society: Series B (Methodological), 30(2), 205-232.

Detiček, A., Locatelli, I., & Kos, M. (2018). Patient access to medicines for rare diseases in European countries. *Value in Health*, *21*(5), 553-560.

Djulbegovic, B., Hozo, I., & Greenland, S. (2011). Uncertainty in clinical medicine. In *Philosophy of medicine* (pp. 299-356). North-Holland.

Donia, M., Kimper-Karl, M. L., Høyer, K. L., Bastholt, L., Schmidt, H., & Svane, I. M. (2017). The majority of patients with metastatic melanoma are not represented in pivotal phase III immunotherapy trials. *European Journal of Cancer*, *74*, 89-95.

Doucet, M., & Sismondo, S. (2008). Evaluating solutions to sponsorship bias. *Journal of Medical Ethics*, *34*(8), 627-630.

Elkin, L. (2017). Imprecise probability in epistemology (Doctoral dissertation, lmu).

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The quarterly journal of economics*, 643-669.

EMA. (2018). *Real world evidence (RWE) – an introduction; how is it relevant for the medicines regulatory system?*.

EUPATI (2015). Health Technology Assessment process: Fundamentals. Available at: https://www.eupati.eu/health-technology-assessment/fundamentals-of-health-technology-assessment-process/ (Retrieved: July 12, 2019).

Evidence-Based Medicine Working Group. (1992). Evidence-based medicine. A new approach to teaching the practice of medicine. *Jama*, *268*(17), 2420.

FDA, (2018b). *Drug Trials Snapshots Summary Report.* Available at: https://www.fda.gov/drugs/drug-approvals-and-databases/drug-trials-snapshots (Retrieved: 3 August, 2019).

FDA. (2010). *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*. Available at: https://www.fda.gov/regulatory-information/search-fdaguidancedocuments/guidance-use-bayesian-statistics-medical-device-clinical-trials (Retrieved: November 2, 2019)

FDA. (2017). *Drug Trials Snapshots Summary Report.* Available at: https://www.fda.gov/drugs/drug-approvals-and-databases/drug-trials-snapshots (Retrieved: 3 August, 2019).

FDA. (2018a). *Framework for FDA's Real-World Evidence Program*. Available at: https://www.fda.gov/media/120060/download (Retrieved: 12 October, 2019).

FDA. (2019). *Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics.* Guidance for Industry: Draft Guidance. Available at: https://www.fda.gov/media/124795/download (Retrieved: 12 October, 2019).

Fischer, K. E. (2012). A systematic review of coverage decision-making on health technologies—evidence from the real world. *Health Policy*, *107*(2-3), 218-230.

Fried, C. 1974. *Medical Experimentation: Personal Integrity and Social Policy*, Amsterdam: North-Holland Publishing.

Friedman, L. M., Furberg, C., DeMets, D. L., Reboussin, D. M., & Granger, C. B. (2010). *Fundamentals of clinical trials* (Vol. 4). New York: Springer.

Gifford, F. (2011). Philosophy of Medicine: Introduction. In *Philosophy of Medicine* (pp. 1-12). North-Holland.

Gilboa, I., & Marinacci, M. (2016). Ambiguity and the Bayesian paradigm. In *Readings in formal epistemology* (pp. 385-439). Springer, Cham.

Gill, Jennifer, Panos Kanavos, Bernard Avouac, Robert Duncombe, John Hutton, Karina Jahnz-Różyk, Wolfgang Schramm, Federico Spandonaro, and Michael Thomas. "The use of Real World Evidence in the European context: An analysis of key expert opinion." (2016).

Goldenberg, M. J. (2009). Iconoclast or creed?: Objectivism, pragmatism, and the hierarchy of evidence. *Perspectives in Biology and Medicine*, *52*(2), 168-187.

Griffiths, E. A., Macaulay, R., Vadlamudi, N. K., Uddin, J., & Samuels, E. R. (2017). The role of noncomparative evidence in health technology assessment decisions. *Value in Health*, *20*(10), 1245-1251.

Guyatt, G. H. (1998). Evidence-based management of patients with osteoporosis. *Journal of Clinical Densitometry*, *1*(4), 395-402.

Halpern, J., Brown Jr, B. W., & Hornberger, J. (2001). The sample size for a clinical trial: A Bayesian–decision theoretic approach. *Statistics in Medicine*, *20*(6), 841-858.

Haute Autorité de Santé. General Method for Assessing Health Technologies. https://www.has-sante.fr/upload/docs/application/pdf/general_method_eval_techno.pdf *published December 2007.* (Retrieved August 3, 2019)

Heath, A., Manolopoulou, I., & Baio, G. (2015). Efficient High-Dimensional Gaussian Process Regression to calculate the Expected Value of Partial Perfect Information in Health Economic Evaluations. *arXiv preprint arXiv:1504.05436*.

Heath, A., Manolopoulou, I., & Baio, G. (2017). A review of methods for analysis of the expected value of information. *Medical decision making*, *37*(7), 747-758.

Heath, A., Manolopoulou, I., & Baio, G. (2018). Efficient Monte Carlo estimation of the expected value of sample information using moment matching. *Medical Decision Making*, *38*(2), 163-173.

Henry, S., Bond, R., Rosen, S., Grines, C., & Mieres, J. (2019). Challenges in Cardiovascular Risk Prediction and Stratification in Women. *Cardiovascular Innovations and Applications*, *3*(4), 329-348.

Higgins, J. P., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. 4). John Wiley & Sons.

Hill, G. B. (2000). Archie Cochrane and his legacy: An internal challenge to physicians' autonomy?. *Journal of clinical epidemiology*, *53*(12), 1189-1192.

HTAi Global Policy Forum (2019). Real-world evidence in the context of health technology assessment processes – from theory to action: Draft for consultation. Available at: https://htai.org/wp-content/uploads/2018/11/Policy_Brief_GPF_2019_051118_final_line-numbers.pdf (Retrieved: 12 October, 2019)

IQWiG Website. *Search. https://www.iqwig.de/en/search.1029.html*

IQWiG. General Method. *Version 5.0, published July 2017 ;* SMC. Guidance to submitting companies for completion of New Product Assessment Form (NPAF). *Published June 2019.* https://www.scottishmedicines.org.uk/media/4527/20190626-guidance-on-npaf.pdf

Ivandic, V. (2014). Requirements for benefit assessment in Germany and England–overview and comparison. *Health economics review*, *4*(1), 12.

Jack Lee, J., & Chu, C. T. (2012). Bayesian clinical trials in action. *Statistics in medicine*, *31*(25), 2955-2972.

Jager, K. J., Zoccali, C., Macleod, A., & Dekker, F. W. (2008). Confounding: what it is and how to deal with it. *Kidney international*, *73*(3), 256-260.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, *106*(4), 620.

Kabisch, M., Ruckes, C., Seibert-Grafe, M., & Blettner, M. (2011). Randomized controlled trials: part 17 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, *108*(39), 663.

Kanavos, P., Tzouma, V., Fontrier, A. M., Kamphuis, B., Parkin, G. C., & Saleh, S. (2018). Pharmaceutical pricing and reimbursement in the Middle East and North Africa region.

Knight, F. H. (2012). *Risk, uncertainty and profit*. Courier Corporation.

Kristensen, F. B., Lampe, K., Wild, C., Cerbo, M., Goettsch, W., & Becla, L. (2017). The HTA Core Model®—10 years of developing an international framework to share multidimensional value assessment. *Value in Health*, *20*(2), 244-250.

Leibovici, L. (2001). Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial. *Bmj*, *323*(7327), 1450-1451.

Lovelace., Berekely. (2019). "This is 'the most important decision' the FDA will make in 2020, says analyst". Published on CNBC website at December 27, 2019. Available at: https://www.cnbc.com/2019/12/27/this-is-the-most-important-decision-the-fda-will-make-in-2020-analyst.html (Retrieved December 27, 2019)

Lundh, A., Lexchin, J., Mintzes, B., Schroll, J. B., & Bero, L. (2017). Industry sponsorship and research outcome. *Cochrane Database of Systematic Reviews*, (2).

Makady, A., de Boer, A., Hillege, H., Klungel, O., & Goettsch, W. (2017). What is real-world data? A review of definitions based on literature and stakeholder interviews. *Value in health*, *20*(7), 858-865.

Makady, A., ten Ham, R., de Boer, A., Hillege, H., Klungel, O., & Goettsch, W. (2017). Policies for use of real-world data in health technology assessment (HTA): a comparative study of six HTA agencies. *Value in Health*, *20*(4), 520-532.

Makady, A., van Veelen, A., Jonsson, P., Moseley, O., D'Andon, A., de Boer, A., ... & Goettsch, W. (2018). Using real-world data in health technology assessment (HTA) practice: a comparative study of five HTA agencies. *Pharmacoeconomics*, *36*(3), 359-368.

Marchau, V. A., Walker, W. E., Bloemen, P. J., & Popper, S. W. (2019). Decision Making under Deep Uncertainty. *Journal Article*

Masic, I., Miokovic, M., & Muhamedagic, B. (2008). Evidence based medicine–new approaches and challenges. *Acta Informatica Medica*, *16*(4), 219.

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.

Montori, V. M., & Guyatt, G. H. (2008). Progress in evidence-based medicine. *Jama*, *300*(15), 1814-1816.

Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *BMJ Evidence-Based Medicine*, *21*(4), 125-127.

Nagel, E., & Lauerer, M. (Eds.). (2015). *Prioritization in Medicine: an international dialogue*. Springer.

Nardini, C. (2014). The ethics of clinical trials. *Ecancermedicalscience*, *8*.

Newton, W. (2001). Rationalism and empiricism in modern medicine. *Law and Contemp. Probs.*, *64*, 299.

NICE (2013). *Guide to the methods of technology appraisal 2013,* published April 2013. https://www.nice.org.uk/process/pmg9/chapter/evidence (Retrieved July 2, 2019)

NICE (2014). Topic selection. Available at: https://www.nice.org.uk/about/what-we-do/our-programmes/topic-selection (Retrieved: August 27, 2019).

NICE. Single technology appraisal: User guide for company evidence submission template. https://www.nice.org.uk/process/pmg24/chapter/clinical-effectiveness published January 2015, updated April 2017. (Retrieved July 2, 2019)

Nicod, E. (2017). Why do health technology assessment coverage recommendations for the same drugs differ across settings? Applying a mixed methods framework to systematically compare orphan drug decisions in four European countries. *The European Journal of Health Economics*, *18*(6), 715-730.

O'Hagan, T. (2004). Dicing with the unknown. *Significance*, *1*(3), 132-133.

Oortwijn, W., Jansen, M., & Baltussen, R. (2020). Use of evidence-informed deliberative processes by health technology assessment agencies around the globe. *International Journal of Health Policy and Management*, *9*(1), 27-33.

Papineau, D. (1994). The virtues of randomization. *The British journal for the philosophy of science*, *45*(2), 437-450.

Patsopoulos, N. A. (2011). A pragmatic view on pragmatic trials. *Dialogues in clinical neuroscience*, *13*(2), 217.

Pearl, J. (2009). *Causality*. Cambridge university press.

Petrisor, B. A., & Bhandari, M. (2007). The hierarchy of evidence: levels and grades of recommendation. *Indian journal of orthopaedics*, *41*(1), 11.

Pezeshk, H. (2003). Bayesian techniques for sample size determination in clinical trials: a short review. *Statistical Methods in Medical Research*, *12*(6), 489-504.

Pimple, K. D. (2017). *Research ethics*. Routledge.

Prevost, T. C., Abrams, K. R., & Jones, D. R. (2000). Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in medicine*, *19*(24), 3359-3376.

QWiG. General Method. *Version 5.0, published July 2017 ;* SMC. Guidance to submitting companies for completion of New Product Assessment Form (NPAF). *Published June 2019.* https://www.scottishmedicines.org.uk/media/4527/20190626-guidance-on-npaf.pdf ;

Reiss, J. (2014). What's wrong with our theories of evidence?. THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia, 29(2), 283-306.

Russo, F., & Williamson, J. (2011). Epistemic causality and evidence-based medicine. *History and philosophy of the life sciences*, 563-581.

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't.

Savage L.J. (1954). *The Foundations of Statistics.* New York: Wiley.

Savage, L. J., Barnard, G., Cornfield, J., Bross, I., Good, I. J., Lindley, D. V., ... & Dempster, A. P. (1962). On the foundations of statistical inference: Discussion. *Journal of the American Statistical Association*, *57*(298), 307-326.

Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T., Hunter, N. L., ... & Shuren, J. (2016). Real-world evidence—what is it and what can it tell us. *N Engl J Med*, *375*(23), 2293-2297.

Sorenson, C., & Chalkidou, K. (2012). Reflections on the evolution of health technology assessment in Europe. *Health Economics, Policy and Law*, *7*(1), 25-45.

Steel, S., & Ibbetson, D. (2011). More grief on uncertain causation in tort. *The Cambridge Law Journal*, *70*(2), 451-468.

Straus, S. E., Ball, C., Balcombe, N., Sheldon, J., & McAlister, F. A. (2005). Teaching evidence-based medicine skills can change practice in a community hospital. *Journal of general internal medicine*, *20*(4), 340-343.

Susukida, R., Crum, R. M., Stuart, E. A., Ebnesajjad, C., & Mojtabai, R. (2016). Assessing sample representativeness in randomized controlled trials: application to the National Institute of Drug Abuse Clinical Trials Network. *Addiction*, *111*(7), 1226-1234.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical methods in medical research*, *10*(4), 277-303.

Teira, D. (2011). Frequentist versus Bayesian clinical trials. In *Philosophy of medicine* (pp. 255-297). North-Holland.

Thall, P. F., & Wathen, J. K. (2007). Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, *43*(5), 859-866.

Thompson, R. P., & Upshur, R. E. (2017). *Philosophy of medicine: An introduction*. Routledge.

Vandenbroucke, J. P. (2011). Why do the results of randomised and observational studies differ?.

Veatch, R. M. (2007). The irrelevance of equipoise. *Journal of Medicine and Philosophy*, *32*(2), 167-183.

Vere J. W. (2018). Evidence Based Medicine –A Critical Analysis. *A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy*. The University of Sheffield.

Vreman, R. A., Bouvy, J. C., Bloem, L. T., Hövels, A. M., Mantel-Teeuwisse, A. K., Leufkens, H. G., & Goettsch, W. G. (2019). Weighing of evidence by health technology assessment bodies: retrospective study of reimbursement recommendations for conditionally approved drugs. *Clinical Pharmacology & Therapeutics*, *105*(3), 684-691.

Wareham, K. J., Hyde, R. M., Grindlay, D., Brennan, M. L., & Dean, R. S. (2017). Sponsorship bias and quality of randomised controlled trials in veterinary medicine. *BMC veterinary research*, *13*(1), 234.

West, S., King, V., Carey, T. S., Lohr, K. N., McKoy, N., Sutton, S. F., & Lux, L. (2002). Systems to rate the strength of scientific evidence. *Evidence report/technology assessment*, *47*, 1-11.

Williamson, J. (2005). *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press.

Woertman, W., Sluiter, R., & van der Wilt, G. J. (2014). Synthesis of Evidence for Reimbursement Decisions: A Bayesian Reanalysis. *International journal of technology assessment in health care*, *30*(4), 438-445.

World Health Organization (2018). *Pricing of cancer medicines and its impacts:* Technical Report. Geneva. Licence: CC BY-NC-SA3.0 IGO.

Worrall, J. (2002). What evidence in evidence-based medicine?. *Philosophy of science*, *69*(S3), S316-S330.

Worrall, J. (2007). Evidence in medicine and evidence-based medicine. *Philosophy Compass*, *2*(6), 981-1022.

Worrall, J. (2011). Causality in medicine: getting back to the Hill top. *Preventive medicine*, *53*(4-5), 235-238.

Zeilstra, D., Younes, J. A., Brummer, R. J., & Kleerebezem, M. (2018). Perspective: fundamental limitations of the randomized controlled trial method in nutritional research: the example of probiotics. Advances in Nutrition, 9(5), 561-571.

APPENDICES

Appendix A: *Comparative Review of Healthcare System Structure and Health Technology Assessment in the Five Selected Countries*

Country	Healthcare System – General Characteristics		Health Technology Assessment								
			Organizational Structure			Non-clinical Considerations within HTA				Measures of Cost-Effectiveness	
	Health Insurance Model	Drug Reimbursement Scheme	HTA Body conducting clinical effectiveness assessment	Body Making Final Decision	Pre-selection process	Economic Evaluation by legislation	Patient's Aspect Beside Clinical Benefit	Social Aspects	Ethical & Legal Aspects	QALY	HRQoL
France	Universal and compulsory, provided by the state.	Drugs covered under the Social Health Insurance are specified in a positive list. Rate of reimbursement of prescription drugs may vary by type of care and level of effectiveness and therapeutic value.	(HAS) Haute Autorité de Santé	Health Ministry Inclusion in refundable list UNCAM (Union Nationale des Caisses d'Assurance Maladie) reimbursement rate	HTA is mandatory for all EMA approved drugs	Yes	Yes	Yes	Yes	Yes	Yes
England	Universal Coverage provided by NHS England (Single-Payer)	Covered drugs are specified in both negative and positive lists. Drugs provided in hospitalization are fully funded, outpatient prescriptions involves co-payments.	(NICE) National Institute of Health and Clinical Excellence (in absent of report by NICE, appraisal is conducted by local clinical commission groups)	National Health Services of England (NHS England)	Yes, criteria explicitly defined	Yes	Yes	No	No	Yes	Yes

	Health Insurance Model	Drug Reimbursement Scheme	HTA Body conducting clinical effectiveness assessment	Body Making Final Decision	Pre-selection process	Economic Evaluation (by legislation)	Patient's Aspect Beside Clinical Benefit	Social Aspects	Ethical & Legal Aspects	QALY	HRQoL
Germany	Universal health insurance which is mandatory and provided by two systems: (1) non-governmental, non-for-profit health funs (2) private health insurance (The Bismarck Model)	**Drugs with unproved added benefit**: grouped and set reference price, serving as a maximum level of reimbursement. **Drugs with proven added benefit** Sickness Funds negotiates a rebate on manufacturer's prices, lowering the price below reference price.	(IQWiG) The independent Institute for Quality and Efficiency in Health Care	G-BA Federal Joint Committee	HTA agency does not initiate appraisal but does so at the request of the G-BA	Yes	No	No	No	No	Yes
Canada	National Health Insurance Model Universal health insurance administered by the various provinces and territories	All prescription drugs provided in hospitals are fully covered publicly, with outpatient coverage varying by province or territory.	(CADTH) Canadian Agency for Drugs and Technologies in Health	Local bodies in each province or territory	Yes criteria explicitly defined	Yes	Yes	Yes	No	Yes	Yes
Scotland	Universal coverage provided by NHS Scotland (single-Payer), provided by nine Health Boards	Drugs provided in hospitalization are fully funded, outpatient prescriptions involves co-payments.	(SMC) Scottish Medicines Consortium	National Health Services of Scotland (NHS Scotland)	Yes no criteria explicitly defined	Yes	Yes	Yes	No	Yes	Yes

Sources: European Commission (2017). *Mapping the HTA Methodologies in EU and Norway. Written by Science & Policy, Author Finn Børlum Kristensen. The Commonwealth Fund Website. International Health Care System Profiles. https://international.commonwealthfund.org/*

***Appendix B**: Documents Analysis*

Issue	NICE (UK)	IQWiG (Germany)	CDATH (Canada)	SMC (Scotland)	HAS (France)
Types of Admissible Evidence	RCTs, Observational Studies " NICE prefers RCTs that directly compare the technology with 1 or more relevant comparators."	RCTs. **In exceptional circumstances only:** Observational Studies	**All types of Data:** " Potential sources for informing parameter estimates for effectiveness (e.g., clinical effects, detection, harms) could include RCTs, observational studies, administrative databases, non-comparative studies, or expert input. "	**Priority:** Active control RCT **In the absent of active-controlled:** placebo controlled, uncontrolled studies), **Under special circumstances:** experts' opinion.	**Level I**: high powered RCTs **Level II:** low powered RCT, comparative observational studies **Level III**: Case control studies **Level IV** Retrospective studies, Case series, controlled data with biased (data considered non-relevant: Animal studies, Experts' opinion)
Hierarchy of evidence	Explicit	Explicit	None	Explicit	Explicit
Role of RCT	**Establishing Causal Relationship:** "Randomised controlled trials (RCTs) minimise potential external influences to identify an effect of one or more interventions on outcom" "RCTs are therefore considered to be most appropriate for measures of relative treatment effect."	**Establishing Causal Relationship:** "RCTs provide a basic precondition for the demonstration of causality."	N/A	N/A	N/A
Limitation of RCT	"The relevance of RCT evidence to the appraisal depends on both the external and internal validity of each trial." " However, such evidence may not always be available and may not be sufficient to quantify the effect of treatment over the course of the disease."	"Even if patient groups in an RCT differ from everyday health care, this does not mean the external validity of study results must be questioned".	"A key issue is the extent to which the data obtained from an RCT reflect the effectiveness likely to be achieved in a real-world setting (i.e., the external validity of the trial). For the evaluation to be relevant to the decision-maker, the effects …should reflect the effectiveness of the intervention rather than its efficacy."	N/A	N/A

Issue	NICE (UK)	IQWiG (Germany)	CDATH (Canada)	SMC (Scotland)	HAS (France)
Role of non-RCT	**Complementary:** "Data from non-randomised and non-controlled studies may be needed to supplement RCT data."	**Second best:** "Study types other than RCTs are usually not suited to demonstrate causality. In nonrandomized comparative studies, as a matter of principle structural equality of groups cannot be assumed. They therefore always provide a potentially biased result and mostly cannot answer with sufficient certainty the relevant question as to whether a difference observed is caused by the intervention tested. The use of non-randomized studies as proof of the causality of an intervention therefore requires particular justification or specific preconditions and special demands on quality"	**Complementary:** "Critical to making a judgment about incorporating real-world factors into the analysis is the strength of the available data linking potential intervention effect-modifying factors with important patient outcomes. Researchers should present these linkages in a transparent manner and provide justification. "	**Second best:** "If active-controlled studies are not available, details of placebo-controlled or uncontrolled studies that provide evidence of the clinical benefits of the medicine in its licensed dose within the indication(s) under review should be included… Where data from studies are insufficient to provide values for relevant variables, and such values can be obtained from expert opinion, then SMC will consider this as a valid source of evidence."	**Complementary:** "Comparative observational studies might be used in the case of added value, in terms of relevance or bias limitation"

Sources: NICE. *Guide to the methods of technology appraisal 2013*, published April 2013. https://www.nice.org.uk/process/pmg9/chapter/evidence , NICE. Single technology appraisal: User guide for company evidence submission template. https://www.nice.org.uk/process/pmg24/chapter/clinical-effectiveness published January 2015, updated April 2017; IQWiG. General Method. *Version 5.0, published July 2017 ;* SMC. Guidance to submitting companies for completion of New Product Assessment Form (NPAF). *Published June 2019.* https://www.scottishmedicines.org.uk/media/4527/20190626-guidance-on-npaf.pdf ; CDATH. Guidelines for the Economic Evaluation of Health Technologies*:* Canada. 4th edition, published March 2017. ; *HAS. Haute Autorité de Santé.* General Method For Assessing Health Technologies. https://www.has-sante.fr/upload/docs/application/pdf/general_method_eval_techno.pdf *published December 2007.*

Appendix C: *Categorization of the features of pivotal studies*

Data Sources:

Market Authorization Reports:

EMA:	*Main Study*" component of the "Clinical Efficacy" section in the "EMA Public Assessment Report (EPAR)".
FDA:	"*Clinical/Statistical-Efficacy*" component (Section 7) in the "Summary Review" report, or the "*Statistical Evaluation"* component in the "Statistical Review(s)" report.
HTA Appraisals:	
NICE:	"*Clinical evidence*" component in the "Community discussion" section (section 3) of the Technology appraisal guidance.
HAS:	"*Clinical Data*" section in the english version of the "Brief Summary of the Transparency Committee Opinion" report.
IQWIG:	Assessment section (section 2) in the English version of the "Addendum to Commission A16-0" report or the "Extract of dossier" report.
SMC:	"*Why has SMC said this*?" component in the public summary report ("Decision Explained")
CADTH:	*Summary of pERC deliberation*" component of the "Expert Review Committee - initial recommendation" reports.

Complementary Data:

- **US National Library of medicine** ClinicalTrial.gov at (https://clinicaltrials.gov/), search by study name. When necessary, ClinicalTrial.gov Archive site was used to trace the relevant study record according to the date of the report.
- **PubMed** (US National Library of Medicine) at https://www.ncbi.nlm.nih.gov/pubmed/ search by keywords.
- **EU Clinical Trials Register** at https://www.clinicaltrialsregister.eu/, search by study name.

***Appendix D**: List of Drugs Granted Market Authorization by either EMA or FDA Based on Non-RCT pivotal trial*

Medicine name	Therapeutic Area	Accelerated Assessment	Orphan Medicine	Marketing Authorization Year	Trial No. (ClinicalTrials.gov Identifier, when applicable)	Phase
Lumark	Metastatic Prostate Cancer	no	no	2015	NCT00195039	2
Zykadia	Carcinoma, Non-Small-Cell Lung	no	no	2015	NCT01283516	1
Strensiq	Hypophosphatasia	no	yes	2015	NCT00952484	2
Elocta	Hemophilia A	no	no	2015	NCT01181128	3
Blincyto	Precursor Cell Lymphoblastic Leukemia-Lymphoma	no	yes	2015	NCT01466179	2
Praxbind	Hemorrhage	yes	no	2015	NCT02104947	3
Obizur	Hemophilia A	no	no	2015	NCT01178294	2/3
Kolbam	Metabolism, Inborn Errors	no	yes	2015	NCT00007020	3
Vistogard	overdose of capecitabine or fluorouracil	yes	yes	2015	401.10.001	3
Venclyxto	Leukemia, Lymphocytic, Chronic, B-Cell	no	no	2016	NCT01889186	2
Kovaltry	Hemophilia A	no	no	2016	NCT01311648	3
trientine	Hemophilia B	no	yes	2016	NCT01361126	1/2
Briviact	Epilepsy	no	no	2016	NCT00150800	3
Sialanar	Sialorrhea	no	no	2016	NCT00425087.	3
Alprolix	Hemophilia B	no	yes	2016	NCT01027364	3
Coagadex	Factor X Deficiency	yes	yes	2016	NCT00930176	3
Tagrisso	Carcinoma, Non-Small-Cell Lung	yes	no	2016	NCT02094261	2
Strimvelis	Severe Combined Immunodeficiency	no	yes	2016	NCT00598481	2
Zalmoxis	Hematopoietic Stem Cell Transplantation, Graft vs Host Disease	no	yes	2016	NCT00423124	1/2
Idelvion	hemophilia B	no	no	2016	NCT01496274	2/3
Venclexta	chronic lymphocytic leukemia	yes	yes	2016	NCT01889186	2

Medicine name	Therapeutic Area	Accelerated Assessment	Orphan Medicine	Marketing Authorization Year	Trial No. (ClinicalTrials.gov Identifier, when applicable)	Phase
Defitelio	hepatic veno-occlusive disease	yes	yes	2016	NCT00358501	3
Tecentriq	Carcinoma, Transitional Cell, Carcinoma, Non-Small-Cell Lung	no	no	2016	NCT02951767 (Cohort 1), NCT02108652 (Cohort 2)	2
Afstyla	Hemophilia A	no	no	2017	NCT01486927	2/3
Cuprior	Hepatolenticular Degeneration	no	no	2017	"Lariboisière study"	NA
Qarziba	Neuroblastoma	no	yes	2017	APN311-303	NA
Bavencio	Neuroendocrine Tumors	no	yes	2017	100070-003	2
Brineura	Neuronal Ceroid-Lipofuscinoses	yes	yes	2017	NCT01907087	1/2
Zubsolv	Opioid-Related Disorders	no	no	2017	NCT01903005	3/4
Alecensa	Carcinoma, Non-Small-Cell Lung	no	no	2017	NCT01871805	1/2
Chenodeoxycholic acid Leadiant	Xanthomatosis, Cerebrotendinous, Metabolism, Inborn Errors	no	yes	2017	CDCA-STUK-15-001	NA
pembrolizumab	Carcinoma, Non-Small-Cell Lung	no	no	2017	NCT02335424	2
Keytruda	measurable urothelial carcinoma	no	yes	2017	NCT02335424	2
Bavencio	Neuroendocrine Tumors	na	yes	2017	NCT02155647	2
Imbruvica	Waldenström’s Macroglobulinemia	yes	yes	2017	NCT01614821	2
Imbruvica	marginal zone lymphoma	yes	yes	2017	NCT01236391	2
Aliqopa	relapsed follicular lymphoma (FL)	yes	yes	2017	NCT 01660451	2
Calquence	mantle cell lymphoma (MCL)	yes	yes	2017	NCT02213926	2
IDHIFA	elapsed or refractory acute myeloid leukemia (AML)	yes	yes	2017	NCT01915498	1/2
Rebinyn	hemophilia B	no	no	2017	NCT01333111	3
Alkindi	Adrenal Insufficiency	no	no	2018	NCT02720952	2
Adynovi	Hemophilia A	no	no	2018	NCT01736475	2/3
Yescarta	Lymphoma, Follicular, Lymphoma, Large B-Cell, Diffuse	no	yes	2018	NCT02348216	2
Rubraca	Ovarian Neoplasms	no	no	2018	NCT01891344	2
Myalepta	Lipodystrophy, Familial Partial	no	yes	2018	NIH 991265	2

Medicine name	Therapeutic Area	Accelerated Assessment	Orphan Medicine	Marketing Authorization Year	Trial No. (ClinicalTrials.gov Identifier, when applicable)	Phase
Kymriah	Precursor B-Cell Lymphoblastic Leukemia-Lymphoma, Lymphoma, Large B-Cell, Diffuse	no	yes	2018	NCT02435849	2
Keytruda	head and neck cancer	yes	no	2018	NCT01848834	1
Libtayo	metastatic cutaneous squamous cell carcinoma (CSCC)	no	no	2018	NCT02760498	2
Revcovi	adenosine deaminase severe combined immune deficiency (ADA-SCID)	no	yes	2018	NCT 01420627	3
Copiktra	refractory chronic lymphocytic leukemia	yes	yes	2018	NCT01882803	2
Elzonris	blastic plasmacytoid dendritic cell neoplasm (BPDCN)	no	yes	2018	NCT 02113982	1/2
Lumoxiti	relapsed or refractory hairy cell leukemia	no	yes	2018	NCT01829711	2
Tibsovo	relapsed or refractory acute myeloid leukemia (AML)	yes	yes	2018	NCT02074839	1
Gamifant	primary hemophagocytic lymphohistiocytosis (HLH)	no	yes	2018	NCT01818492	2/3
Trogarzo	immunodeficiency virus type 1 (HIV-1)	no	yes	2018	NCT02475629	3
Annovera	contraceptives.	yes	yes	2018	NCT00263341	3
Lorbrena	metastatic non-small cell lung cancer	yes	yes	2018	NCT01970865	1/2

Appendix E : *Specification of the Econometric Models*

Dependent variable	Model	Model No.	Regression*
The probability of being evaluated			
$Y_i = \begin{cases} 0 & Not\ Evaluated \\ 1 & Evaluated \end{cases}$	**Mixed-effects logistic regression**	(1)	$P(Y_i = 1) = F\left[\beta_1\left(RCT_{t_o}\right)_i + \beta_2\left(HTA\_Agency_j\right)_i + u_i \cdot Z_i + \varepsilon_{i,j}\right]$
		(2)	$P(Y_i = 1) = F\left[\beta_1\left(RCT_{t_o}\right)_i + \beta_2\left(HTA\_Agency_j\right)_i + \beta_3(Oncology)_i + \beta_4(Orphan)_i + \beta_5(AcceleratedAssessment) + u_i \cdot Z_i + \varepsilon_{i,j}\right]$
		(3)	$P\left(Y_{i,} = 1\right) = F[\beta_1\left(RCT_{t_0}\right)_i + \beta_2\left(HTA\_Agency_j\right)_{i,} + \beta_3(Oncology)_i + \beta_4(Orphan)_i + \beta_5(AcceleratedAssessment) + \beta_6\left(HTA_{Agency_j} \times RCT_{t_o}\right)_i + u_i \cdot Z_i + \varepsilon_{i,j}]$
The probability of obtaining favorable reimbursement recommendation			
$Y_i = \begin{cases} 0 & Unfavorable \\ 1 & favroable \end{cases}$	**Mixed-effects logistic regression**	(4)	$P(Y_i = 1 \mid Evaluated = 1) = F\left[\beta_1\left(RCT_{t_1}\right)_{i,j} + \beta_2\left(HTA\_Agency_j\right)_i + u_i \cdot Z_i + \varepsilon_{i,j}\right]$
		(5)	$P(Y_i = 1 \mid Evaluated = 1) = F\left[\beta_1\left(RCT_{t_1}\right)_{i,j} + \beta_2(HTA\_Agency)_i + \beta_3(Oncology)_i + \beta_4(Orphan)_i + \beta_5(AcceleratedAssessment) + u_i \cdot Z_i + \varepsilon_{i,j}\right]$
		(6)	$P(Y_i = 1 \mid Evaluated = 1) = F[\beta_1\left(RCT_{t_1}\right)_{i,j} + \beta_2(HTA\_Agency)_{i,j} + \beta_3(Oncology)_i + \beta_4(Orphan)_i + \beta_5(AcceleratedAssessment) + \beta_6\left(HTA_{Agency_j} \times RCT_{t_1}\right)_i + u_i \cdot Z_i + \varepsilon_{i,j}]$
The probability of obtaining specific reimbursement recommendation			
$Y_i = \begin{cases} 1 & Negative \\ 2 & Resricted \\ 3 & Positive \end{cases}$ $k = \{1,2\}$	**Mixed-effects ologit regression**	(7)	$P(Y_i > k \mid Evaluated = 1) = F^*\left[\beta_1\left(RCT_{t_1}\right)_{i,j} + \beta_2\left(HTA\_Agency_j\right)_i + u_i \cdot Z_i + \varepsilon_{i,j}\right]$
		(8)	$P(Y_i > k \mid Evaluated = 1) = F^*\left[\beta_1\left(RCT_{t_1}\right)_{i,j} + \beta_2(HTA\_Agency)_i + \beta_3(Oncology)_i + \beta_4(Orphan)_i + \beta_5(AcceleratedAssessment) + u_i \cdot Z_i + \varepsilon_{i,j}\right]$
		(9)	$P(Y_i > k\ Evaluated = 1) = F^*[\beta_1\left(RCT_{t_1}\right)_{i,j} + \beta_2(HTA\_Agency)_{i,j} + \beta_3(Oncology)_i + \beta_4(Orphan)_i + \beta_5(AcceleratedAssessment) + \beta_6\left(HTA_{Agency_j} \times RCT_{t_1}\right)_i + u_i \cdot Z_i) + \varepsilon_{i,j}]$

* Where i is index for the drug, j is index for the HTA agency, u_i is random drug-intercept, F and F^* are logistic functions of the following type: $\frac{1}{1+\exp(\Sigma_i \beta_i x_i + \varepsilon_i)}$