The Epistemology and Ethics of Early Stopping Decisions
in Randomized Controlled Trials

by

Roger Stanev

B.Sc. Physics, University of São Paulo, 1996
M.S.E., Seattle University, 2001
M.A. Philosophy, Virginia Polytechnic Institute and State University, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES

(Philosophy)

The University of British Columbia

(Vancouver)

June 2012

© Roger Stanev, 2012
Abstract

Philosophers subscribing to particular principles of statistical inference and evidence need to be aware of the limitations and practical consequences of the statistical approach they endorse. The framework proposed (for statistical inference in the field of medicine) allows disparate statistical approaches to emerge in their appropriate context. My dissertation proposes a decision theoretic model, together with methodological guidelines, that provide important considerations for deciding on clinical trial conduct. These considerations do not amount to more stopping rules. Instead, they are principles that address the complexity of interpreting and responding to interim data, based on a broad range of epistemic and ethical factors. While they are not stopping rules, they would assist a Data Monitoring Committee in judging its position with regard to necessary precautionary interpretation of interim data. By vindicating a framework that accommodates a wide range of approaches to statistical inference in one important setting (clinical trials), my results pose a serious challenge for any approach that advocates a single, universal principle of statistical inference.
Preface

A version of the case study in section 4.2 (Early stop due to efficacy) has been published: Stanev, Roger (2011) “Statistical decisions and the interim analyses of clinical trials”, Theoretical Medicine and Bioethics, Vol. 32:61-74.


A version of the case study in section 4.4 (Early stop due to harm) has been accepted for publication: Stanev, Roger (2012b) “Modeling and Evaluating Statistical Decisions of Clinical Trials”, Journal of Experimental & Theoretical Artificial Intelligence, (forthcoming).
# Table of contents

Abstract .................................................................................................................. ii

Preface .................................................................................................................. iii

Table of contents ................................................................................................... iv

List of tables .......................................................................................................... ix

List of figures ......................................................................................................... x

List of abbreviations ............................................................................................. xi

Acknowledgements .............................................................................................. xiii

Dedication ............................................................................................................. xiv

Introduction ......................................................................................................... 1

Chapter 1. Stopping rules .................................................................................... 5

1.1. Introduction ................................................................................................... 6

1.2. The problem of stopping rules .................................................................... 9

1.3. Why stopping rules are irrelevant for Bayesians ....................................... 11

1.4. Bayes vs. error-statistics: historical context for the debate ....................... 18

1.5. The problem of early stopping of RCTs ...................................................... 24

1.5.1. The inadequate reporting of RCTs .......................................................... 25

1.5.2. The skepticism about RCTs in developed countries ............................. 27

1.5.3. The public mistrust of RCTs in developing countries ............................. 28

1.5.4. Data monitoring committee ..................................................................... 29
1.6. Why early termination of RCTs can be a problem for ES .......................... 31

1.6.1. Balancing early vs. late-term effects....................................................... 32

1.6.1.1. The upshot for ES ........................................................................ 41

1.6.2. Rationing and the allocation of statistical boundaries in RCTs .............. 43

1.6.2.1. The upshot for ES ........................................................................ 48

Chapter 2. Disagreements over early stopping of RCTs ..................................... 55

2.1. Introduction ................................................................................................. 56

2.2. The controversy over AZT in asymptomatic HIV-infected individuals ....... 60

2.2.1. Scientific uncertainties .......................................................................... 67

2.2.1.1. Uncertainties surrounding HIV virology ........................................ 68

2.2.1.2. Uncertainties surrounding the human immune system .................... 70

2.2.1.3. Uncertainties surrounding surrogate markers ................................. 72

2.2.1.4. Scientific uncertainty: some general observations .......................... 76

2.2.2. Difference in judgments about RCT monitoring policies ..................... 77

2.3. Toxoplasmic encephalitis (TE) in HIV-infected individuals ..................... 81

2.3.1. Differences in judgments over the standards of proof in RCTs ............. 84

2.4. Vitamin A as prevention of HIV transmission from mother-to-child through breastfeeding ................................................................. 98

2.4.1. Differences in judgments about the balance of long-term versus short-term treatment effects ................................................................................. 103

2.5. Conclusion.................................................................................................. 105
Chapter 3. Qualitative framework

3.1 Preliminaries

3.1.1. Basic requirements

3.1.2. Restrictions

3.2 Overview of the framework

3.2.1. Classification

3.2.1.1. Efficacy

3.2.1.2. Harm

3.2.1.3. Futility

3.2.2. Identification

3.2.3. Reconstruction

3.2.4. Evaluation

3.2.4.1. Comparing DMC decisions

3.3 Classification and identification

3.3.1. Early stop due to efficacy

3.3.1.1. E-A cases

3.3.1.2. E-B cases

3.3.2. Early stop due to futility

3.3.2.1. F-A cases

3.3.2.2. F-B cases
3.3.3. Early stop due to harm ................................................................. 147
3.4 Reconstruction .................................................................................. 152
3.4.1. Loss function .................................................................................. 155
3.4.2. Decision criteria ............................................................................. 160
3.4.2.1. Maximum expected utility (MEU) .................................................. 161
3.4.3. Loss function, factors and decision criteria: a paradigm case .......... 163
3.5 Evaluation .......................................................................................... 165
3.5.1. Policy stage and the choice of stopping rule .................................... 166
Chapter 4. Quantitative framework ............................................................... 172
4.1 Decision-theoretic framework (DTF) ....................................................... 173
4.1.1. Observation model ........................................................................... 176
4.1.2. Parameter space $\Theta$ .................................................................... 178
4.1.3. Set $A$ of basic (allowable) actions .................................................. 178
4.1.4. Decision monitoring rule $\delta(y):Y \rightarrow A$ and the set $\Delta$ of such rules ......................................................................................... 179
4.1.5. Loss function ................................................................................... 185
4.1.6. Decision criteria .............................................................................. 187
4.2 Early stop due to efficacy ..................................................................... 188
4.2.1. Simulation ....................................................................................... 188
4.2.2. Discussion ..................................................................................... 194
4.3 Early stop due to futility ..................................................................... 197
4.3.1. Simulation .............................................................. 197
4.3.2. Discussion ............................................................. 204

4.4 Early stop due to harm ..................................................... 206
4.4.1. Simulation .............................................................. 207
4.4.2. Evaluation .............................................................. 212
4.4.2.1. Comparison of stopping rules using a loss function ........ 215
4.4.3. Discussion .............................................................. 217

Chapter 5. Toward a pluralistic framework of early stopping decisions in RCTs .... 220

5.1 Summary ................................................................. 221
5.2 Some objections ......................................................... 226
5.3 Some achievements ..................................................... 231
5.3.1. Contribution to philosophy of medicine ....................... 231
5.3.2. Contribution to philosophy of statistics ....................... 233
5.4 Future projects .......................................................... 234

Final Remarks .................................................................. 236

Bibliography .................................................................... 237

Appendix ........................................................................ 250

Appendix A: Likelihood Principle Example ................................ 250
Appendix B: Additional Graphs and Tables ............................. 252
List of tables

**Table 1.1**  Simplified free throw experiment ................................................. 13

**Table 3.1**  Types and summary of early stopping decisions .......................... 121

**Table 3.2**  RCT monitoring decision with outcomes in abstract .......................... 153

**Table 3.3**  Representation of decision outcomes and their payoffs ..................... 154

**Table 3.4**  Example of an ‘ethical’ loss function .............................................. 159

**Table 3.5**  Example of a ‘scientific’ loss function ............................................. 159

**Table 3.6**  Contextualized loss function .......................................................... 164

**Table 4.1**  Interim data ...................................................................................... 199

**Table 4.2**  Predictive power at interim .............................................................. 201

**Table B.1**  Overall type I error according to $p((y_2 - y_1) | H_0)$ ............................ 252
List of figures

Figure 3.1  Representation of the framework’s main elements ............................. 115
Figure 3.2  Stringency and efficacy levels as weighing dimensions ........................... 138
Figure 4.1  U.S. decision monitoring rule as a function of \((y_2 - y_1)\) .......................... 182
Figure 4.2  Concorde decision monitoring rule as a function of \((y_2 - y_1)\) ................. 185
Figure 4.3  Weighted-losses of stopping vs. continuing ........................................... 190
Figure 4.4  Two rules under different loss functions .................................................... 194
Figure 4.5  Cut-offs CPs as a function of utility .......................................................... 203
Figure 4.6  Comparison of two stopping rules with 8 interims .................................... 210
Figure 4.7  Comparison of two stopping rules with 3 interim ....................................... 211
Figure 4.8  Comparison of expected pay-offs ............................................................... 219
Figure B.1  Example of frequentist stopping rule and its error probabilities ............... 252
Figure B.2  Three families of stopping rules implemented and plotted ....................... 253
Figure B.3  Stopping rules and two ways of computing conditional power .................. 253
Figure B.4  Example of an early trend of efficacy and a stopping rule ....................... 254
Figure B.5  Example of an early trend of harm and a stopping rule ............................ 254
List of abbreviations

ACTG    AIDS Clinical Trials Group
AIDS    Acquired Immune Deficiency Syndrome
AZT     Azidothymidine (also known as Zidovudine)
CD4     Cluster of Differentiation-4
CONSORT Consolidated Standards of Reporting Trials
CP      Conditional Power
DMC     Data Monitoring Committee
DSMB    Data and Safety Monitoring Board
DTF     Decision Theoretic Framework
EBM     Evidence Based Medicine
ES      Error Statistics
FDA     U.S. Food and Drug Administration
GCTS    U.S. FDA Guidance for Clinical Trial Sponsors
HIV     Human Immunodeficiency Virus
ICH     International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use
IRB     Independent Review Board
JAMA    Journal of the American Medical Association
NIAID   U.S. National Institute of Allergy and Infectious Diseases
NIC     U.S. National Institute of Cancer

1 Protocol used for the identification of white blood cells.
NIH  U.S. National Institutes of Health
NP   Neyman-Pearson
RCT  Randomized Controlled Trial
TE   Toxoplasmatic Encephalitis
Acknowledgements

My thanks go first of all to my supervisor Paul Bartha who has been instrumental in making this dissertation a reality. Such success as I have had so far in examining, articulating and addressing several of the issues that are the subject of this dissertation is due in large part to his guidance, insight and patience.

For helpful comments and criticism of various parts of this dissertation I thank Alan Richardson and Hubert Wong. For less direct but nevertheless helpful input I thank John Beatty, Scott Anderson, Christopher Stephens, Leslie Burkholder, Lang Wu, Jeremy Simon, Rachel Ankeny, Sylvia Berryman, Yuichi Amitani, Roger Clarke, Eric Nadal, Jie Tian, Alirio Rosales, Christopher French, Andrew Inkpen, Julian Reiss, David Teira, Jan Vandenbroucke, Teddy Seidenfeld, Jesús Zamora-Bonilla, Peter Danielson, Alex Mesoudi, and Jan-Willem Romeijn.

I must also record my intellectual debt to Deborah Mayo, Richard Burian, Joseph Pitt, and Andrea Woody for originally introducing me to the subject of history and philosophy of science and technology.
Dedication

For Dana and Maximilian
Introduction

For ethical, epistemic and financial reasons, data are monitored and reviewed periodically during the course of a trial by a data monitoring committee (DMC). The two main responsibilities of a DMC are to: (i) safeguard the interests of study participants by assessing the safety and efficacy of interventions, and (ii) keep vigilance over the conduct of the trial so that the trial meets the need for objectivity and reproducibility of results. These imply providing recommendations about stopping, modifying or continuing the trial on the basis of interim data analyses (Ellenberg, Fleming, and DeMets 2003).

Because a clinical trial is a public experiment, a rationale is required for its design and interim monitoring decisions. The primary goal of this dissertation is to develop a framework (first informal, then formal) for reconstruction and retrospective assessment of DMC early stopping decisions. This goal is amply justified because of the problem of lack of transparency in DMC reports; namely, RCT results continue to be published in top medical journals without sufficient detail that would allow their proper assessment. The dissertation thus proposes a decision theoretic framework that makes the articulation and justification of clinical trials decisions more explicit. The framework provides ways for comparing statistical monitoring rules based on a set of ethical and epistemic constraints. It allows disparate statistical monitoring methods to emerge in their appropriate context, while also offering important considerations for deciding on trial conduct based on a range of factors.

2 Data Monitoring Committees are also known as Data and Safety Monitoring Boards.
A secondary goal of the dissertation is to provide a decision theoretic framework (first informal, then formal) for discourse by the DMC members. The “framework for discourse” is to assist the DMC with clinical trial planning by highlighting epistemic and ethical factors that the DMC must or should take into account. My hope is that the proposed framework may also be used prospectively—for design of experiments—and the “framework for discourse” idea goes part of the way.

Finally, I should add that even though the proposed decision theoretic framework is largely concerned with statistical applications in clinical trials, my goal in proposing the framework includes the identification of general lessons and principles underlying the statistical procedures clinical researchers have come to value on pragmatic grounds.

The dissertation is structured as follows. Chapter 1 introduces the problem of stopping rules for RCTs (randomized controlled trials), and the principal treatments of the mainstream views from philosophies of statistics. The chapter gives particular attention to the reason why these rules can be a problem for Bayesians, and presents a strong case for ES (error statistics) over Bayesian statistical inference. The chapter then explores the problem of early termination of RCTs. In particular, it offers reasons why early termination of RCTs can be a problem for ES, and provides a strong case for Bayesian statistical inference over ES.

Chapter 2 explains the increased dissatisfaction, among clinical researchers and epidemiologists, over the proper use and interpretation of classical statistical methods
applied to clinical trials. Dissatisfaction with practice and the need for guidelines is motivated by three case studies involving HIV/AIDS studies. The first case-study is a pair of treatment trials concerning two data monitoring committees arriving at two different conclusions: the U.S. AIDS Clinical Trial Group (Protocol 019) study of zidovudine (AZT) in asymptomatic HIV-infected individuals, and a similar study run by the European Concorde group. The issue here is \textit{early stopping due to treatment efficacy}. The second case-study is a trial involving an \textit{early stop due to futility}. This was an RCT designed to determine whether pyrimethamine would reduce the incidence of a brain-infection in HIV-infected individuals; the DMC recommended that the trial be stopped, and the chair that it be continued. The third case-study is about a trial that examined vitamin A supplementation for the prevention of mother-to-child transmission of HIV through breastfeeding. This was an instance of an \textit{early stop due to harm}, as the vitamin A appeared to have the opposite effect, namely, increasing the chances of HIV transmission from mother to child. These case studies are used to identify and explain important factors in early stopping decisions, in a qualitative way.

In Chapter 3, I articulate a comprehensive, and mostly qualitative, approach to the evaluation of early stopping decisions of RCTs. This is a non-technical treatment of the theoretical framework for decision-making.

Chapter 4 introduces the decision theoretic framework (DTF). The chapter includes the different formal elements constituting the DTF (i.e., observation data model, possible states of nature, set of available actions, loss functions, set of monitoring rules), and
provides examples of computer simulations that implement the framework to compare alternative stopping decisions and data monitoring policies.

In Chapter 5 it is argued—in light of previous discussions—that different families of statistical inference have different strengths and that they can coexist in serving the practical demands for designing and monitoring clinical trials.
Chapter 1. Stopping rules

This chapter begins by introducing the problem of stopping rules and the principal treatments of the mainstream views from philosophies of statistics. It devotes particular attention to the reason why these rules can be a problem for Bayesians, and reviews why they furnish a strong case for error-statistics (ES) over Bayesian statistical inference. The chapter then explores the problem of early termination of RCTs. In particular, it offers reasons why early termination of RCTs can be a problem for ES, and provides a strong case for Bayesian statistical inference over ES. The main reason is that even though RCTs are experiments designed to assess evidence for efficacy, other issues which are assessed during the trial, such as the possibility of side-effects and the rationing of scarce resources, play an important role in assessing evidence for efficacy. These factors are best captured by a Bayesian decision framework.
1.1. Introduction


At the most basic level, a stopping rule is a rule that dictates when to stop an experiment and start inferring from the data. It can be characterized as a “formal definition of the circumstances under which a trial would be stopped depending on the results obtained.” (Grant et al 2005, xi) Sometimes stopping rules are referred to as ‘approaches’ to early termination, and sometimes simply as ‘stopping procedures.’ Clinical researchers at times refer to stopping rules as “guidelines,” i.e. guidelines helping researchers to make decisions during the course of an experiment. Yet at other times stopping rules are referred to by specific names, e.g. the Pocock rule, or O’Brien-Fleming stopping boundaries.

DeMets et al (1999) offer a more liberal characterization. Stopping rules are “statistical monitoring procedures assisting the DSMB [Data and Safety Monitoring Board] in repeatedly inspecting outcome data while controlling for false claims for treatment
benefit or harm ... although useful, they are not absolute rules, but objective guidelines to aid DSMB decisions.” (DeMets et al 1999, 1983)

Whether interpreted strictly or as guidelines, stopping rules are typically formulated in technical terms that derive from the error-statistical approach to statistical inference. Here is a first example of a stopping rule cited in a clinical trial: “An interim analysis [is] planned when half of the patients had been enrolled. A difference of $p \leq .001$ (two sided) in [disease-free survival] between the two groups [is] considered sufficient to stop patient accrual.” (Frustaci et al 2001, 1240) Referring to this particular stopping rule, here is an oncologist commenting on it:

Trials using the group sequential design are common in the published literature. One example of its use is a trial reported by Frustaci and others, in which 190 patients were to be accrued to detect a 20% difference in 2-year disease free survival (60% on the adjuvant chemotherapy treatment arm vs. 40% in the control arm undergoing observation alone). An interim analysis was planned after half of the patients were accrued with the stopping rule in terms of adjusted $P$ value. The trial was stopped as this criterion was met, thereby saving 50% of the planned patient accrual. The observed difference was found to be 27% (72% on the treatment arm vs. 45% on the control arm), 7% higher than what was hypothesized initially at the design stage. Therefore, the risk of treating additional patients with suboptimal therapy was avoided. (Mazumdar 2004, 526)

The commentator identifies a typical reason for early stopping (patient welfare); the stopping rule itself is formulated in terms of $p$-value in patient survival. The example also follows what is, according to DeMets et al, the most frequently used approach, “the group sequential design, which requires more extreme result for early termination than conventional criteria such as $p < 0.05$. How extreme these results are depends on the
specific method, but the approach of O'Brien-Fleming, as implemented by Lan and DeMets, is widely used.”³ (DeMets et al 1999, 1984)

The following HIV/AIDS studies provide further example of stopping rules:

An interim analysis was planned when the number of primary endpoints reached half of the expected total, and a stopping rule was determined a priori, following the O'Brien and Fleming principle. This interim analysis was done by an independent board who recommended study termination on March 17, 1998, because results showed a 49% reduction in severe events in the co-trimoxazole group (p=0.001). (Anglaret et al 1999, 1465) ⁴

This stopping rule involves the use of the method of Pocock dividing the type II error rate (beta=0.14) equally among three planned interim analyses. Given the assumptions of equal numbers of person-years of follow-up in all the study groups, the 1.14 boundary is the limit of the point estimate for the hazard ratio that would lead to the rejection of the null hypothesis as the final analysis. (Gulick et al 2004, 1852) ⁵

Today, there is an array of statistical approaches to stopping rules in clinical trials. While according to several epidemiologists and biostatisticians no single approach addresses all of the possible issues that DMCs may face, the diverse approaches do provide a useful set of tools to assist DMCs in deliberating about interim data analyses. Existing approaches to statistical stopping rules in RCTs can be classified into different groups, with the main approaches often cited as the following: Sequential group, Bayesian, and

³ The authors add that such an approach “requires very strong early evidence of a treatment difference (i.e., a very small p value), and slackens that criterion as the trial progresses, nearly returning to the boundary of conventional significance level at the final analysis. (…) These methods appeal because they are conservative in the final scheduled analysis, require negligible increase in sample size to allow for repeated looks at data, and allow for unscheduled analyses.” (ibid 1984)
Curtailment approach. (cf. Ellenberg, Fleming, DeMets 2003) I shall say more on the difference among such approaches to stopping rules in section 1.2.2 when examining why they are a problem for ES.

1.2. The problem of stopping rules

As a simple illustration of the relevance of stopping rules to statistical inference, and in particular to the contrast between Bayesian and ES approaches, consider the following Free Throw Experiment. Suppose I tell you that I am good at playing basketball. By good, I mean that I can free-throw baskets with a relative frequency of successful conversions that exceeds the relative frequency expected by chance alone—assuming here that chance alone means the hypothesis H of 50% successful conversions. Let’s say this excess is a number sufficient to attain a 5% significance difference. Suppose further that, despite your disbelief in my free-throwing ability, you agree to let me prove otherwise. You give me an official size ball and point me to a basketball court. 40 minutes later I come back with the following data: 56% successful free-throws, a statistically significant result. I guarantee that all my throws were clean. Would you find the data to be evidence in support of my ability? If you say yes, your intuition agrees with Bayesian intuitions.

Would you find it relevant to know that after 50 free-throws, having failed to reach enough successful conversions to reach the 5% statistically significant difference initially agreed upon, I went on to make 30 more throws; failing yet again, I went on to another 30, and then 20, until stopping at a total of 180 throws, when I finally attained a statistically
significant result? If you find this information relevant, i.e., if you find the lack of a fixed stopping rule relevant for assessing my ability, then your intuition might be at odds with the earlier Bayesian intuition. For ES, it matters that a try-and-try-again procedure was in place when evaluating the results.

Although the relationship between Bayesian accounts and stopping rules can be complex (cf. Steel 2003), in general, Bayesians regard stopping rules as irrelevant to what statistical inference should be drawn from the data. This position clashes with classical statistical accounts of inference. For classical statistics, stopping rules do matter to what inference should be drawn from the data. “The dispute over stopping rules is far from being a marginal quibble, but is instead a striking illustration of the divergence of fundamental aims and standards separating Bayesians and advocates of orthodox statistical methods.” (Steel 2004, 195)

The fundamental problem of stopping rules, then, is, in the first place, whether they ought to play an essential role in experimental design and statistical inference. If they should play such a role, the problem then bifurcates depending on one’s approach to statistical inference. For a Bayesian, the problem is how to make sense of stopping rules. For an error-statistician, the problem is about the best choice of stopping rule.

6 Steel (2003) argues for the possibility of a Bayesian account to make stopping rules relevant, given some special confirmation measure, but at the cost of rejecting two important principles: the likelihood and the sufficiency principle.

7 Likelihoodists side with Bayesians on the subject of stopping rules, i.e., stopping rules should not matter for an account of statistical evidence. This is despite important differences between Likelihoodist and Bayesian notions of evidence. See Sober (2008, 72-78) for more.
In the next section, I explain why, on a Bayesian account of evidence, stopping rules are irrelevant, and why this leads to a fundamental clash with ES approaches.

1.3. Why stopping rules are irrelevant for Bayesians

For a Bayesian account of evidence, knowledge of the stopping rule should make no difference to your appraisal of the data. For Bayesians, evidence used in the appraisal of a hypothesis H must be expressed in a statement. Here are three ways of describing the result e of the basketball trial:

\( e_1: \) 101 successful and 79 failed throws, in a trial designed to have 180 free throws.
\( e_2: \) 101 successful and 79 failed throws, in a trial designed to end after a statistically significant result is attained (in this case 101 successful throws).
\( e_3: \) 101 successful and 79 failed throws, in a trial designed to end when it starts to rain.

In technical terms, the reason why all three descriptions of the trial provide exactly the same evidential support for H is that their likelihoods \( L(H \mid e) \) are proportional to one another. ⁸ Since all Bayesian inference follows from the posterior probabilities, the

---

⁸ When it comes to referring to ‘likelihoods’ not every philosopher of science is careful enough about what he or she means by it. I should remind the reader that the likelihood function, likelihood principle, law of likelihood, likelihood test, likelihood ratio test, and the method of maximum likelihood are all very different things. For the purposes of understanding the debate about stopping rules, the important concepts are as follows: (1) The likelihood function, \( L(\theta \mid x) \), which is a conditional probability function—\( P(x \mid \theta) \)—considered as a function of its 2nd argument (i.e. \( \theta \), the parameter of interest) with its 1st parameter (i.e. \( x \), the outcome) fixed—including any other function proportional to such function (a.k.a. an equivalence class); thus \( L(\theta \mid x) = \alpha P(x \mid \theta) \). Even though in non-technical
important point here is that, if you assume the likelihood principle\(^9\) no matter what your prior probability is for \(H\), the likelihood should contain all the sample information that matters to a Bayesian. If we have two sets of data \(x_1\) and \(x_2\), and there is a constant \(c > 0\) such that \(P(x_1 | H) = c P(x_2 | H)\) for all values of \(H\), then any two posterior probabilities \(P(H | x_1)\) and \(P(H | x_2)\) are equal.

In the basketball example, \(e_1\) reports the number of successful free-throws among a fixed number of throws \(n\), which is modeled by a binomial;\(^10\) that makes \(P(e_1 | H)\) depend on factor \(C(n, r)\). However, that factor cancels out in the application of Bayes's theorem (Howson and Urbach 1996, 365). The situation is exactly the same with \(e_2\), except that, \(e_2\) says that we are counting the number of throws \(n\) needed until the \(r\)th successful free-throw, which is modeled by a negative-binomial;\(^11\) that makes \(P(e_2 | H)\) depend on the factor \(C(n-1, r)\). Once again, this cancels out in applying Bayes's theorem. Thus, any two descriptions of the trial for which the likelihoods \(P(e | H)\) are proportional to one another yields the same posterior probability for \(H\), according to Bayes’s theorem (see Appendix A for a proof).

---

\(^{9}\) An explicit statement of the likelihood principle: “all information about \(θ [H]\) obtainable from an experiment is contained in the likelihood function for the actual observation \(x [e]\).” Berger and Wolpert (1984, 19)

\(^{10}\) \(C(n,r)θ^r(1-θ)^{n-r}\) where \(n = 180\) (fixed number of trials) and \(r = 101\) (number of success).

\(^{11}\) \(C(n-1,r)θ^r(1-θ)^{n-r}\) where \(n = 180\) (number of trials, i.e. successful + failed attempts) and \(r = 101\) (number of success).
To illustrate what I just explained about likelihoods and their relation to the different descriptions of a trial providing exactly the same evidential support for H, consider the following simplified version of the *Free Throw Experiment* consisting of a shorter experiment, i.e. \(n = 3\) free throws. Let \(X = \{0, 1, 2, 3\}\) represent the number of successful conversions, and \(H\) the probability of successful conversion with \(\theta = \{0.5, 0.4\}\). Consider two experiments \(E_1\) and \(E_2\), both of which consist of observing \(x_1\) and \(x_2\) with the possible outcomes \(X\) and the same \(\theta\) value, but with the following difference: \(E_1\) is an experiment modeled with a binomial distribution (a fixed number of free throws, \(n=3\)), and \(E_2\) is an experiment modeled with a negative binomial (where \(n\) is *not* fixed, but happens to be 3), i.e. the probability that the \(r\)th successful free throw \((X = x)\) occurs when the number of free throws equals 3. See Table 1.1 below.

\[
E_1: \text{modeled as a binomial}^{12} \\
E_2: \text{modeled as a negative binomial}^{13}
\]

<table>
<thead>
<tr>
<th></th>
<th>(x_1)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>(L(\theta=0.5</td>
<td>x_1))</td>
<td>0.125</td>
<td>0.375</td>
<td>0.375</td>
</tr>
<tr>
<td>(L(\theta=0.4</td>
<td>x_1))</td>
<td>0.216</td>
<td>0.432</td>
<td>0.288</td>
</tr>
<tr>
<td><em>Ratio</em></td>
<td>0.579</td>
<td>0.868</td>
<td>1.302</td>
<td>1.953</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(x_2)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>(L(\theta=0.5</td>
<td>x_2))</td>
<td>0.125</td>
<td>0.250</td>
<td>0.125</td>
</tr>
<tr>
<td>(L(\theta=0.4</td>
<td>x_2))</td>
<td>0.216</td>
<td>0.288</td>
<td>0.096</td>
</tr>
<tr>
<td><em>ratio</em></td>
<td>0.579</td>
<td>0.868</td>
<td>1.302</td>
<td>1.953</td>
</tr>
</tbody>
</table>

**Table 1.1** Free throw experiment consisting of 3 free throws, with experiment \(E_1\) modeled as a binomial and experiment \(E_2\) modeled as a negative binomial.

---

12 \(C(n, x)\theta^x(1-\theta)^{n-x}\) where \(n\) is the number of free throws (fixed) and \(x\) is the number of successful free throws.

13 We are observing a sequence of free throws until a predefined number of successful conversions has occurred. \(X=1\) in our \(E_2\) table refers to 1 success after 2 misses (failures). What is the probability of taking \(k=3\) trials to get, 1 success (\(k-r=1\)), i.e., having one success after two failures (\(r=2\))? The model is \(C(k-1,k-r)(1-p)^rp^r\). According to the notation, \(k=3\), \(r=2\) and \(p\) is probability of success. For \(p=0.4\) \(P(X)=0.096\), for \(p=0.5\) \(P(X)=0.125\).
Suppose $E_1$ is performed and $x_1 = 1$ is observed. The likelihood principle states that the information about $\theta$ depends only upon the likelihoods $\left( L(\theta=0.5| x_1), L(\theta=0.4| x_1) \right)$, in this case $(0.375, 0.432)$, with ratio $(0.375/0.432) = 0.868$. Now consider the same data from the perspective of $E_2$. Suppose $x_2 = 1$ is observed, that is, I had to throw 3 baskets until I made one of them. Given that the likelihoods are $\left( L(\theta=0.5| x_2), L(\theta=0.4| x_2) \right) = (0.250, 0.288)$, their ratio is $(0.250/0.288) = 0.868$, which means that $x_2 = 1$ provides the same evidence about $H$ as does $x_1$ via $E_1$. The likelihood ratio is the same in both experiments. Note that the likelihood ratios for the two experiments are the same when either 0 or 2 successful conversions are observed, as well as when 3 successful free throws are observed. Thus, regardless of which experiment is performed, for Bayesian inference, because of the likelihood principle and the proportionality of the likelihoods (and hence the equality of the likelihood ratios), the exact same conclusion about $H$ should be reached for a given outcome ($x_1 = x_2$), whether the experiment is modeled as a binomial or a negative binomial. This illustrates why, for Bayesians, the stopping rule does not matter.

From the point of view of error-statistics, $x_1$ and $x_2$ tell us different things. To see why this is the case, consider a statistical test that states: accept $\theta=0.5$ if observing either 2 or 3 successful free throws and reject $\theta=0.5$ (i.e. accept $\theta=0.4$) if observing otherwise (either 0 or 1 successful free throws). The statistical test has error probabilities (of type I and type II error, respectively) 0.5 and 0.352 for $E_1$, and 0.375 and 0.16 for $E_2$, thus reflecting different properties. Given the same outcome ($x_1 = x_2$), the same statistical test

---

14 Type I error = $(0.375+0.125) = 0.5$, and type II error = $0.288+0.064 = 0.352$; thus power = 0.648.
15 Type I error = $(0.25+0.125) = 0.375$, and type II error = $0.096+0.064 = 0.16$; thus power = 0.840.
would report different evidence from the two experiments. This exemplifies the reason why, for ES, the way in which the experiment is conducted—either as a binomial (E₁) or a negative binomial (E₂)—is relevant for evaluating H—contrary to Bayesian inference. This consequence was already understood by Edwards, Lindman and Savage when they wrote:

The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money or patience. (1963, 193)

The Bayesian response to this predicament is that stopping rules reflect private intentions of the experimenter. These intentions—which we have no access to—should not matter for an account of evidence (Savage 1962, 18). For scientists—ideal scientists—would not regard such personal intentions as proper influences on the support which data give to a hypothesis (Howson and Urbach 1996, 212). As Howson and Urbach put it:

(...) was the intention to draw a sample of size n, or to continue sampling until a simultaneously spun coin produced three heads or until boredom set in, or what? ... it seems most implausible and unrealistic that anyone's confidence in an estimate would or should in general be influenced by a knowledge of such facts about the experimenter's state of mind. The subjective-confidence interpretation of confidence-interval statements is in conflict with these, to our mind compelling, intuitions. (1996, 241)

In technical terms this difference between Bayesians and error-statistics means that stopping rules do not impact likelihoods for Bayesians, but they impact a procedure’s error probabilities for statistical accounts of inference, such as Mayo’s (1996) error-statistics (ES).
For ES, stopping rules as test specifications cannot be relegated to a particular experimenter’s intention. Stopping rules as part of the experimental design considerations impact error rates, which are operating characteristics of the particular test procedure. A try-and-try-again procedure with an optional stopping point could lead to high or maximal overall significance levels. Stopping rules affect both the reported error rates and what should be expected to happen in subsequent repetitions of the experiment. That is, they are not just important for correctly reporting what occurred in obtaining the experimental result, but also in determining what should be expected to occur in subsequent repetitions of the experiment performed (Mayo 1996, chapters 9 and 10). Other experimenters seeking to check or repeat the results observed run the risk of being misled when stopping rules are ignored. And for this reason, it “seems natural, and only to be expected” that stopping rules should affect the assessment of the hypothesis H of interest (Gillies, 1990, 94).

In contrast to Bayesians, Mayo and Kruse (2001) argue that to the error statistician the situation is exactly the reverse of what we have with the likelihood principle. That is, “the stopping rule is relevant because the persistent experimenter is more likely to find data in favor of H, even if H is false, than one who fixed the sample size in advance.” (2001, 389) As put years before by Armitage, in responding to Savage (at the 1959 forum):
My own feelings go the other way. I feel that if a man deliberately stopped an investigation when he had departed sufficiently far from his particular hypothesis, then 'Thou shall be misled if thou dost not know that'. If so, probability methods seem to appear in a less attractive light than frequency methods, where one can take into account the method of sampling. (1962, 72)

For Mayo, error statistical philosophy insists that data only provide genuine or reliable evidence for H if H survives a severe test—more on this criterion in the next section. Severity of the test, as a probe of H, depends upon the test’s ability to reveal that H is false when it is. Therefore, according to error statistics, H is not considered put to a stringent test when a researcher allows trying and trying again until the data are far enough from the null-hypothesis to reject it in favor of H. (Mayo and Kruse 2001)

Bandyopadhyay and Brittan (2006) have proposed a Bayesian account of a severe test, while providing quantitative comparisons between error-statistics severity and their Bayesian severity counterpart. According to their account, T counts as a severe test for hypothesis H just in case it yields data x such that: (i) $P(H| x)$ is high, i.e., the posterior probability of H is high—a.k.a., the confirmation condition, and (ii) the ratio of the likelihoods, $P(x| H)/P(x| H')$ is high, where H and H’ are competing hypotheses—a.k.a., the evidential condition. (2006, 263-264) By making a distinction between evidence for H and confirmation of H, Bandyopadhyay and Brittan argue that their Bayesian account is to be preferred based on quantitative comparisons made with respect to standard diagnostic tests in medical sciences.
1.4. Bayes vs. error-statistics: historical context for the debate

The difficulty in attempting either to unify or to supplement either account of statistical inference (as with Bandyopadhyay and Brittan) is that Bayesianism and ES reflect profound differences in how probabilities, as quantitative measures of uncertainty, are interpreted and used. This should be no surprise given the fundamental difference in their aims. The main goal of a Bayesian approach is to say how well (and when) evidence e supports (and confirms) hypothesis H. What one learns about a hypothesis H from evidence e is measured by the conditional probability of H given the evidence e. This conditional probability is computed via Bayes’s theorem. Whether we are referring to logical Bayesians or subjective Bayesians, “the aim of the Bayesian school is to find methods for calculating P(H|e).” (Gillies 2000, 36) According to Spiegelhalter et al (2004), in spite of differences among Bayesian schools, the three concepts that distinguish Bayesian from conventional methods are: “coherence, exchangeability and the likelihood principle.” (p 113) Howson and Urbach’s type of Bayesianism aims at an account of evidential support in terms of how accepting evidence e as true affects an individual’s degree of belief in some hypothesis H. One salient point is that “it is an inductive logic” (Howson 1997b, S187) providing a “theory of (degree of belief) consistency ... and inference from given data and prior distribution of belief to a posterior distribution.” (Howson & Urbach 1996, 419)

16 A common formulation is: P(H | e) = P(e | H)P(H) / P(e)
17 From this school of Bayesianism where probabilities are personal degrees of actual belief, we find also: Ramsey, De Finetti, and R. Jeffrey. More on this view later.
ES, on the other hand, aims at methods with certain desirable operating characteristics, treating probabilities as frequencies. These probabilities are investigated by using methods with certain operating characteristics. One example is the probability that a certain test will reject a hypothesis H erroneously—a.k.a. an error probability. There is no single “logic” underlying measures of evidence. There is no unified scheme for relating beliefs or hypotheses to data. Instead there are only methods of various types with certain operating characteristics. The preference is for an assortment of methods that enable users to detect deviations or discrepancies of interest and “for communicating results in a form that promotes their scrutiny and their extension by others.” (Mayo 1997, S198)

The dispute often plays out with Bayesians criticizing classical statistical methods on the grounds that their approach conflicts with intrinsic and appealing norms of rationality. In the meantime, the response from classical statistics is to maintain that stopping rules as part of a pre-designed strategy are necessary for assessing the reliability of our statistical inferences, adding that if to attain this desired standard of reliability requires rejecting some alleged norms of rationality, then “so much worse for those norms.”

Giere observes, “the differences between Howson and Urbach’s conception of scientific reasoning and that of Mayo are not mere technicalities. They are as deep and profound as theoretical differences ever get.” (1997, 181) According to Giere, the Bayesian approach can be seen as an up-to-date version of Carnap’s program for inductive logic, without its foundationalist ambitions. By contrast, ES can be seen as a development of Reichenbach’s inductive approach, also without foundationalist ambitions. This contrast
might illustrate the differences in how probabilities are used and interpreted in these competing methodologies.

Reichenbach advocates the view of probability as the limit of a relative frequency (1949). For him, “the merits of this interpretation are obvious; what we have to study are its difficulties.” (1956, 236) The two essential difficulties, according to him are: probability assignments to “single cases”, and the problem of induction. With respect to the former, the meaning of it is “harmless”, a “useful linguistic habit” that is both “fictitious” and capable of “leading to correct evaluations of future events”. (1956, 240) As for the latter problem, a probability statement is a predictive statement and best interpreted as a posit—a statement which we treat as true although we do not know whether it is so.

18 Under this interpretation, the assertion that the probability of an outcome \( s \) from a trial \( X \) is \( r \), simply means that, in a very large sample of trials of the sort \( X \) (say \( n \) runs of trial \( X \)), there are \( m \) outcomes \( s \), and the limit of \( m/n \) equates to \( r \), i.e., \( P(s)=r \). With minor reservations, the view is similarly shared by other members of Reichenbach’s logical empiricist circle (Berlin Circle), such as Richard von Mises.

19 It is not clear what Reichenbach meant by a “harmless” and “useful linguistic habit”. He never addressed the issue directly. Although just a conjecture, perhaps by useful linguistic habit, Reichenbach meant something akin to what Edwards (1972) had said about the usefulness of abstractions, such as “the notion of a random choice” for understanding the frequentist interpretation of probability, i.e. “[t]he only concept of probability which I find acceptable”. I quote the pertinent passage from Edwards below:

The only concept of probability which I find acceptable is the frequentist one involving random choice from a defined population. Though the notion of a random choice is not devoid of philosophical difficulties, I have a fairly clear idea of what I mean by ‘drawing a card at random’. That the population may exist only in the mind, an abstraction possibly infinite in extent, raises no more (and no less) alarm than the infinite straight line of Euclid. I am only prepared to use probability to describe a situation if there is an analogy between the situation and the concept of random choice from a defined population. The extent of the analogy is a matter of opinion. The adequacy of an inference based on a probability argument is proportional to the adequacy of the analogy, or, as I shall call it, the probability model. In so far as the assessment of the model is subjective, so are probability statements based on it. (Edwards 1972, xv)

20 Just as mathematicians “speak of the infinitely distant point at which two parallels intersect” single-case-probability-statements “need not make the logician unhappy” since they can always account for useful linguistic habits—again, it is not entirely clear what Reichenbach might have meant by “useful linguistic habits.” At any rate, I part company with Reichenbach with regards to single-case-probability statements or events, since I do consider such cases as being problematic cases to frequentist interpretations of probability and not simply an inconvenient linguistic difference as Reichenbach suggests.

21 Once inductive inference is conceived as a method of posits with inferences reduced to induction by enumeration, it solves Hume’s problem, says Reichenbach, since it “must lead to success if success is ever attainable.” The
One way of understanding Reichenbach’s probability is to consider it as attempting to provide us with a “guide of life”. Where truth is not available, or when waiting until the future becomes fact is not a practical option either, the best we can do is to try to act in such a way that we will succeed as often as possible. In these situations, says Reichenbach, posits are our instruments of action. (1956, 246) But not any posit will do. Induction is the instrument of finding the best posit. The only function of a probability is in supplying a rating of the posit; “telling us how good the posit is.” The degree of probability has nothing to do with the truth of statements, but … “it functions as advice on how to select our posits.” (ibid., 240-41) If our aim is predicting, for instance when finding the limit of a frequency, then induction is the best means to attain that aim, given that there is a limiting frequency. And this is the picture that is left of the scientist when deciding too. The scientist, never knowing a priori whether his posits will come true, can tell you only his best posits. Reichenbach writes:

He is a better gambler, though, than the man at the green table, because his statistical methods are superior. And his goal is staked higher—the goal of foretelling the rolling dice of the cosmos. If he is asked why he follows his methods, with what title he makes his predictions, he cannot answer that he has an irrefutable knowledge of the future; he can only lay his best bets. But he can prove that they are best bets, that making them is the best he can do—and if a man does his best, what else can you ask of him? (1956, 249)

justification of induction rests on induction being “the best instrument of action known to us”. As space precludes further elaboration, the reader is referred to original publications (1949 and 1956, 240-49). I should note, however, that I am setting aside the problem of induction, while not conceding that Reichenbach has solved the problem satisfactorily.
Carnap (1950), on the other hand, distinguishes two notions of probability: ‘probability\textsubscript{1}’ (logical probability) and ‘probability\textsubscript{2}’ (statistical probability). Despite several points of agreement with Reichenbach, reflected by the inclusion of the frequency or statistical interpretation of probability, Carnap’s work emphasizes the semantic relation between statements and the role of logical probability in explicating and formulating a system of inductive logic.\textsuperscript{22}

Another way of looking at a continuation of Carnap’s inductive logic as “a theory of logical probability providing rules of inductive thinking” (Carnap & Jeffrey 1971, 7) is to consider the work of Carnap’s student Richard C. Jeffrey. Jeffrey is an advocate of Bayesianism and proponent of subjectivist models of learning where probabilities are best understood as subjective degrees of belief. According to Howson and Urbach, “Jeffrey inaugurated a new chapter of Bayesian research, into what has come to be called probability kinematics” (1996, 105) i.e., probabilities about consistent belief revision in the presence of new data. On the one hand, this is different from the narrowest kind of Bayesianism, which insists that conditionalization is the only rational way to change one’s mind. Jeffrey does not share this view. Neither is Jeffrey a logical or rationalistic Bayesian, “according to which there exists a logical and a priori probability distribution that would define the state of mind of a perfect intelligence, innocent of all experience”. Bayes, Laplace, Keynes, and Carnap might be located in this camp (Jeffrey 1992, 3). Jeffrey’s Bayesianism, of the radical probabilist type, runs contrary to a Carnapian project where evidence statements are “given” and treated as certain. Jeffrey sees probability all the way down, and

\textsuperscript{22} “Inductive logic is here understood as the theory of probability in the logical or inductive sense, in distinction to probability in the statistical sense, measured by frequencies.” (Carnap & Jeffrey 1971, 35)
therefore allows even for observation statements that are uncertain. It is this latter move by Jeffrey, according to Salmon, that avoids some of the undesirable features of earlier inductive logics. (1967, 142) 23

Moreover, Jeffrey, like Howson and Urbach, takes there to be identifiable “canons of good thinking that get used on a large scale in scientific inquiry at its best”; and also takes Bayesianism “to do a splendid job of validating the valid ones and appropriately restricting the invalid ones among the commonly cited methodological rules.” (Jeffrey 1992, 77)

As we have seen, there is a fundamental divide between two competing approaches to statistical inference. The Bayesian approach, according to some advocates, “avoids one of the most perverse features of classical statistical methods, namely, a dependence on stopping rules.” (Howson & Urbach 1996, 381) This feature is regarded as a “considerable merit” and a point in favor of their account. ES, on the other hand, regards stopping rules as an important factor in determining what inference should be drawn from the data. From an ES perspective, the Bayesian disregard for stopping rules is a critical point against that approach. So the choice between these two approaches depends crucially on the importance of stopping rules. This point has been widely acknowledged. Looking ahead, on this point, this dissertation will side with the ES perspective, in opposition to Bayesians.

____________________

23 According to Salmon, Carnap found his logical interpretation of probability “growing closer to the personalistic conception” where he approved the coherence requirement and its justification. Carnap differs from the personalists, “in his insistence upon many additional axioms beyond those of the mathematical calculus itself”. Salmon finds Carnap’s intuitive justifications of these additional axioms to be insufficient. (1967, 82-83)
The issue about stopping rules also vividly brings the Carnap vs. Reichenbach debate to the forefront—with respect to the sorts of considerations drawing some people, towards inductive logic, and others to the practical (frequentist approach)—while crystallizing the divide between Bayesians and orthodox approaches to statistical evidence and inference, providing an important test case for the two competing approaches.

What is less apparent, however, is that, in addition to the traditional problem of stopping rules, there is a second problem having to do with ongoing data monitoring and early termination of RCTs. This furnishes a second important test case for these two broad approaches to statistical evidence and inference. The next section expands on the problem of early termination of RCTs, and suggests that the issue actually favors a decision-theoretic approach, e.g. Bayesian decision theoretic methods, over an ES approach.

1.5. The problem of early stopping of RCTs

In recent years, there has been a growing concern about the proper conduct and reporting of RCTs. We can identify three problems as being of interest and importance to the analyses of interim monitoring decisions of RCTs, namely: (a) the inadequate reporting of RCTs; (b) the public mistrust about RCTs in developing countries; and (c) the skepticism about the conduct of RCTs in developed countries. Let me briefly describe each of the problems.24

24 Here I am simply reporting concerns that have been expressed by a wide variety of people. I do not necessarily endorse any of the positions.
1.5.1. The inadequate reporting of RCTs

High on the agenda of clinical trialists\(^\text{25}\) is the inadequacy of reporting on RCT, particularly trials stopped early due to benefit, harm or futility. Regulatory agencies (e.g. NIH, U.S. FDA) have issued recommendations and general guidance for both the conduct (e.g. ICH E6, ICH E9)\(^\text{26}\) and reporting (e.g. ICH E3)\(^\text{27}\) of RCTs. Independent initiatives have been launched by international researchers and journal editors (e.g. CONSORT).\(^\text{28}\) Despite these efforts, recent systematic reviews of RCTs show that top medical journals continue to publish trials without requiring authors to report details for readers to evaluate early

\(^{25}\) ‘Clinical trialist’ is a term often used in communities such as those associated with the Society for Clinical Trials (SCT). In this society one finds members from a wide variety of environments, including academia, the pharmaceutical industry, government agencies (e.g. NIH, U.S. FDA, CIHR), medical organizations (e.g. medical care centers), patient advocacy groups, and other clinical research entities.

\(^{26}\) Also known as the “Good Clinical Practice” document, ICH E6 is a document that provides general guidance for trial sponsors. It is an international standard for “designing, conducting, recording, and reporting trials that involve the participation of human subjects”; where compliance with it “provides public assurance that the rights, safety, and well-being of trial subjects are protected, consistent with the principles that have their origin in the Declaration of Helsinki, and that the clinical trial data are credible.”

ICH E9, on the other hand, is a document known as “Statistical Principles for Clinical Trials” and is intended “to provide recommendations to sponsors and scientific experts regarding statistical principles and methodology which, when applied to clinical trials for marketing applications, will facilitate the general acceptance of analyses and conclusions drawn from the trials.”

\(^{27}\) This document provides guidance for industry on the “Structure and Content of Clinical Study Reports”. Its main objective is to “facilitate the compilation of a single core clinical study report acceptable to all regulatory authorities of the ICH region”—i.e. the U.S., European Union, and Japan. The guideline is intended to assist sponsors in the development of a report that is “complete, free from ambiguity, well organized, and easy to review.” Among the things the study report should provide is “a clear explanation of how the critical design features of the study were chosen and enough information on the plan, methods, and conduct of the study so that there is no ambiguity in how the study was carried out.”

With regards to the specifics of interim analyses and data monitoring, “all interim analyses, formal or informal, preplanned or ad hoc, by any study participant, sponsor staff member, or data monitoring group should be described in full, even if the treatment groups were not identified.” (section 11.4.2.3)

\(^{28}\) Consolidated Standards of Reporting Trials (CONSORT) is an initiative aiming at authors and journal editors to improve reporting of RCTs, for the purposes of “enabling readers to understand a trial’s conduct and assess the validity of its results.” (Moher et al 2001, 1)
stopping decisions carefully. (cf. Montori et al 2005; Mills et al 2006) Many experts have argued that, in order to understand the results of an RCT, readers must be able to have information about its design, conduct, analyses and interpretation. Clarity in reporting is paramount. (cf. Moher et al 2001)

This elementary problem, lack of explicit justification for DMC decisions, is widespread. As just one example, consider “the best AIDS-prevention news in years,” a recent placebo-control trial (CAPRISA 004) of a vaginal gel showing a 39% reduction in the incidence of HIV infection in women. What the reader notices is that despite its encouraging results (released at a world AIDS conference and published in Science), “experts are pondering the many questions raised by the news”, including the fact that “if it was known after the first year that the gel was working, why wasn’t the trial stopped?” (ibid) The author of the NYTimes article reports that Dr. Karim—the principal investigator of the study—said the trial did not stop “because the independent review board that could have done so wanted results so overwhelming that they would have equaled the results of two generally favorable trials.” Given the importance of such a study and its findings, particularly in light of the ongoing fight to find effective HIV/AIDS preventions, it is disappointing to say the least to find that no explanation is ever given in the RCT’s final publication; not even a sentence is given regarding the two interim decisions issued from the two preplanned interim analyses in the study. No description of the adopted stopping rule is provided either, except one note, that “stringent stopping guidelines” (Karim et al 2010, 1170) were used.

The problem of inadequate reporting of RCTs is further aggravated by yet another problem, namely, the skepticism about RCTs in developed countries—which I briefly explain next.

1.5.2. The skepticism about RCTs in developed countries

A second set of concerns about RCTs is represented by those expressing sentiments of skepticism about the merits of current guidelines and regulations of RCTs in developed countries. The thinking here is that in order “to best strike the balance between regulation and cost” (Duley et al 2008, 40) we should get rid of most current guidelines, since they are ineffective anyways. Premised in terms of cost-saving practices meeting market demands, the argument is that there is little evidence that “the increasing bureaucratic and regulatory sludge” has improved the quality of RCTs; what it has done, instead, is continue to “increase RCT costs.” Consequently, what is needed is a “radical re-evaluation of existing trial guidelines [by] eliminating unnecessary documentation and reporting without sacrificing validity or safety.” (ibid) One reason for such ‘radical measures’ is that current guidelines “do not take into account how a trial should be conducted in a resource-poor setting compared to a setting where resources are more generously available.” (ibid 48)
1.5.3. The public mistrust of RCTs in developing countries

As a result of competition among drug researchers, RCTs have shifted to developing countries, typically countries with little research experience and weak regulatory agencies. (Petryna 2009) A common public suspicion has been that industry-sponsored trials conducted in such countries can bypass the stringent ethical guidelines and regulatory demands found in the U.S. and the U.K. Allegations that such trials often employ questionable cost-saving practices (e.g. cheaper labor costs, accelerated patient recruitment practices, weaker evidential standards) add to a growing public distrust about the validity and legitimacy of such trials.

For RCTs that are sponsored by the U.S. government and conducted in developing countries—whether or not by U.S. researchers—further concerns have been expressed by leading U.S. medical journals, including the issue of “whether it is appropriate to apply the same set of ethical standards and procedures that is used for trials in the U.S. to trials conducted in developing countries, where the context may be different” and whether such

---

30 For a review of the ethical and scientific questions raised by globalization of clinical trials, see Glickman et al (2009). Among the issues identified by the authors is this: “[w]e know little about the conduct and quality of research in countries that have relatively little clinical research experience.” (ibid 818) The issue is further aggravated by the fact that local regulatory agencies, because they are structured to monitor the quality of data and the safety of drugs “in their domestic jurisdictions” only, provide limited information and impact on several important aspects of trial conduct outside their jurisdiction.

31 According to a recent conference that focused on the issues involving “clinical investigators’ incentives to participate in industry-sponsored RCTs,” two of the three most fundamental challenges that investigators face in making “industry-sponsored trials more attractive to participants” are: (a) “ensuring the publication of the results of any clinical trial” and (b) “development of fair, standardized clinical trial agreement.” (proceedings of the Clinical Trial Magnifier December 2010 conference in Malaysia)
trials “pose unique ethical issues that must be addressed.” (Editorial Board *NEJM* 2001, 139)

1.5.4. Data monitoring committee

The idea of a data monitoring committee (DMC) is a natural response to the epistemic, ethical and economic complexities just noted. Several well-known RCT experts, while acknowledging the importance of current guidelines such as ICH E9, and the availability of current statistical methods for helping assess new interventions, note that in *practice*, the decision whether—and exactly when—to stop an RCT is a difficult problem. It is difficult due to the range and complexity of epistemic, ethical, and economic factors involved in such decisions. Because of their complexities, early stopping decisions are best served by the assistance of an independent DMC. Although this is a common view among many leading practitioners, it is not shared by all.32

The DMC is an external and independent group—presumably the only group reviewing the outcome data by treatment assignment. It assumes responsibility for the safety of trial participants and the future patients who might potentially use the new intervention. In fulfilling these duties the DMC has an obligation to make recommendations

32 In a case against independent monitoring committees, Harrington et al (1994) argue that “design changes”, “routinely looking for flaws”, “recommending changes in data management or standard monitoring”, or “any decisions about the scientific value of a trial in the larger context of national or international research” should be left to and “are the responsibility of the scientists conducting the trial.” (ibid 1413) If there are any problems reflecting “unskilled or inept investigators” or “insufficient or weak peer review”, or “lax or overburdened funding agency”, they should be addressed and corrected “at their sources”, that is, not by adding another “bureaucratic” review body, as implied by the addition of an independent DMC to oversee the conduct of RCTs.
to trial sponsors and principal investigators on the proper conduct of RCTs, whether for government or industry-sponsored trials.\textsuperscript{33}

The fact that many truncated trials by researchers do not provide proper justifications for their interim decision in their respective publications is already an indictment of researchers’ early stopping decisions and reporting. If we succeed in deriving a means of giving plausible ‘reconstructions’ of early stopping decisions, then justifications for pertinent RCT decisions can be provided. Without such justifications, the validity and generalizability of the DMCs’ decisions and the subsequent analysis of their decisions are problematic. By providing a common framework that structures justifications in individual cases (and permits comparisons between RCT cases) we shall have the beginnings of a general model for interim decision-making.

The reasons behind these concerns, here only briefly summarized, merit close scrutiny. But taking too narrow an approach can be misleading. Instead, by taking an approach that is both general and rigorous, the decision theoretic framework that is presented in chapters 3 and 4 aims at representing and evaluating early stopping decisions based on a clear and consistent set of criteria. In the next section, I explain why early stopping of RCTs can be a problem for ES.

\textsuperscript{33} According to a recent report from the U.S. Department of Health and Human Services “too many researchers are not adhering to standards of good clinical practice. The FDA has identified cases in which researchers failed to disqualify subjects who did not meet the criteria for a study, failed to report adverse events as required, failed to ensure that a protocol was followed, and failed to ensure that study staff had adequate training. These were not isolated incidents on the fringes of science. Instead, these troubling problems occurred at some of our most prestigious research centers and involved leaders in their fields of study. There can be no shortcuts when it comes to the protection of human subjects. Good clinical practices are neither esoteric nor frivolous. When adverse events are properly reported, the FDA and other bodies charged with the oversight of research can assess the safety of a particular study, as well as similar studies, and look for trends. In that way, subjects are better protected.” (“Protecting research subjects—what must be done” \textit{NEJM}, vol.343, number 11, p.809)
1.6. Why early termination of RCTs can be a problem for ES

In this section, I raise two problems for ES when it comes to explaining and justifying early stopping of RCT experiments.

(1) The first problem pertains to the need for balancing between early and late-term effects in RCTs. ES is silent about this balance. Severity (and other ES concepts) provides inadequate guidance in a case of early unfavorable trends because of the possibility of trend reversal. Because ES is an incomplete guide to early stopping, it needs to be supplemented in RCTs. Below, I show how the method of conditional power functions as an add-on to ES; we thus have direct evidence from practitioners of the inadequacy of ES. I examine this problem for ES in section 1.6.1.

(2) The second problem is that ES is silent on how to prospectively control error rates in experiments requiring multiple interim analyses. Because ES is silent on how to control error rates, ES provides inadequate guidance in cases of early stopping. The notion of a spending function is introduced and explained as a necessary add-on to ES for RCTs with multiple interim analyses. I examine this problem for ES in section 1.6.2.
1.6.1. Balancing early vs. late-term effects

In this section I explain the principle of severity according to ES, and then proceed to the main argument, which is that severity (and other ES concepts) provides inadequate guidance in a case of early unfavorable trends because of the possibility of trend reversal. Severity (and ES in general) needs to be supplemented by *stochastic curtailment testing*, a form of statistical testing which I explain in this section. The method of conditional power is one possible add-on to ES that can fulfill this need.

According to ES, in scientific experiments researchers use probabilities and statistical considerations to explain how the data provide evidence for (or against) the scientific hypothesis at stake. Probabilities are not used to supply degrees of credibility or support to hypotheses but as ways to characterize the test’s error probabilities (Mayo 2000, S196).

At the core of ES, we find the *principle of severity*. Severity has two conditions. We say hypothesis H passes a severe test T with data $x_0$ if the following two conditions are met:

(S-1) $x_0$ agrees with H; and

(S-2) With very high probability, test T would have produced a result that accords less well with H than $x_0$ does, if H were false. (Mayo and Spanos 2006, 329)
When conditions S-1 and S-2 are met, ES says that data x₀ is good evidence for H, if T severely passes H with x₀. Severity can be construed and interpreted as a triad relation <T, H, data x₀>. It is important to emphasize that severity is not a property of the test T exclusively, but a property of the test T passing H with data x₀. Let me illustrate the severity principle with an example of statistical testing.

A statistical hypothesis is an assertion about a particular parameter (e.g. H: θ=θ₀). Consider a statistical parameter θ, e.g. the decrease in level of viral load (for a certain population of HIV+ individuals). The test hypothesis H₀: θ=θ₀ asserts that the antiviral therapy does not cause a relative decrease in a patient’s viral load level. While H₀ asserts that there is no difference in the reduction of viral load level between the members of control vs. treatment group, the alternative hypothesis H₁: θ > θ₀ asserts that the antiviral treatment causes a positive decrease in viral load level.

Suppose at week 8 an unfavorable trend is observed, i.e. the test fails to reject the H₀ hypothesis. If a statistically non-significant result is observed at that interim—i.e. the p-value happens to be greater than the predefined significance level (alpha) for that particular interim analysis—then the statistically non-significant result could mean one or more of the following—the list is not meant to be exhaustive:

34 A more realistic formulation of H₀ is that the antiviral therapy does no better than the control, i.e. the proportion of patients among the treatment group experiencing a substantive decrease in the level of viral load, is no greater than the proportion of patients in the control group experiencing a substantive decrease in the level of viral load.
(i) The effect investigators are trying to detect—drop in viral load level—is simply not statistically significant.

(ii) The sample size used up to that interim is just too small to be able to indicate the effect of interest.

(iii) Investigators are measuring the effect of interest in a wrong manner, e.g. the chosen test statistic (which stands for the discrepancy or departure of interest from the $H_0$) is inappropriate for measuring the effect size.

ES interprets statistically insignificant results in the following way: since statistically insignificant differences can occur with a certain frequency, even in populations where the antiretroviral treatment does reduce the level of viral load, failing to find a statistically significant difference with a given experimental test is not the same as having evidence for asserting that $H_0: \theta = \theta_0$ is true. Informally, the ES analog argument from error says that we should not declare an error (e.g. discrepancy from $H_0$) is absent if a procedure had little chance of finding the error even if the error existed. Severity gives the formal analog of such argument. Even though “by the severity requirement, a statistically insignificant result does not warrant ruling out any and all positive [decrease in viral load level] ... severity does direct us to find the smallest positive [decrease] that can be ruled out.” [Emphasis is mine] (Mayo 1996, 197)
In post-data evaluations, the *severity* evaluation is framed in terms of a discrepancy \( \gamma \geq 0 \), from \( H_0 \). That is, for the case when the Neyman-Pearson test has accepted \( H_0 \) (failed to reject \( H_0 \)) the severity evaluation is framed in terms of the smallest warranted discrepancy \( \gamma \geq 0 \), measured on the scale of \( \theta \), with its magnitude assessable on *substantive* grounds. (Mayo and Spanos 2006) Since the Neyman-Pearson statistical test does not provide any information about the substantive or practical importance of its outcome, *severity* evaluates the extent to which a substantive or practical claim, such as \( \theta \leq \theta_0 + \gamma \) (associated with acceptance of \( H_0 \)) is warranted on the basis of the particular test and outcome.

For instance, if the test had a high probability (high power), e.g. 0.8, of rejecting \( H_0 \), if the decrease in the level of viral load were 3 times larger among treatment than control group, then a failure to reject \( H_0 \) *does* indicate that the actual decrease in the level of viral load is not as large as 3-fold.\(^{35}\) That is because in the case of accept \( H_0 \) (i.e. not rejecting \( H_0 \)) “high power entails high severity.” (Mayo 1988, 499) Consequently, not rejecting \( H_0 \) does not rule out decreases as large as 3-fold, if there is a small probability (low power) of rejecting \( H_0 \) even if the decrease in the level of viral load were as large 3-fold.\(^{36}\)

---

\(^{35}\) One way to understand the claim ‘decrease in the level of viral load’ is to read it as the proportion of patients who experienced a decrease in their level of viral load. Therefore, the 3-fold comparison refers to the difference in the proportions of patients experiencing a decrease in viral load, treatment vs. control group; e.g. 5% of the control individuals vs. 15% of those under treatment.

\(^{36}\) A technical note for completeness: according to Mayo (1988) the difference between power and severity is that while severity is a function of the particular observed difference (\( D_{obs} \)), the power is a function of the smallest difference judged significant (\( D^* \)) by a given test, i.e. \( D^* \) is the critical boundary beyond which the result is taken to reject \( H \). The power of test \( T \) against an alternative \( H_1 \) equals the probability of a difference as large as \( D^* \), given that \( H_1 \). The severity, in contrast, substitutes \( D_{obs} \) for \( D^* \). (see Mayo 1988 footnote 15)
In other words, a failure to reject $H_0$, according to ES, provides reason to say that the data provide good grounds that the drop in viral load level is no greater than such and such “upper bound”. And if you are interested in finding a plausible “upper bound” ES requires determining the extent of decrease in viral load level that with high probability (high power) would have resulted in the rejection of $H_0$—i.e. the decrease against which the observed difference had high severity. (Mayo 1988, 499)

The essential point I want to stress is this: even though the Neyman-Pearson test addresses the question of whether the observed data are reasonably consistent with the assumed null-hypothesis ($H_0$), the particular discrepancy from $H_0$ observed, whether at interim or at any stage, may or may not be of substantive importance. This follows from the simple fact that what is substantively important (e.g. clinical significance, health policy significance) in any particular context depends on factors outside what the Neyman-Pearson test itself can tell investigators, which is limited to statistical significance. In addition to this limitation, the value of severity assessments in post-data analysis, which is found in the evaluation of the extent to which a particular substantive or practical claim—expressed in terms of the distribution under the assumption of $H_0$ (e.g. $\theta \leq \theta_0 + \gamma$)—is warranted, is made on the basis of the particular test and outcome.

Even though this is a desirable feature that severity provides, which in certain contexts can be interesting and informative in assessing substantive claims, it is of limited value in the contexts of early stopping, where a DMC faces the possibility of making a poor or incorrect decision causing an inferential error with serious consequences. For example, the DMC may decide to end an antiviral treatment trial early on the basis of an early
unfavorable trend, leading to an incorrect conclusion about the overall long-term effects of the treatment under evaluation. In such a case, three considerations are worth mentioning. (1) Because of experimental considerations such as predictive uncertainties involved in cases where surrogate markers are involved; (2) because there could be a reversal of trend in the trial, i.e. from an early unfavorable trend to a late favorable trend; and (3) because the stakes are high, the DMC should use all of the relevant information that it might have at its disposal during interim for making a well-informed decision as to whether to stop or continue the trial. Its decision should be safeguarded by an assessment of all the relevant data, including varying a range of reasonable assumptions about the treatment effect and its surrogate marker.

Relevant information that the DMC might consider using could include results from the secondary interim endpoint, as well as information from other ongoing trials, when these are available, despite putting more weight on the primary end point in the particular study. To borrow the words from Ian Hacking, “[s]tatistical data are never the only data bearing on a practical problem. ‘All things being equal except the statistical data’ is empty.” (1965, 112)

Take for instance the uncertainties involved in the expected treatment effect at week 8, and expected treatment effects during subsequent weeks, and how these uncertainties can be dealt with using statistical methods that differ from the statistical tests (and tools) amenable to direct severity assessments. While it might not be feasible to plan for every contingency during a trial, when facing an early unfavorable trend, the DMC might have the means to make a judgment about the likelihood of a trend reversal.
The statistical method of conditional power—also known as stochastic curtailment testing—can be used to assess whether an early unfavorable trend can still reverse itself, in a manner that is sufficiently clear to show a statistically significant favorable trend at the end of the trial. Conditional power allows one to assess whether the trend reversal is possible and how likely the reversal might be to occur, which could be informative for scenarios involving an early unfavorable trend, or an early beneficial trend. It also allows one to judge whether the ongoing trial should be extended beyond its planned termination in order to be able to answer the question at stake.

The method of conditional power was first introduced by Halperin et al (1982) and Lan et al (1982) “in assisting with the decision whether a trial should be stopped before its planned termination, when there is a suitably high conditional probability on an ultimate rejection or acceptance of the null-hypothesis.” (Andersen 1987, 68) Andersen (1987) himself was the first to apply the same ideas of stochastic curtailment to the decision of whether an extension of the trial would be necessary for detecting significant differences under situations where survival is the primary endpoint.

The concept of statistical power is used because it is the probability of detecting a treatment difference, at the end of the trial, if such a difference exists—i.e. if the alternative hypothesis is true. Power is defined as 1 – (probability of type II error). Power is a function of the sample size, alpha, the specific alternative hypothesis, and the test statistic. But conditional power—as opposed to the initial statistical power of the test—is termed conditional because it takes into account the data observed up to that point in the trial. By
taking into account the early unfavorable trend, conditional power can compute the probabilities that the current trend could revert yielding a statistically significant result at either a subsequent interim, or at the originally scheduled end of the trial. It allows these probabilities to be computed under different assumptions of the effect size originally hypothesized, when the statistical test was initially designed. The method also allows assumptions about the effect sizes (different alternatives) to differ from the hypothesized $H_0$, with the possibility of having different weight assigned to each alternative hypothesis.

Conditional power can be of value to DMC members in deciding whether to stop or continue the trial in the following ways. If the results from calculating the conditional power, for a range of reasonable treatment effect sizes, indicate that the probabilities of a reversal are small, then the DMC might have further reasons to discontinue the trial since the treatment is unlikely to show statistically significant benefit effects. On the other hand, if the results from calculating the conditional power show that the reversal is quite probable, then the DMC might have a further reason to continue the trial.

It is important to note, however, that there are different variations in the method of conditional power computations. Because when computing conditional power investigators must make an assumption about the distribution of the future data yet to be observed, either using the assumptions from the original study design or assuming that the future effect of interest will have the same effect size (or an effect size reasonably close to but not exactly the same) as that estimated from the current trend in the data, there are different versions of conditional power computations. Each assumption about the effect size can yield a different conditional power value.
The original frequentist method, simply known as the *conditional power* method, typically computes the probability of rejecting the null-hypothesis under a pre-specified effect size (θ) conditional on the data observed up to that moment. If, however, investigators desire to compute conditional power considering a range of reasonable effect sizes with different weights attached to them, then the hybrid Bayesian-frequentist method of conditional power may be appropriate. This hybrid method is known as the *predictive power* approach, and involves averaging the conditional power function over the range of reasonable effect sizes. Most Bayesian uses of conditional power assessments that one finds discussed in the biostatistics literature—journals such as *Statistics in Medicine*, *Controlled Clinical Trials*, *Biometrics*—are examples of this hybrid Bayesian-frequentist approach. Spiegelhalter et al (2004) and Choi and Pepple (1989) are often cited as references for this particular approach to clinical trials.

Conditional power has been used in a wide range of RCTs. In chapter 2, section 2.3, I examine a case-study involving an early stop due to futility trial where the method of conditional power was essential in helping the DMC solve its dilemma. DMCs have come to rely on this approach, not in an exclusive manner, but in a crucial and complementary way to hypothesis testing when making an informed decision on whether to continue or stop trials early, whether for futility, harm or efficacy.

---

37 Given that $Z(t)$ is the standard statistic at information fraction $t$, and $Z_{α/2}$ the critical value for type I error, then conditional power (CP) for some $θ$ is given by: $P(Z(1) \geq Z_{α/2} | Z(t), θ) = 1 - \Phi[Z_{α/2} - t^{1/2}Z(t) - θ(1 - t) |(1 - t)^{1/2}$
1.6.1.1. The upshot for ES

If the value of a methodological rule is found in “understanding how its applications allow investigators to avoid certain experimental mistakes,” including “being able to amplify differences between the expected and actual data trend,” (Mayo 1996) then conditional power is not only desirable, but also a method that meets the experimentalists’ demand for avoiding or circumventing error. If the method of conditional power is helpful to DMCs when trying to safeguard their decisions against the possibility of error during an early unfavorable trend in the trial, then conditional power, as a class of statistical techniques, should be added to the experimentalist’s tool kit of techniques of learning from error.

So far I have argued that from an RCT point of view, one does not want to be dogmatic about a set of methods. Different problems demand different empirical methods. Flexibility over appropriate experimental methods is possible. I focused on ES because that is a popular and influential methodological view in philosophy of statistics that has the potential of being applied to the problem of early termination of RCTs. The point is not to refute ES and champion a particular class of statistical method, e.g. conditional power or hybrid Bayesian-frequentist methods. By showing the complexity involved in balancing early vs. late effects during ongoing trials, my general strategy has been to advocate a cautionary approach to clinical trial methodology. My arguments in favor of the role and importance of conditional power methods give no reason to prefer Bayesianism over ES or ES over Bayesianism as providing the best overall methodology to RCTs.
Although I have raised issues with ES, showing how conditional power can help DMCs during an early unfavorable trend, this neither suggests nor claims that severity assessments are inapplicable to post-data analysis in RCTs or inapplicable to RCTs which were early stopped. Although non-Bayesian approaches to evidence and inference have dominated statistical theory and experimental practices for most of the past century, the last two decades have seen a flourishing of hybrid Bayesian-frequentist and 'pure' Bayesian approaches (Gelman et al 2004) also applied to RCTs. What is more, there are aspects of RCT which other statistical methods capture much better than ES. Generally, there is good reason to expect “that scientists can be perfectly comfortable using Bayesian and classical methods side-by-side … [leading] to a pluralistic theory of scientific inference” (Steel 2001, S163).

My point about ES being limited for the purposes of early stopping will become clearer as we examine three real RCT examples in chapter 2. But before passing to the examples, I will raise one more objection against ES and explain why one should be suspicious about the value of severity for explaining and justifying the early stopping of RCTs. The next sub-section introduces the problem of controlling the overall error rate in RCTs. This is a problem that researchers face when designing and monitoring RCTs, and it gives us yet another objection to the view that ES is sufficient to make early stopping decisions.
1.6.2. **Rationing and the allocation of statistical boundaries in RCTs**

In this section I explain the problem of controlling error rates in RCTs with multiple interim analyses. I then introduce the notion of a spending function, give examples of such functions and explain how these functions address the problem of controlling error rates. I then proceed to the main argument, which is that ES is silent on how to prospectively control error rates in experiments requiring multiple interim analyses. Because ES is silent on how to control error rates it provides inadequate guidance in cases of early stopping. Because ES is an incomplete guide to early stopping, it needs to be supplemented. Rationing is introduced and explained as an add-on to ES and an essential principle for early stopping.

One of the main difficulties when monitoring RCTs with multiple interim analyses is that the overall alpha error rate increases with the number of hypothesis tests that are performed while data accumulates. (I shall explain the issue of error rate increase below.) It is one matter to have a single statistical evaluation of the data, usually performed at the conclusion of the study, and quite another matter to have an evaluation strategy that has to deal with the possibility of a type I error taking place at either the end of the study or

---

38 The overall alpha error is simply the overall probability of a type I error (incorrectly rejecting a true null-hypothesis) given a decision procedure that includes multiple statistical tests. It is the type I error rate for the entire collection of statistical tests applied to a dataset. By increasing the number of statistical tests (e.g. tests of significance, hypothesis testing) conducted over a fixed dataset for instance, where the null-hypothesis is true, there is a corresponding increase in the probability of a statistically significant result. A variety of statistical procedures have been developed to allow the researcher to control for the overall alpha error. Bonferroni method is an example of such procedure. Bonferroni allows the researcher to control a decision procedure for a desired overall alpha error (e.g. 0.05). For instance, if the researcher wants to perform a significance test 8 times to the same dataset, then in order to have an overall alpha of 0.05, the researcher would set to alpha to 0.00625 for each of the 8 tests to be performed.
during any one of the study’s multiple interim points. Because the study could be
terminated at any one of its interim points, investigators need to understand the
probability that the evaluation procedure meets the criteria of early stopping in any
interim interval.

If a chief concern for evaluating evidence is the control of error probabilities, as ES
has it, then for the purposes of prospective considerations for RCT with multiple analyses,
investigators should also concern themselves with how to control or “spend” the overall
alpha error rate among the multiple interim points in the trial. Although several
approaches to the problem of error rate allocation can be taken, this is an issue that neither
the Neyman-Pearson framework nor severity as an evidential principle for guiding
experiments directly addresses. In practice, the Neyman-Pearson framework (or any ES
approach) has to be supplemented with what I call a spending rule or a ‘rationing’
principle.

Before I explain what a spending rule is and more broadly what I mean by a
rationing principle, let me explain the basic problem of alpha error rate inflation, and the
difficulty researchers face when designing RCTs with interim analyses. The problem is
easily illustrated by a simple scenario. Consider a researcher who has designed a 2 year
RCT with a type I error rate of 0.05 in sight. Rather than evaluating the RCT results only
once at the end of the trial, she tests the accumulating data every 6 months. She tests at
each interim point, for evidence of efficacy, against a p-value of 0.05. That is, for any of the
4 interim analyses she uses an alpha of 0.05 as the critical statistical boundary.\textsuperscript{39} After all, one might reason, if 0.05 is an appropriate measure of evidence for the end of the trial, why isn’t it an appropriate measure during interim?

If we simplify things even further, and assume for the time being that all 4 interim tests are independent of one another,\textsuperscript{40} then we can see why the probability of type I error is no longer the original 0.05, but considerably greater. That is because given the multiple interims, the overall probability of type I error now becomes the probability of type I error either at the end of the trial or at any one of the interim analyses. Because there are 4 interim tests, the researcher must find the probability of at least one type I error. Given that the \( P(\text{no type I error at 1}^{\text{st}} \text{ interim}) = 1 - 0.05 = 0.95 \), we have \( P(\text{no type I error on all tests}) = (0.95)^4 \) and \( P(\text{at least one type I error on all tests}) = 1 - (0.95)^4 = 0.186 \). Thus, the actual type I error rate of the trial has increased from 0.05 to 0.186, a 3.7 fold increase.\textsuperscript{41}

The question for the researcher is simple: “what should I do with this increase of type I error rate?” Assuming she wants to control the overall error rate, that is, that she wants to keep the overall error rate under a certain threshold given the multiple interim analyses, she then needs to reason about how to ‘spend’ or distribute the foreseeable overall type I error rate over her experiment. One further issue worth noting is that she

\textsuperscript{39} A critical boundary tells the range of \( p \)-values that warrants rejection of the null-hypothesis. E.g. \( \alpha=0.05 \), which means any observed \( p \)-value that is less than 0.05 warrants rejection of the null-hypothesis.

\textsuperscript{40} Clearly they are not, since the data accumulates, but the simplified scenario is given only to illustrate the intuition and difficulty with having to deal with the increase of type I error rate.

\textsuperscript{41} Armitage et al (1969), Pocock (1977), O’Brien and Fleming (1979), Proschan et al (2006) and others have considered the problem from the point of view of the tests (interim analyses) being dependent of one another. This adds difficulty to the problem of error rate increase, a dependency which is approached either through numerical integration, e.g. Armitage et al (1969) or through simulation techniques (approximation methods) e.g. Proschan et al (2006). Regardless of the approach used however (whether through integration or simulation), the end results are the same, namely, as the number of interim analyses approaches infinity, the type I error rate slowly approaches to 1.
may or may not know *a priori* when or how often she might need to analyze the accumulating data in the trial. Thus, to put it simply, the problem of controlling or ‘spending’ the overall error rate that a researcher must face can be seen as the problem of controlling how much of the overall error rate (e.g. type I error) should be used at each interim analysis, given what the researcher intends to test in the trial. The answer is: it depends on the context. The context will inform the researcher which rule to adopt.

The *spending rule* is a rule that tells the researcher how much of the overall error rate should be used at each interim as a function of some factor that is contextual and relevant to the trial. For instance, the rule may tell the researcher to distribute the overall error rate evenly over equally spaced and expected calendar intervals (e.g. every 6 months). Another rule may tell the researcher to control the overall error rate as a function of the fraction (t) of the total information expected in the trial. Controlling it as a function of information could mean, for instance, controlling it as a function of the fraction of trial participants (e.g. t=0.25, 0.5, 0.75, and 1) that is expected to go through the trial. It could mean to control it as a function of the fraction of events (e.g. for time-to-death, progression of disease outcomes) of the total events expected in a 2 year trial. Let me give examples of such rules found in RCTs.\textsuperscript{42}

One spending rule is simply to have and perform all interim statistical tests at highly conservative type I errors (e.g. requiring the study to have p-values of 0.001 or less for each

\textsuperscript{42} The term “spending function” or “alpha spending function” has a technical meaning in the biostatistical literature. It typically means a function that approximates a group sequential method to RCTs. I use the term more broadly, as a rule or function that distributes the overall error rate over the intended interim points, whether or not the rule approximates a group sequential method.
interim point to justify early stopping) so that the impact on the final, and thus overall
alpha is considered relatively minimal. Even though this is a conservative approach to the
issue of alpha allocation, it has its place in certain RCT contexts. For instance, if a treatment
is expected to induce a long-term effect on healthy individuals, and the treatment is
somewhat known for presenting a substantial variability for its estimated effect given early
and small samples of individuals, then the approach is justified under the basis that later
evaluation of the effect size presents much less effect variability. This approach is
considered conservative, even when the RCT approaches the planned time of completion.
An example of such approach is the Haybittle-Peto monitoring plan developed for cancer
trials during the mid 1970’s. (See Ellenberg et al 2003 for further discussion)

In the Armitage (1969) group sequential procedure, the number of inspections in
the study is expected to be very large, with an alpha spending rule intended for use with
inspections made after every pair of patients are added to the trial. Two other commonly
used stopping approaches, the Pocock and O’Brien-Fleming approaches, are employed with
a smaller number of inspections in mind: a small but equally spaced number of interim data
analyses. Both the Pocock and O’Brien-Fleming approaches are commonly used in the
monitoring of HIV/AIDS RCTs.

When the chosen primary endpoint is a good predictor of an effect size that is
regular throughout the planned trial, then the Pocock approach which maintains the p-
value constant (i.e. the same critical boundary) over all interim analysis can be justified as
an appropriate choice for a monitoring plan. Since the Pocock procedure allows stopping at
an earlier interim analysis with less extreme treatment differences, it requires larger
treatment differences at the completion of the study than does the O'Brien-Fleming monitoring approach to achieve the desired level of significance. (cf. Proschan et al 2006, 86)

Ellenberg et al (2003) explains the rationale behind the O'Brien-Fleming approach to the spending rule. The rule is first very conservative early in the trial, when the number of events is relatively small compared to the remaining time of the trial. Its underlying rationale is that as more events occur, the information fraction increases and the stopping criteria for statistical significance become correspondingly less stringent as the trial progresses, with a significance level at the end of the study that is close to the overall p-value. When the stakes of making a false positive are high, the O'Brien-Fleming approach has an advantage over the other approaches since the power of the trial is maintained without having to increase the sample size significantly. “This is important because it permits the conduct of interim analyses while maintaining the false positive error rate at accepted levels without having to substantially increase trial size and cost.” (Ellenberg et al 2003, 12)

1.6.2.1. The upshot for ES

Given the problem of controlling the overall error rate, is it unreasonable to expect from a methodology of RCTs guidance on how to allocate or spend relevant error rates? If the chief concern is the control of error probabilities, then guidance with respect to how to spend
fractions of the overall error rate would seem appropriate. Statistical principles of evidence akin to ES are silent on such matters.

The obvious question then, is, “why should it matter to ES?” Why should the prospective allocation of error rates matter to an account of statistical evidence and inference? After all, contextual matters, namely, what is or isn’t substantively relevant to the control of error rates—for instance what is considered of clinical significance in a trial—has no bearing on what an account of statistical evidence should be. And to be fair to ES, it is not that ES denies that in science—e.g. medical research contexts such as phase III RCTs—evidence and contextual values (whether pragmatic, social or political values) are not often intermingled with evidence. But the issue is that ES contends that an account of scientific evidence and inference should extricate itself from what is considered substantively important by keeping the question of what is considered of value (e.g. clinical or epidemiological significance) separate from the question of what the evidence says. According to ES, an account of scientific evidence should not mix together what is considered substantively significant with what is statistically significant.43

---

43 On the general issue of what might be deemed substantively important in statistical testing, Mayo says:

[O]f course, the particular risk increase that is considered substantively important depends on factors quite outside what the test itself provides. But this is no different, and no more problematic, than the fact that my scale does not tell me what increases in weight I need to worry about. Understanding how to read statistical results, and how to read my scale, inform me of the increase that are or are not indicated by a given result. That is what instruments are supposed to do. Here is where severity considerations supply what a text book reading of standard tests does not. (Mayo 1996, 197-198)
But contrary to what ES says about evidence, i.e. that an account of scientific evidence should keep matters of substantive significance separate from matters of statistical significance, I argue that in order to understand the scientific evidence revealed (and reported) by RCTs—particularly in cases involving early stopping—we need to understand the context whereby statistical decisions occurred. And in order to understand the context of statistical interim decisions we need to understand what made such statistical decisions possible from the point of view of the planning of such trials. The ‘rationing’ principle, once identified, can shed light on such decisions.

A rationing principle is a principle used by those conducting an RCT because ‘resources’44 are scarce. The use of spending rules is an illustration of the need for rationing in RCTs. Trial participants (and consequently the information that can be extracted from them) are examples of scarce resources. Because RCTs are public experiments, human experiments on humans, the experiments are highly constrained ones. These experiments do not (and cannot) happen in a vacuum. That is, risking the embarrassment of stating the obvious, an RCT is not a statistical experiment. What we find in these experiments are researchers trying to make an informed decision. Researchers are trying to make the best decision they can given the highly constrained circumstances they find themselves with. RCTs are examples of science on the clock. Human beings are not replaceable. Trials cost money. Time is running out for many trial participants and concurrent patients who need access to better treatments. These truisms reflect reasons why such experiments are (or

44 Individuals (trial participants), information, care, goods and services are all scarce resources. These resources are utilized within the confines of RCTs.
should be) accountable to individuals, including the results (scientific evidence) and the recommendations that typically follow from their results.

For instance, consider two different health systems, i.e. two distinct social and political contexts, evaluating the same HIV treatment for asymptomatic AIDS individuals. Suppose system A operates under no overall financial constraint, with plenty of HIV asymptomatic individuals looking for the opportunity to participate in the RCT. Suppose system A is a form of U.S. health system with a third-party fee-for-service care system having a strong financial incentive to treat HIV asymptomatic individuals, as exemplified by the recent shift in HIV treatment in San Francisco, where “the program is being carried out without a clear picture of its cost to city and state health agencies.”

Suppose system B is a system of public health operating within a limited financial constraint. System A has a tendency to treat HIV+ asymptomatic individuals aggressively, whereas system B tends not to treat individuals immediately, since the costs and toxicity of

45 NYTimes article reporting on the introduction of a recent “controversial” “major shift of HIV treatment policy” in the city of San Francisco, where public health doctors had begun to advise HIV asymptomatics “to start taking antiviral medicines as soon as they are found to be infected, rather than waiting for signs that their immune systems have started to fail.” The author suggests that “the turning point” in the San Francisco’s thinking might have been guided by “a study in the New England Journal of Medicine on April 1, 2009, that compared death rates among thousands of North American H.I.V. patients” showing that “patients who put off therapy until their immune system showed signs of damage had a nearly twofold greater risk of dying from any cause [compared to patients that were healthy and with] T-cell counts above 500.” Leaving aside the fact that antiviral drugs can cost US$12,000 year per patient—taking approximately US$350 million of the state of California annual budget—there remains disagreement among researchers over whether such change in policy is a good idea after all; disagreement illustrated by Dr. Anthony Fauci (director of the NIAID) quoted as claiming that the new policy of early treatment is “an important step in the right direction” whereas Dr. Jay Levy (virologist from UCSF and one of the pioneers in identifying HIV as the cause of AIDS) quoted as saying of the policy, “it’s just too risky”, since “no one knows the effects of taking them for decades”—even though the new drugs may be less toxic than they used to be. The author of the article also reports that “San Francisco’s decision follows a split vote in December [2009] by a 38-member federal panel on treatment guidelines.” Russel, Sabin “City Endorses New Policy for Treatment of H.I.V.” The New York Times, published online April 3, 2010.
early HIV treatment are generally not considered worth the small predictive chance of benefit (e.g. longer survival) unless it is shown to be safe after a multi-year RCT.

One might ask: how much benefit can reasonably be expected from an early HIV treatment in asymptomatics? What level of expected benefit to patients would have to be considered high enough, in relation to costs and risks to these patients, before a treatment is deemed successfully evaluated? Public health officials from systems A and B may have very different answers to these questions. System B may demand that for expensive treatments, like the one in question, a multi-year RCT shall stop early only if the expected benefit from the treatment exceeds a certain threshold, e.g. 0.3 probability of “significant” improvement in terms of survival extension with “reasonable” quality of life compared to the current standard of care. The acceptable threshold for system A might be less conservative, e.g. 0.1 probability of “significant” survival improvement. The issue is not only epistemic, not just a matter of statistics and clinical practice, but a matter of social and public health policy as well.

For system A, which values the extension of lives of individuals very highly, regardless of cost (e.g., if the treatment is perceived to be a matter of individual right), gaining a small probability of longer survival may outweigh considerable discomfort, including possible side effects and financial burden on the individual and his or her family. For system B, on the other hand, we may have a more demanding and comprehensive standard or threshold for an acceptable quality of life for its citizens.
Once we start seeing that rationing decisions are not only morally required in designing and monitoring RCTs, but also morally permissible within the system in which the trial takes place—since RCTs must comply with the system in which they serve—it becomes more difficult to hold the view that disagreements over scientific evidence in RCTs are essentially reflective of theoretical differences in statistical philosophies. Or that disagreements of what should be deemed proper or good evidence in such experiments can be reduced to disagreements of statistical methodologies. It is very difficult to make sense of the statistical evidence—in terms of their error rates—as something isolated from what was considered of substantive importance at the time of the interim decision.

In sum, in this section I have argued that if the control of error probabilities is of central concern for evaluating evidence, then it seems appropriate that the issue of how to control the overall error rate among the multiple interim points in an RCT should also be of concern when evaluating the evidence. Even though different approaches can be taken to the problem of error rate allocation, this is an issue that ES is silent about. More specifically, severity as an evidential principle for guiding experiments is at best limited in certain RCT contexts. I have also explained why rationing is implicit yet essential to RCTs and how ES with its severity criterion of evidence is neither by itself sufficient for informing DMCs on how to prospectively design their trials nor for evaluating allocations of error rate, whereas in both cases, rationing is called for.

Although I have raised issues with ES, showing how conditional power can help DMCs during an early unfavorable trend (section 1.5.1), and showing the importance of complementing ES with a rationing principle for RCTs (section 1.5.2), these objections do
not claim that severity as a meta-statistical principle for experiments is inapplicable to post-data analysis in RCTs or inapplicable to RCTs which were early stopped. Severity can be informative, although its role can also be limited in the ways I have argued regarding RCT considerations.

In the next chapter I will take the point about legitimate disagreements among researchers further. There is a need for a clear framework for representing and evaluating early stopping decisions given typical disagreements seen during the monitoring of RCTs in light of early stopping decisions. In chapter 2, I explain important dissatisfaction among researchers, specifically epidemiologists, over the proper use and interpretation of classical statistical methods—such as the Neyman-Pearson framework—when dealing with clinical trials. Dissatisfaction with practice and the need for guidelines is motivated by three case studies involving HIV/AIDS trials. Discussions of the case studies shall be used to identify and explain important factors in early stopping decisions; these factors will motivate and inform a decision theoretic framework for early stopping decisions.
Chapter 2. Disagreements over early stopping of RCTs

In this chapter I explain the increased dissatisfaction among clinical researchers and physicians, specifically epidemiologists, over the proper use and interpretation of classical statistical methods when dealing with clinical trials. Dissatisfaction with practice and the need for guidelines is motivated by three case studies involving HIV/AIDS trials. The first case-study is a pair of treatment trials concerning two data monitoring committees arriving at two different conclusions: the U.S. AIDS Clinical Trial Group (Protocol 019) study of zidovudine (AZT) in asymptomatic HIV-infected individuals, and a similar study run by the European Concorde group. The issue here is early stopping due to benefit. The second case-study involves an early stop due to futility. This was an RCT designed to determine whether pyrimethamine would reduce the incidence of a brain-infection in HIV-infected individuals, with the DMC recommending the trial to stop and the chair that it be continued. The third case-study is about a trial that examined vitamin A supplementation for the prevention of mother-to-child transmission of HIV through breastfeeding. This was an instance of an early stop due to harm case, with vitamin A having an opposite effect, namely, increasing the chances of HIV transmission from mother to child. These case studies are used to identify and explain important factors in early stopping decisions, in a qualitative way.
2.1. Introduction

An important aspect of a properly designed and conducted RCT is the quality of its reporting, particularly trials stopped early for any reason. Ideally, the reporting of a trial should ensure that the results are interpretable with respect to the study’s primary question of interest and that the DMC’s monitoring decisions are properly justified. In this chapter, I identify and explain four considerations in the justification of early stop decisions of RCTs: (1) Scientific uncertainties; (2) Statistical monitoring plans; (3) Standards of proof; and (4) Balancing long-term and short-term drug effects. I discuss these considerations because they are precisely the factors that come up repeatedly in disagreements among practitioners over interim monitoring decisions of RCTs. These considerations will be introduced and discussed, in the context of the case studies, in sections 2.2.1, 2.2.2, 2.3.1 and 2.4.1.

Throughout the chapter, I give voice to different perspectives of HIV/AIDS researchers, statisticians, regulatory agencies, and their critics. In shedding light on the actual practice and the difficulties involved in the conduct of HIV trials, my aim is to examine the justification of early stop decisions while presenting a comprehensive view of the human dimensions regarding the interim monitoring of RCTs. My strategy is to discuss disagreements in early stopping decisions and interpret them from a point of view of practitioners’ dissatisfaction with current statistical methods and thereby the need for guidelines and frameworks that provide greater clarity for decision-making.
My examples are based on HIV/AIDS related RCTs. The examples are based on HIV/AIDS cases because HIV/AIDS is a domain where the stakes are high. I should clarify this important point. Even though the examples have unique aspects, they are telling in the following sense. Compared to other clinical research areas—where the lack of public information and discussion seem to be the norm—lots of information on HIV/AIDS cases is publicly available. This is largely due to literatures in the sociology of science, clinical research ethics, and writings from HIV/AIDS patients’ advocacy groups. Lots of information on HIV/AIDS cases is available because such cases were highly political, especially during the 1980s and early 1990s when AIDS and its epidemic were not well understood, and because the stakes continue to be so high for many people affected by the disease.

Another point is that, from an epistemological point of view, the HIV/AIDS domain of clinical research is particularly challenging. There are many statistical challenges when designing, conducting, and analyzing the results of such trials, including interpreting the final report of these experiments. But most importantly, there are ethical issues regarding the early stopping of RCTs that are particularly complex, because HIV/AIDS treatments not only offer life-prolonging potential but also are toxic (at some point in history their treatments were highly toxic) and when a patient begins treatment he or she has to stay with the treatment for life, thus making the human costs in such cases particularly high.

The main reason for selecting the particular three HIV/AIDS cases is that they illustrate the three different types of reasons for early stopping decisions, namely, efficacy, harm, and futility. Almost all the factors (epistemic and ethical) that might be relevant to cases of early stopping decisions for RCTs can come out on these three cases. These factors
will in turn motivate and inform the proposal of a framework for representing and analyzing early stopping decisions.

One methodological worry, however, is that by selecting these HIV/AIDS cases, I might be missing other relevant factors that should have been included, or that might have been included in the set of factors relevant to early stopping decisions. Given that I cannot guarantee or prove that my three cases include all and only those necessary relevant factors, this is a point of concern. However, a substantive proportion of the factors that are likely to be relevant to early stopping decisions are captured by the three cases because they represent so well the three typical reasons for early stopping. I will come back to this methodological point in chapter 3, when I introduce the qualitative framework for early stopping of RCTs. There I revisit the concern when explaining the method for selecting relevant factors for the reconstruction of early stopping decisions. For now I should add that the identification of relevant factors for the purposes of reconstructing early stopping decisions is driven primarily by the DMC mandate and concrete early stopping cases similar to the three examples introduced in this chapter.

As an example of considerations (1), (2), (3) and (4) when examining early stop decisions of RCTs, I begin with the interim monitoring experiences of two DMCs arriving at two different conclusions. This first case-study pertains to the U.S. AIDS Clinical Trial Group (ACTG) Protocol-019 trial of zidovudine (AZT) in asymptomatic HIV-infected individuals, and an equivalent trial run by a consortium of European researchers, also known as the Concorde trial. This case-study serves to illustrate differences in judgments about the implications of interim monitoring decisions by tracing different uncertainties underlying
the pertinent RCTs. This case-study also illustrates the sort of dilemma that investigators face when dealing with a situation of early stop due to benefit. The dilemma underlying cases of early stop due to benefit is further examined in more detail in chapter 3—see section 3.2.1.

My second case-study is about an HIV trial with the DMC recommending that the trial no longer continue while the primary investigator recommended the study to continue. This case-study illustrates primarily considerations (2) and (3) in early stop decisions of RCTs. The disagreement over the conduct of the trial is traced to uncertainties about the future (primary) event of interest, that is, the number of toxoplasmatic encephalitis cases that was reasonably expected for the remainder of the trial. This trial was an instance of an early stop due to futility scenario. It serves to illustrate the importance and impact of statistical monitoring plans when justifying early stop decisions. While the trial illustrates an early stop due to futility dilemma, the generality of the dilemma itself is further examined in chapter 3—see section 3.2.2.

The third and last case-study is an example of an RCT examining vitamin A as prevention of HIV transmission from mother-to-child through breastfeeding. It exemplifies primarily considerations (4) and (1) when examining early stop decisions of RCT. This case-study also illustrates the sort of dilemma that investigators face when dealing with a situation of an early stop due to harm. The dilemma underlying cases of early stop due to harm is further examined in more detail in chapter 3—see section 3.2.3.
2.2. The controversy over AZT in asymptomatic HIV-infected individuals

In the 1980s, when AIDS was first identified and growing at alarming levels, physicians, researchers, and health officials seemed incapable of responding quickly enough to the epidemic. Their inability spawned a "credibility crisis" surrounding the clinical sciences, resulting in grassroots movements and leading to HIV/AIDS treatment activism (Epstein 1996). This activism focused on the pace and organization of AIDS research, pressuring researchers to develop more clinically relevant trials with flexible designs that subjects would find more expedient yet ethically acceptable.

The situation led to the formation of the U.S. ACTG to oversee trials via data monitoring committees that "had to develop new operational approaches to deal with the many new challenges posed by HIV/AIDS trials," among them "an unprecedented involvement of patient representatives and advocacy groups in the design of trials and interpretation of results, and scientific and political pressures to identify effective treatments as quickly as possible." (Ellenberg et al 2003, 9) The DMC’s responsibilities included a duty to "review interim data on clinical trial performance, treatment safety and efficacy, and overall study progress. If the interim results provide convincing evidence of either excessive adverse effects or significant treatment benefit, the [DMC] may recommend early termination of the trial to the NIAID and the study investigators." (DeMets et al 1995, 409) In short, researchers had to find ways to compromise between on the one hand the activists’ demands for expediting drug evaluation and approval, and on the other maintaining the scientific validity and integrity of drug trials. This indicates the fundamental dilemma of the RCT.
A separate but closely related dilemma originated shortly after AZT was shown effective in prolonging survival of AIDS patients. There was a question dividing clinical opinion: whether asymptomatic HIV-infected individuals should start treatment early or later. Should asymptomatics start treatment before the development of AIDS—since, presumably, patients could tolerate AZT’s toxicity more easily, keeping them healthy for longer, thus increasing their chances of survival—or should patients delay a highly toxic treatment that could increase liver damage, as well as induce early drug resistance, thus reducing their chances of later disease resistance at a time they could need AZT the most?

The activists’ demands included more flexible trial designs and replacement of clinical end points with surrogate markers such as CD4 cell count. In 1989, as a result of this activism, the ACTG decided to accept its DMC recommendation to stop a placebo-controlled trial of AZT for asymptomatics, since based on the first interim analysis, the rate of progression to AIDS in the placebo group was more than twice that with AZT. “It was thought unethical to continue the original trial, given the magnitude and statistical significance of the observed differences.” (Pocock 1993, 1466) As expressed by the study investigators:

In August 1989, the first full interim analysis was presented to an independent board monitoring data and safety. On the basis of evidence that treatment at the lower dosage delayed the progression of disease in patients with initial CD4⁺ cell counts below 500 per cubic millimeter, and guided by rules established for the sequential termination of portions of the trial, the board recommended that the placebo arm of that sub-study be terminated. [Emphasis added] (Volberding et al 1990, 942)
The final conclusion of the study was that AZT was effective and safe in asymptomatics, including two disclaimers from the investigators: “because the subjects in the current trial were followed for no more than 2 years, the data provide no information about the possible long term benefit or safety of AZT” and “it is possible that even if AZT persistently delays the onset of AIDS, it may not have an ultimate effect on survival. Continued observation of the subjects in this trial may help resolve these crucial issues.” (Volberding et al 1990, 948)

There were mixed reactions. U.S. researchers praised the expediency and new initiatives in that trial, whereas European researchers were more cautious, suggesting that stopping the trial early prevented them from judging AZT’s long-term effects. The Europeans were thus in conflict with the opinion of those more optimistic US researchers who claimed the benefit in asymptomatic patients out-weighed AZT's potentially toxic long term effects.

Moreover, even after the results from the U.S. study had been reported, Concorde researchers decided to continue with their equivalent ongoing trial. Epstein quotes Jean-Pierre Aboulker, head of the Concorde study, as saying, “the results we have seen do not allow us to give a strict recommendation to give AZT.” (Epstein 1996, 245) Two years later, at the end of the three-year Concorde study, the results showed “no statistically significant difference in clinical outcome,” and “similarly, there was no significant difference in progression of HIV disease”; the researchers concluded that “the results of

---

46 Concorde Coordinating Committee: Concorde: MRUANRS randomized double-blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection. Lancet 343:871-881, 1994
Concorde do not encourage the early use of zidovudine in symptom-free HIV-infected adults. (...) The optimum time to start zidovudine remains unclear.” (Concorde 1994, 879)

The European study, while interested in the same set of questions of early vs. late AZT treatment for asymptomatic HIV-infected patients, according to DeMets et al (1995), at the end of the trial had shown that:

While disease progression differences in treated vs. placebo recipients existed at 1 year, all benefit was lost by 3 years. Furthermore, evidence over the extended follow-up period indicated that the earlier administration of AZT did not provide survival benefits. Finally, CD4 counts did not prove useful as a surrogate marker, since the large sustained improvement in CD4 counts following early AZT gave misleading predictions about treatment effect on the true clinical outcomes. (DeMets et al 1995, 417)

Behind the differences between the U.S. and the European studies lurk legitimate and important questions. Given that the trials were run in parallel for the same HIV-intervention, when one trial stops early, should the other stop too? How can we account for the difference between the U.S. and the European interim monitoring decisions? How is it that two committees have come to radically different decisions about AZT in asymptomatics? Were they the result of differences in ethical principles, different statistical monitoring procedures and their corresponding statistical thresholds, or different judgments about the consequences of particular actions? If the latter, was the difference a consequence of assuming different losses incurred by their action, or distinct loss functions altogether? Or were the committees really interested in a different set of questions after all?
According to a leading British researcher, Stuart Pocock, “in HIV trials, especially in the United States, the push towards individual ethics at the expense of collective ethics”, as in leading to “stopping trials too soon” “has been detrimental to determining the most effective therapeutic policies.” (1993, 1466) Although Pocock is never clear on what “individual” versus “collective” ethics really mean, in a more recent report entitled “Issues in Data Monitoring and Interim Analysis of Trials” (Grant et al 2005) in which Pocock is listed as one of the authors, the suggestion seems to be that U.S. researchers with their “individual ethics” approach put the interest of individual participants within a trial before those of future patients who may or may not benefit from the results of the trial. This approach is supposedly in contrast with the “collective ethics” approach adopted by Europeans, which puts the interests of future patients who may or may not benefit from the results of the trial before the interests of the individual trial participants.47

Although most investigators do not dispute that DMCs should have ethical responsibilities towards their trial participants and to others outside the trials, I do not think that the language used by Pocock, especially the distinction between ‘individual’ and ‘collective’ ethics, is of much help when trying to understand differences between DMC monitoring decisions. The notions of “individual ethics” and “collective ethics” are ambiguous. For one, I do not think it is necessary to see the differences between stopping and continuing an RCT as an issue of differences between the interests of those individuals in the trial versus those outside of the trial. More is needed to make these two terms consequential to understand the possible disagreements between DMCs.

47 See the glossary in the report (p. x)
For instance, a DMC can take the interest of those in the trial before the interest of the larger HIV asymptomatic population and yet be reluctant to stop the trial despite early evidence for efficacy. That is because there are other aspects to an RCT that go beyond just producing immediate evidence for efficacy, to include balancing the possible long-term side effects of the drug. The converse equally applies. Stopping a trial due to early evidence of efficacy can be consistent with a view of collective ethics, as in having the importance of reducing the spread of HIV/AIDS take precedence over the interest of trial participants that might want to know the long-term effects of the drug for their chances of survival. Thus, continuing a trial despite early evidence for efficacy can be consistent with either one of the two ethical views. It can be consistent with the view of safeguarding the interest of trial participants over the larger population, and it can be consistent with the view of prioritizing the interest of a larger population at risk over those of trial participants.

It seems dubious to me that individuals, including potential DMC members, would participate knowingly in a clinical trial that puts the interests of future patients—who may or may not benefit from the trial results—before the interests of individual participants within a trial. Anything short of prioritizing trial participants would seem at odds with today’s FDA guidance for clinical trials, with its insistence that the first and most fundamental reason for establishing a DMC is “to enhance the safety of trial participants in situations in which safety concerns may be unusually high.”

48 An important disclaimer is made in the guidance for clinical trial sponsors: “This guidance represents the Food and Drug Administration’s (FDA’s) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations. If you want to discuss an alternative approach, contact the appropriate FDA staff. If you cannot identify the appropriate FDA staff, call the appropriate number listed on the title page of this guidance.” (GCTS p.iii)
I have the impression that Epstein is correct when suggesting an alternative source of possible differences between the US and the European researchers. Epstein contends that the difference between the decisions is due to differences in the judgments about the implications of what was and was not known by the investigators. He makes this point when he puts the disagreement between their actions in the following way:

In practice, everyone agreed on which questions they would like to have answered—questions about long-term risks and benefits, about the development of resistance, about whether early use would squander the limited efficacy of the drug against the virus. The different actions taken had to do with different judgments about the implications of what was known and what wasn’t. How and when should a particular balance of certainty and uncertainty about a drug inform clinical practice? And what is the feasibility of continuing a placebo-controlled trial, once a specific—but not conclusive—benefit had been shown? (1996, 245)

Looking at different judgments about “the implications of what was known and what wasn’t” helps us understand the disagreements over the different interim monitoring decisions and corresponding AZT recommendations. I take this much to be uncontroversial. But what was known and what wasn’t needs much unpacking before we can have a proper reconstruction of their respective interim monitoring decisions. In my attempt to reconstruct what was known and what wasn’t, I identify four potential sources of epistemic differences. All four will be examined in this case study. The differences can be seen as:
(1) Differences in judgments about the basic sciences underlying RCTs (section 2.2.1);

(2) Differences in judgments about the competing policies guiding RCTs (section 2.2.2);

(3) Differences in judgments over the ‘standards of proof’ required by their RCTs (section 2.3.1); and

(4) Differences in judgments about the way to balance long-term versus short-term drug effects (section 2.4.1).

Let us begin with the differences in judgments about the sciences underlying RCTs.

### 2.2.1. Scientific uncertainties

The basic idea here is that there are different types of sciences relevant to running an RCT, and correspondingly, different types of scientific uncertainty underlying RCTs. That is to say that there are considerations from different sciences brought to bear on decisions by RCTs. This point about the different sciences also serves as a reminder to the reader about the complexity of interwoven issues that investigators face when dealing with the evaluation of HIV-interventions in RCTs. That is because different sciences—e.g. virology, immunology, epidemiology, statistics, the science of drug and regimen adherence—with their own set of uncertainties can combine in ways such that disagreements over the design and conduct of trials may be expected to occur. Furthermore, the type and degree of uncertainties in these experiments are quite distinct from what is seen in physics
experiments for instance, experiments that have become paradigmatic for philosophers of science.

In order to motivate this point, that the different uncertainties affecting RCTs help us to understand interim monitoring decisions, I shall consider three types of scientific uncertainty typically present in the evaluation of HIV-interventions. I begin with uncertainties surrounding the virus itself, followed by uncertainties about the immune system, and then surrogate markers.

2.2.1.1. Uncertainties surrounding HIV virology

Consider the situation of virology back in the 1980s. Back then there were—and until this day continue to be—uncertainties about HIV virology. One question that seems to linger until today is, when and how exactly does HIV replicate? The particular issue of when HIV starts its rapid replication is yet unsettled. This bit of science is important because it helps investigators deal with the problem of mutated strains of HIV, namely, that it is difficult if not impossible to treat HIV-infected individuals with a treatment or treatment ‘cocktail’ when the virus has mutated.

Consider the uncertainty over viral replication and its implications for understanding the possible disagreement over interim monitoring decisions. Suppose HIV starts its replication shortly after infection. Suppose that investigators agree and expect this much. But suppose there are uncertainties over (a) when exactly rapid replication...
initiates and (b) what causes viral mutations to occur. With regards to (a), the uncertainty is over when, after infection, genetic mutations begin and when they accelerate. This is an essential piece of information for delaying the development of AIDS. The presupposition is that the suppression of viral load and its replication are positively correlated with delaying the development to AIDS. With regards to (b) the uncertainty is whether HIV mutation is driven by drug selection pressure or by some type of stochastic genetic drift. Thus, if you think that viral mutation is essentially due to genetic drift, and that by controlling HIV replication sooner rather than later you reduce the rate of virus mutations, then you may advocate an early and aggressive AZT treatment on behalf of HIV symptom-free individuals. Your reason is that you do not want HIV-infected individuals waiting too long before they get treatment, because by treating quickly and aggressively you reduce the chances of viral mutation that can make it impossible to treat later with the same drug.

On the other hand, if you think that HIV mutation is due to drug selection pressure, and that mutation mostly occurs much later after infection, than you may want to advocate a different course of action on behalf of HIV-infected individuals. Your reason is that by avoiding early AZT treatment you also avoid drug-resistant mutations due to the absence of drug selection pressure. Thus, withholding AZT treatment until later in the course of HIV infection would seem like a better idea than following a treatment approach of early and aggressive AZT,49 because withholding AZT should increase the patient’s chance of survival.

49 Today there seems to be an agreement that HIV mutation is driven primarily by drug selection. See “Hit HIV-1 hard, but only when necessary” Harrington et al (2000), The Lancet, vol. 335, Issue 9221 p.2147-2152
2.2.1.2. Uncertainties surrounding the human immune system

Another source of uncertainty regarding the underlying sciences in HIV/AIDS RCTs is due to uncertainties about human biology, and more specifically, about how the immune system works. Here the locus of uncertainty is over the rate of progression from HIV infection to AIDS given the development of certain opportunistic infections. Because HIV progressively attacks a patient’s immune system by damaging its cells, namely, reducing its CD4+ cell count,\(^{50}\) HIV weakens the patient’s immune system to the point of full vulnerability to other infectious diseases. This process can take years before HIV has damaged a person’s immune system significantly enough that the development of opportunistic infections and AIDS is fully expressed. Today we understand that the asymptomatic stage in HIV-infected individuals takes on average 10 years before the individual becomes symptomatic. Back in the 1980s investigators did not know how long this could take. As a matter of fact not much was known about the pathogenesis of the disease.

Different infections would typically occur at different stages of HIV infection. For instance, in early stages of disease progression patients could develop bacterial related infection—e.g. tuberculosis, bacterial pneumonia—whereas in later stages of the disease, when the immune system was very weak, fungal and other types of disease developed. Thus, if you think that by intervening early and aggressively with a treatment that has shown promising results in keeping CD4 cell count close to normal (akin to the ACTG

\(^{50}\) Also known as T helper lymphocyte cells, they play a key role in maintaining a person’s immune system healthy.
Protocol 019 trial) you can reduce the speed of destruction of the patient’s immune system, then you might conclude that progression from infection to AIDS will also be delayed by an early intervention of AZT.

Suppose, on the other hand, that you think the benefit of early intervention with AZT should be weighed together with AZT’s highly toxic effects. Because AZT’s toxicity can make the patient’s immune system even more vulnerable to opportunistic infections than it currently is, by withholding AZT treatment you think a patient has a better chance of delaying the destruction of an already vulnerable immune system and the full onset of AIDS. I return to this point about the balancing of benefit and harm in section 2.4 when discussing a case of early stop due to harm, and disagreements over what the right course of action might have looked like. For now it suffices that we notice that various uncertainties over disease progression and its impact on the immune system can help us to understand different approaches among those sympathetic to stopping the trial early due to efficacy, with a treat early and aggressively approach in mind, and those who might be more cautious and want to know more about the toxicity of AZT before releasing it to asymptomatic populations.51

51 In a report about the ‘optimum’ time to start HIV treatment, published in 1996, Phillips et al reports that Those in favor of early treatment argue that as HIV is actively replicating throughout the clinically asymptomatic period it is probably vital, as in other infections, to treat as early as possible to gain the best results from treatment. This may be true, say those in favor of late treatment, but given that only around 50% of people develop AIDS 10 years after infection, it must be proved that any benefits of early treatment outweigh the possible harm from long term treatment. In other words a survival benefit for early treatment compared with late treatment has to be clearly shown in a randomized controlled trial. To date, the clinical trials that have addressed this question have not resolved this controversy. (BMJ 1996; 313:608-609)
2.2.1.3. Uncertainties surrounding surrogate markers

There are also uncertainties of a different kind of science: uncertainties over the predictive ability of particular surrogate markers. According to the FDA, a surrogate marker is “a laboratory measurement or physical sign that is used in therapeutic trials as a substitute for a clinically meaningful endpoint that is a direct measurement of how a patient feels, functions, or survives and that is expected to predict the effect of the therapy.”52 With respect to HIV-interventions, today, the FDA acknowledges two types of surrogate markers for the purposes of interim monitoring HIV-interventions: CD4 cell count, and plasma HIV RNA. I focus on CD4, since it illustrates well the point about uncertainties.

An important decision that investigators face when designing RCTs is the selection of a meaningful end point. In considering an end point, it is important that investigators consider the time necessary to evaluate the results of the trial. There are many instances in which the use of clinical end points—e.g. mortality, survival time—as the primary end point may be simply impractical. For instance, simply put, if patients are not dying quickly enough or in enough numbers, or when there is an ethical need to expedite drug evaluation, then the adoption of surrogate markers can be seen a means to evaluating the new intervention.

52 57 Federal Registry 13234, 13235 (April 15, 1992)
According to today’s FDA fast track drug approval process—a process that was originally put in place in order to deal with situations where no treatment alternatives are available, like the fast approval of AZT had—“if a drug can be shown to reduce the amount of HIV virus detectable in the blood of AIDS patients, it can be approved on the basis of its short-term effect on this surrogate endpoint, even though the drug’s durable effect on the virus and its ultimate effect on health and survival requires extended studies.”

However the adoption of surrogate markers is not without difficulty. For instance, researchers may disagree over a surrogate marker’s ability to correlate with a clinical outcome. Because an essential motivation for treating an HIV-infected individual is to increase the chances of maintaining CD4 cell count over certain thresholds, one reasonable concern many clinicians have is that we may keep an individual’s CD4 cell count under control with the help of a treatment for some period of time—e.g. over 500 per cubic mm for 2 years—but in the long-run, CD4 cell count may not imply clinical benefit, that is, it may not increase the chances of survival, or improve mortality rates—as suggested by the differences in conclusions and recommendations shown by the U.S. trial and the European researchers.

Because a surrogate marker is a biomarker intended to substitute for a clinical endpoint which is typically the primary end point of the trial—e.g. mortality, survival, onset of AIDS—having evidence on a clinical end point is different from having evidence on a surrogate marker. At an important HIV/AIDS conference in 1989, during the discussion

---

session over potential issues in the design of clinical trials, the topic of surrogate markers was discussed. U.S. based NIAID researcher Susan Ellenberg pointed out the following problem with the use of surrogate markers in HIV/AIDS treatments:

I would like to note that the presence of a statistical association between a marker and a clinical end point does not necessarily mean that the marker will be useful to distinguish between treatments. Demonstrating the prognostic value of a marker is different from demonstrating that it can be used reliably as an end point in a trial. There are actually two steps in the process of validating end points: one is to find a marker that is clearly prognostic for clinical outcome; the other is to determine if it actually differentiates treatment effect. (Journal of Acquired Immune Deficiency Syndromes, 1990, S75)

In other words, what Ellenberg and others before her have expressed is that, when a drug has a particular effect on CD4 count, we may not know if it has any effect on pathogenesis. It is one thing to have a good prognostic marker of the future course of a disease, yet quite another to know the effect of a treatment on the individual’s prognosis just by knowing the effect of the treatment on the individual’s marker. The surrogate marker may not be meaningful as an endpoint of a clinical drug trial. By knowing the individual’s CD4 cell count investigators may be able to predict the future onset of AIDS, but just knowing the effect of the treatment on the patient’s CD4 cell count may not tell investigators anything meaningful about his or her chances of survival.54

54 There is another point that might be obvious from this discussion, but perhaps worth reminding the reader. That is that even when the surrogate marker works as a good prognostic marker of the future of the disease, this only means that it is a good predictor in statistical terms. That is, even though CD4 cell counts can be predictive on average for a group of asymptomatics, CD4 cell count may not be predictive for any single individual in that group.
Additional uncertainties have been expressed by investigators about CD4 cell count and its ability to predict clinical end points. One is that investigators do not know what it means when the CD4 count goes up in response to an intervention like AZT; a closely related point is that “there is no evidence that antiretroviral therapy causes a biologically significant rise of CD4 count.” (ibid, Dr. Lawrence Corey, S71) I return to this difference between statistical and scientific evidence in section 2.3.1 when I explain the differences in the standards of proof demanded in RCTs.

Another source of uncertainty, related to the previous ones, involves the surrogate marker's natural variability. CD4 cell count, like other markers, has an inherent variability. CD4 cell count can exhibit substantial fluctuations within an individual. In the case of CD4 level, the variability may be due to natural fluctuations in patients depending on the time of day, his or her diet, hours of sleep, and other factors. Thus, the concern that investigators have is that when a highly variable parameter must be used as an end point, how should they deal with it, and how reduce its natural variability? Suppose the end point is the time when CD4 counts drop below 300 cells per mm. Although there is no consensus among researchers, one possibility is to require that the patient have two successive counts below 300 within a certain interval of time before considering her to have reached the end point. Different researchers may demand different criteria.55

2.2.1.4. Scientific uncertainty: some general observations

The focus of uncertainty over what was known and wasn’t can be seen as an expression of trade-offs between uncertainties about (a) HIV virology, namely, how to avoid or delay HIV mutation that might become drug resistant; and (b) the immune system and how best to delay the onset of AIDS. If my characterization of these potential uncertainties underlying HIV/AIDS trials is correct—or at least plausible—it seems as though that an approach to RCTs that would allow an explicit appreciation of the different sciences and their different uncertainties can be helpful in understanding interim monitoring decisions. A decision theoretic framework that explicitly recognizes the importance of context—in line with what I propose in the next chapter—might be particularly suitable for understanding conflicting interim monitoring decisions of RCTs.

With the different types of sciences and their uncertainties in mind, I hope to have shown that the evaluation of an HIV intervention via clinical trials is complex.

To sum up: two general observations regarding the scientific uncertainty in RCTs are in order:

A. Trade-offs in uncertainty among different sciences can play an important part in justifications in the early stopping decisions of these RCTs;
B. There is no good manner to accommodate such uncertainties into traditional statistical methods, thus there is need for a more general decision-theoretic approach with ways of representing the relevant sorts of uncertainty.

I shall now turn attention to a different but equally important source of difference among investigators when justifying early stop decisions: differences in judgment about the appropriate monitoring policies for guiding RCTs. This will serve to show the impact of a monitoring management style in accounting for possible disagreements in interim monitoring decisions.

2.2.2. Difference in judgments about RCT monitoring policies

According to the GCTS, “a DMC guided by a pre-specified statistical monitoring plan acceptable to both the DMC and the study leadership, will generally be charged with recommending early termination on the basis of a positive result only when the data are truly compelling and the risk of a false positive conclusion is acceptably low.” (GCTS, 17)

While CONSORT states that RCTs should report any plan for monitoring for benefit and rules for stopping the trial because of benefit, the ICH9 states that “the protocol should contain statistical plans for the interim analysis to prevent certain types of bias.”

56 “An interim analysis is any analysis intended to compare treatment arms with respect to efficacy or safety at any time prior to formal completion of a trial. Because the number, methods, and consequences of these comparisons affect the interpretation of the trial, all interim analyses should be carefully planned in advance and described in the
readers of RCTs find however is that time and time again—at least more often than one would like to see happening—both recommendations are either marginally followed or completely ignored, making it difficult to understand the reasons for early stopping.

The reason why a difference in judgments about RCT monitoring policies is an important factor in understanding early stopping decisions, is *not* that it is a consideration frequently brought up in disputes about justification per se, but rather that the existence of a statistical monitoring plan is mandated by official policy documents, and yet it is a policy rarely adhered to. Moreover, when DMCs do follow the policy, there are large disputes—e.g. between the U.S. and Europe—over the appropriate statistical monitoring policy.

Ellenberg says that although there is almost unanimity among researchers about the ethical necessity of monitoring interim data, “there are still widely disparate views on early stopping criteria.” (2003, 586) She adds that some—mostly in Europe—hold the view that trials should rarely, if ever, stop early since extreme evidence is needed to have an impact on clinical practice. In contrast, others—mostly in the US—believe that “once it has become clear that the question addressed by the trial has been answered with pre-specified statistical precision, the trial should be terminated.” (2003, 586)

It is interesting to note that in contrast to the U.S. ACTG protocol 019, the Concorde’s monitoring management style was incredibly casual:

---

protocol. Special circumstances may dictate the need for an interim analysis that was not defined at the start of a trial. In these cases, a protocol amendment describing the interim analysis should be completed prior to unblinded access to treatment comparison data.” (ICH 9, section 4.5 on Interim Analysis and Early Stopping)
The Data and Safety Monitoring Committee (DSMC) reviewed summary data on the primary endpoints every 4 to 6 months, mainly to assess toxicity, and reviewed two full interim analyses. The protocol specified that the DSMC should report to the chairmen of the Coordinating Committee if at any time or for any category of participants there was “clear evidence” of benefit (in terms of overall prognosis) or new information (e.g., results obtained from other studies), that might significantly alter clinical practice. The statistical and biological criteria for “clear evidence” were not specified precisely and were left to the discretion of the DSMC. [Emphasis added]

During the proceedings of a two day workshop held at the NIH on January of 1992 on the ‘Practical Issues in Data Monitoring of Clinical Trials’—sponsored by the NIAID, NHLBI and NCI—different perspectives for monitoring were presented by “clinical trials leaders from diverse settings” including academic, federal government, FDA, and the pharmaceutical industry, “both within the U.S.A and in Europe.” (1993, 415) During the opening remarks of the conference, despite the common belief that DMCs with their monitoring policies are “primarily responsible for the protection of the subjects in the trials, and secondarily responsible for assurance of the scientific validity of the trials” it is noted that “there is practically no available guidance in the literature on how [DMCs] are organized and how they go about their business.” (ibid 417)

During one of the discussion sessions in the conference, identifying another difference between regulatory practices in the U.S. and in Europe, Stuart Pocock is quoted as saying:

I think the regulatory science in Europe may have specific implications for trial practice that do not apply to the types of major trials we have been discussing which do not have a specific regulatory connotation. I think the regulatory authorities in Europe are generally less precise than the FDA as regards what they expect in the way of interim analyses. We're trying to get a single statistician within the Department of Health in Britain to have some function in regulatory concerns. There
are 50 or 60 in the U.S. FDA and there is not a single one in the British Medicines Control Agency, only external advisers. I think we stand a better chance of statistician input in the broader European realm with the CPMP developments. Recent progress has been made in Germany and in Sweden with statisticians now involved in the regulatory authorities, but the scale of statistical support in regulatory affairs is considerably less than over here. (Pocock 1993, 522)

So, why are the statistical monitoring rules important? They are important because they link the preliminary planning of the RCT with subsequent interpretations of its findings. If one of the goals of a normative framework for decision-making for RCTs is to help an analyst to interpret and carry out comparative analyses of early stopping decisions, then accounting for adopted RCT statistical rules is necessary. The supposition is that a proper interpretation of the DMC's decision requires one to know why the DMC did what it did, knowing the procedure by which the decision was made and then was (or wasn't) followed. There is much more to come in later chapters about statistical monitoring rules and their importance.

In my next case study I complement this brief analysis of the use of statistical monitoring plans for understanding early stopping decisions, by examining another consideration. The related consideration reflects the existence of differences and uncertainties over the meaning of statistical interim data: a consideration that I refer to as the difference in ‘standards of proof’ in RCTs. I examine this consideration by looking at an RCT case of early stop due to futility.
2.3. Toxoplasmic encephalitis (TE) in HIV-infected individuals

This case-study illustrates the sort of dilemma that investigators face when dealing with a case of early stop due to futility. I examine the type of dilemma underlying cases of early stop due to futility at a general level in chapter 3—see section 3.2.2.

During the early 1990s an important multi-center RCT was conducted in order to evaluate an intervention that could prevent a brain-infection in HIV+ individuals. Because toxoplasmic encephalitis (TE) is a major cause of morbidity and mortality among patients with HIV/AIDS and it had been known by then to be “the most common cause of intracerebral mass lesions” (Jacobson et al 1994, 384) an RCT was set to evaluate pyrimethamine in the prevention of TE.

The study was designed with a target of 600 patients followed for a period of 2+½ years, with an estimate that “a 50% reduction of TE with pyrimethamine could be detected with a power of .80 at a two-sided significance level of .05 if 30% of patients given placebo developed TE during follow-up.” (Jacobson et al 1994, 385) Survival was the primary end point of the study. By March 1992, at the time of the fourth interim analysis, while patients were still being enrolled, the investigators found that:

[T]he committee recommended that the study be terminated because the rate of TE was much lower than expected in the placebo group and was unlikely to increase appreciably during the planned duration of the study and because there was a trend toward rates of both TE and death being higher for patients given pyrimethamine than for patients given placebo. Thus, it was thought unlikely that pyrimethamine as used in this study was an effective prophylaxis against TE. (Jacobsen et al 1994, 386)
This RCT is a good example of how conditional power was important in assisting the DMC. Even though the original publication did not report on the specific type of conditional power computations that led to the early termination decision, nor did it specify the statistical monitoring plan adopted, it is clear from the publication that due to the unexpectedly low TE event rates observed during interim analysis, which compromised the original power of the study, and due to an early unfavorable trend in survival, the DMC recommended the early stopping of the trial. But the decision to early stop was not unanimous and was far from uncontroversial.

In a subsequent article authored by the original investigators, they explain that “a Haybittle-Peto interim analysis monitoring plan for early termination” was used by the DMC. (Neaton et al 2006, 321) Despite the DMC recommending that the trial be terminated, “the chair advocated continuing the study due to the uncertainties about the future TE event rate and about the association of pyrimethamine with increased mortality” while “the DMC reaffirmed their recommendation to stop the trial.” (Neaton et al 2006, 324) Among the “lessons learned” investigators expressed a desire “that procedures should be in place for adjudicating differences of opinion about early termination.” (Neaton et al 2006, 327)

Assuming that the chair and the DMC used the same statistical monitoring plan in arriving at their respective conflicting recommendations, including the same ethical principles, and similar judgments about the consequences of continuing or stopping the trial, how can we account for the difference between their recommendations, assuming both recommendations were reasonable?
One way is to base the difference on “the uncertainties about the future TE event rate” expressed by the chair, as quoted earlier. This consideration suggests that the chair might have adopted a different statistical method for computing conditional power when arriving at his recommendation. Although this is only a conjecture, it is a plausible one given that not only is it consistent with the limited report available to us through the original publication, but also that conditional power calculations are frequently made by DMCs when considering an unfavorable data trend. (Ellenberg et al 2003, 129) If my conjecture is correct, whereas the DMC used the frequentist conditional power based on a single pre-specified effect size, the chair computed conditional power considering a range of reasonable effect sizes with different weights to them. This method is a hybrid Bayesian-frequentist method of conditional power, and is known as the *predictive power* approach, which involves averaging the conditional power function over a range of reasonable effect sizes. In chapter 4, section 4.3 I propose a formalization of this dispute by following this conjecture.

One important note regarding the classification choice for this case: because this is a case where the choice of statistical inference is relevant to the disagreement—the DMC adopting a traditional way of computing conditional power vs. the principal investigator adopting a hybrid Bayesian-frequentist version of conditional power—I classify this case as first, a case of disagreement over standards of proof. I shall adopt this classification scheme even when the case reflects a disagreement about scientific uncertainty per se—as illustrated by the “uncertainties about the future TE event rate.” Whenever the choice of statistical inference rule is relevant to the dispute, classifying the case as a dispute over
standards of proof shall take precedence over classifying the case as a dispute about scientific uncertainty.

Given that there are rival approaches to statistical inference and evidence (Chapter 1) and because a common dissatisfaction among practitioners is that conventional statistical approaches to RCTs have become intertwined in practice—being mistakenly regarded as a single approach to inference—at this point, I would like to pursue this conjecture further. The suggestion is that disagreements between practitioners over interim monitoring decisions might be explained by differences in the statistical approach that investigators—explicitly or not—adopt. My conjecture here is that fundamental uncertainties over the very meaning of statistical evidence might be at play in some disagreements over monitoring decisions. I refer to such differences in judgments of statistical evidence as disagreements over the appropriate ‘standards of proof’ in RCTs. The following section pursues this theme.

2.3.1. Differences in judgments over the standards of proof in RCTs

Alongside the scientific uncertainties examined in section 2.2.1, there is also uncertainty of a more fundamental type: uncertainty about the meaning of statistical interim data. This is indecision over what the statistical evidence is telling the DMC at any given point during the course of a trial. Different statistical approaches can lead to different interim analyses. For instance, to reject or accept for a Neyman-Pearson test or to compute
a posterior probability for a Bayesian procedure are very different responses to what might be “the same interim data.”

Although many investigators have characterized modern statistics as mostly a split between the classical school and the Bayesian school, and at times with a third school—e.g. the likelihood school (cf. Senn 2003)—other statisticians have pointed out, and rightly so, that it is misleading to dichotomize statistical accounts of evidence and inference as either classical or Bayesian. Although dichotomizing can be helpful in simplifying descriptions, e.g. when used as an introduction to the different uses of probability in scientific reasoning, it does not do justice to the wide range of statistical techniques available to researchers today. Different taxonomies of statistics can be made of statistics as a science in the 21st century. I think a proper characterization would be a finer one. It would delineate the different statistical approaches among classical, likelihoodist, and Bayesian statistics, including techniques best described as hybrids of these schools. Examples of such finer characterization might include approaches such as Fisher’s likelihood, Neyman-Pearson-Wald, empirical Bayesian, orthodox Bayesian, decision-theoretic Bayesian, and classical decision theory. (cf. Spiegelhalter et al 2004)

With so many statistical approaches available, in this section I will focus on two of them only. I discuss the differences between the Fisherian and Neyman-Pearson approaches to statistical inference in relation to dissatisfactions from practitioners with regards to their use in RCTs. I choose these two approaches because they are typically considered the “conventional statistical approaches used in health-care evaluations,” (Spiegelhalter et al 2004, section 4.2) and were employed during the evaluations of HIV-
interventions, whether we are referring to RCTs designed and conducted in North America or Europe. But the reader should keep in mind that other statistical approaches are also used in health-care evaluations, although much less frequently. I shall focus on the pervasive notion of p-value, and the differences between the Fisherian and Neyman-Pearson approaches. Going into the two approaches will illustrate two points: (a) differences of opinion about statistical evidence exist, and (b) neither Fisher nor Neyman-Pearson intended their methods to be applied in an automatic fashion.

The Fisherian approach to statistical evidence and inference can be characterized as an estimation approach. It is based on the idea of the likelihood function, which expresses relative support given by the data to the possible different values of $\theta$—the unknown value of the intervention effect. The problematic aspect of the Fisherian approach comes from its application in experimental design vis-à-vis tests of significance, and from its notion of p-value. Fisher wanted a way to summarize experimental results as evidence against a specified hypothesis, with the strength of evidence expressed in terms of a probability value, the p-value. According to Fisher, “every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis” (1935, 16), so that the p-value

---

57 In his highly influential book *Statistical Methods for Research Workers*, while explaining his rejection of the Bayesian mode of inference, he says of the likelihood method:

> The rejection of the theory of inverse probability should not be taken to imply that we cannot draw, from knowledge of a sample, inferences respecting the corresponding population. Such a view would entirely deny validity to all experimental science. What is essential is that the mathematical concept of probability is, in most cases, inadequate to express our confidence or diffidence in making such inferences and that the mathematical quantity which appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. I have used the term ‘Likelihood’ to designate this quantity. (Fisher 1925, section 2 General Method, Calculation of Statistics)
becomes the measure of evidence against the specified hypothesis, i.e. against the null-hypothesis or the hypothesis of no difference between two groups.

A Fisherian p-value is a probability value, namely, the probability of getting a result equal to or more extreme than the one observed were the specified hypothesis true. As an inferential approach it judges the validity of the specified hypothesis via a rejection of the hypothesis, rather than emphasizing a ‘positive’ claim in favor of an informative alternative hypothesis. This idea of rejecting the null-hypothesis is best understood by considering how a p-value relates to yet another important notion, i.e. the meaning of a rare event.

According to Fisher the importance of a statistical test is to help the research determine whether two groups, treatment and no-treatment, can be considered groups from the same hypothetical population—thus the need to specify the hypothesis of no difference (on some characteristic of interest) between the two groups. The researcher is to infer, from a single occurrence of the experiment, whether the difference between two samples is “significant” (thus the name significance test), namely, whether the difference between the two samples is such that the researcher is licensed to infer that the hypothesis of no difference is to be rejected. Fisher originally suggested the criterion of a rare event as an event that happens less than 5 times in 100—if the hypothesis of no difference is

---

58 Some interpreters of Fisher identify him as a ‘falsificationist’ akin to Popper’s philosophy of science. I think this is a mistake and there are certainly misinterpretations here. A ‘falsificationist’ such as Popper, would not subscribe to a view of the likelihood to justify the validity (or corroboration) of experimental hypotheses. Like Fisher, Popper denies that we are allowed to say that a theory or hypothesis is likely to be true. But unlike Fisher, Popper would add that researchers are also not allowed to say that a theory or hypothesis is reliable or that it can be justified. Popper’s critical stance implies that it is rational to adopt a hypothesis tentatively, if our critical efforts to falsify it are unsuccessful, but that is because there is no more rational course of action available to us. Popperian falsificationism is well documented by philosophers of science, and not without its problems. But Fisher’s “falsificationism” is different, with its own set of issues.
assumed as true. This rejection criterion was for convenience only and never intended to work as a fixed criterion. It did not have to be made before the running of the experiment either. The issue of judging whether the event observed was or was not considered rare enough to warrant rejection of the null-hypothesis was supposed to be a matter of judgment and interpretation for the researcher to make, once the experiment was performed and the samples available, and not as an automatic and sacred criterion for testing hypotheses.

In his other equally important and influential work, *The Design of Experiments* (1935), Fisher puts his point about significance tests the following way:

It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. It is obvious that an experiment would be useless of which no possible result would satisfy him. (...) No such selection can eliminate the whole of the possible effect of chance coincidence, and if we accept this convenient convention, and agree that an event which would occur by chance only once in 70 trials is decidedly “significant,” in the statistical sense, we thereby admit that no isolated experiment, however, significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the “one chance in a million” will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us. In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. (Fisher 1935, pp 13-14)
“Fisher argued that p-values are good measures of the strength of evidence against a hypothesis.” (Berger 2003, 6) A p-value should indicate to the researcher the amount of evidence against the hypothesis of no difference between the two groups—i.e. that the two groups were coming from the same hypothetical population. From a Fisherian perspective to inference—to make a parallel case with the Neyman-Pearson approach which I examine in a minute—there are two alternative hypotheses between which the researcher have to choose, i.e. either “reject” the hypothesis of no difference or “accept” the hypothesis that the results are “inconclusive” given the experimental data. With the Neyman-Pearson approach, on the other hand, the choices are between “reject” the hypothesis of no difference in favor of a particular alternative, and fail to reject (“accept”) the hypothesis of no difference.

The adoption of a Fisherian approach has a clear impact on the conduct of RCTs. According to Spiegelhalter et al (2004), the Fisherian approach is perhaps best exemplified in RCTs that are designed and conducted by the Clinical Trial and Services Unit (CTSU) in Oxford, UK, in which protocols generally state that the DMC should only alert the steering committee to stop the trial on efficacy grounds if there is both “(a) proof beyond reasonable doubt that for all, or for some, types of patient one particular treatment is clearly indicated” and “(b) evidence that might reasonably be expected to influence the patient management of many clinicians who are already aware of the results of other studies”. (ibid 203) The authors contend that if one follows what the CTSU prescribes, |

“there is no formal expression of what evidence is required to establish ‘proof beyond reasonable doubt’ ” although ‘extreme p-values’ is mentioned as a criterion. This last point is illustrated by the description found in section 5.6 of the CTSU’s report on the use of “statistical methods” in the following way:

In an unbiased overview of many properly randomized trial results the number of patients involved may be large, so even a realistically moderate treatment difference may produce a really extreme p-value (e.g. 2p<0.0001). When this happens, the p-value alone is generally sufficient to provide proof beyond reasonable doubt that a real treatment effect does exist. In contrast, p-values that are not conventionally significant, or that are only moderately significant (e.g. 2p=0.01), may be much more difficult to interpret appropriately. (CTSU report on statistical methods, section 5.6; emphasis is mine)⁶⁰

On the other side of the Atlantic, more specifically in the U.S., the reliance on the part of regulatory agencies such as the NIH and the FDA is based on the very different approach due to Neyman-Pearson (NP).⁶¹ NP provides a more formal testing approach than the Fisherian significance test and its corresponding notion of evidence through a p-value. Even though the NP approach was built on similar ideas to those originally expressed by Fisher, the NP approach grew out of an attempt to provide tests within a more formal epistemological rationale than Fisher’s by considering alternative hypotheses—and consequently introducing the probability of a type II error. This leads to a very different perspective on statistical inference.

⁶⁰ http://www.ctsu.ox.ac.uk/reports/ebctcg-1990/section5

⁶¹ Senn (2003) reports that “mathematical statistics as practiced in the USA in its frequentist form can be largely regarded as being the school of Neyman.” (2003, 84) Although I am speculating, I think part of this predicament might be due to Neyman having spent most of his professional career in the U.S., as opposed to Fisher’s career which he spent mostly in the UK. Neyman moved to Berkeley, CA in 1938 where he was to spend the rest of his life including much of his prolific professional career.
Some statisticians say that neither Neyman nor Pearson was interested in scientific inference per se, but rather in scientific decisions. Although this distinction is correct, it can be quite limited in explaining their departure from Fisher. The distinction is meant to characterize the fact that the NP approach gives importance to the difference between two possibly erroneous actions, i.e. “accepting” and “rejecting” the hypothesis of no difference. The researcher chooses a level of significance based on the two actions with (presumably) their consequences in mind. By contrast, the Fisherian approach has nothing to say about such consequences.

The NP approach—sometimes described as Neyman-Pearson-Wald—contrasts with the Fisherian approach in another essential way. It does not attempt to judge the validity of any given hypothesis. The essential idea is based on a general procedure for making decisions involving a formal rule—with the choice of a level of significance—stated in advance of the experiment yielding a specified proportion of valid decisions in the long-run. For any particular instance of the testing procedure, the results of the test yield a decision according to what the pre-specified formal rule prescribes. The justification for adopting such a procedure is supposedly a prudential one, namely, that making decisions in this manner—according to what the rule dictates—even though it does not commit the researcher to infer anything about the probability, likelihood, or possibility of the hypothesis being true, correct, or not, reduces the chances of the researcher making mistakes in the long-run. No ‘interpretation’ of the experimental results is given other than that the tested hypothesis is either rejected or fails to be rejected (‘accepted’) on each instance of the test.
The potential decisions which the researcher can make are thus no more than guides to action, interpreted in a fixed framework of repeated tests. Neyman gave the interpretation that their approach provides “inductive behavior” to the researcher who is seeking procedures on which he or she could rely for testing hypotheses, with the procedures satisfying certain properties in long-run repeated use.\footnote{Pearson rejected this interpretation though, crediting ‘inductive behavior’ to Neyman only. In a notorious exchange with Fisher, while attempting to explain the epistemic rationale of the Neyman-Pearson approach he says their approach should be seen as one in the spirit of what Fisher envisioned, i.e. as a means of helping the research to learn something about what is tested:}

Another important point of difference between the two approaches relates to how probabilities are used. In NP, probabilities are meant to describe properties of the testing procedure. The focus is on the probability of making different types of error when making decisions on the basis of the observed data. For instance, a common choice of probability of making errors is to fix type I error, \( \alpha \), as 5\% and probability of type II error, \( \beta \), as 20\%, meaning having power of 80\%, i.e. \( 1-\beta \). Fisher rejected such an approach, accusing it as suitable for industrial planning and sampling, but not for drawing inferences in scientific experiments.\footnote{In an interesting exchange between Fisher and Pearson in 1955, Fisher attacked the Neyman-Pearson school as being one akin to Russian industrial planning, that is, the Neyman-Pearson approach is much like

[R]ussians [who] are made familiar with the idea that research in pure science can and should be geared to technological performance, in the comprehensive organized effort of a five-year plan for the nation. (Fisher 1955, 70)}
Fisher's intent was that his significance test and p-value were only guides to the researcher; they were not supposed to be adopted as cook book recipes. Similarly, in their seminal work “On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference” published in 1928, Neyman and Pearson say—in the first paragraph—that their approach is only intended as “tools to help the worker to make a final decision” on his or her hypothesis of interest. They say:

ONE of the most common as well as most important problems which arise in the interpretation of statistical results, is that of deciding whether or not a particular sample may be judged as likely to have been randomly drawn from a certain population, whose form may be either completely or only partially specified. (...) In general the method of procedure is to apply certain tests or criteria, the results of which will enable the investigator to decide with a greater or less degree of confidence whether to accept or reject Hypothesis A, or, as is often the case, will show him that further data are required before a decision can be reached. At first sight the problem may be thought to be a simple one, but upon fuller examination one is forced to the conclusion that in many cases there is probably no single “best” method of solution. The sum total of the reasons which will weigh with the investigator in accepting or rejecting the hypothesis can very rarely be expressed in numerical terms. All that is possible for him is to balance the results of a mathematical summary, formed upon certain assumptions, against other less precise impressions based upon à priori or à posteriori considerations. The tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision; one man may prefer to use one method, a second another, and yet in the long run there may be little to choose between the value of their conclusions. What is of chief importance in order that a sound judgment may be formed is that the method adopted, its scope and its limitations, should be clearly understood, and it is because we believe this often not to be the case that it has seemed worthwhile to us to discuss the principles involved in some detail and to illustrate their application to certain important sampling tests. (Neyman, J. and Pearson, E. S. 1928, pp 175-176)

In the U.S. also the great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money. (Ibid)
According to Hogben (1973) the Fisherian statistical approach is a “backward look” approach—as opposed to the Neyman-Pearson-Wald approach which is a “forward look” approach. That is because the Fisherian approach involves retrospective interpretation of the information produced by a significance test to infer the validity of the specified hypothesis in probabilistic terms; by contrast, the Neyman-Pearson-Wald approach involves the application of probability theory to a given empirical outcome to determine the ratio of correct decisions about future outcomes of tests of a hypothesis.

Besides the objection to automaticity in application of statistical methods (cf. Morrison and Henkel 1973), a common point of dissatisfaction among practitioners is illustrated by those voicing dissatisfaction over the use and interpretation of conventional statistical methods. Here, the blame goes to the ambiguities and complexities underlying such methods. Goodman (1999) claims that despite the incompatibility of the Fisherian and Neyman-Pearson school, the methods “have become so intertwined that they are mistakenly regarded as part of a single, coherent approach to statistical inference.” (997)

---

64 Wald’s approach is often seen as an extension and generalization of the Neyman-Pearson approach. One contrasting feature of Wald’s approach to his predecessors is that:

The theories of testing hypothesis, as developed during the last thirty years by Fisher, Neyman and Pearson, and their schools, deal almost exclusively with the case where experimentation is carried out in a single stage; i.e., it is determined in advance (before experimentation starts) how many and what kind of observations should be made during the whole course of experimentation (Wald 1950, pp 18-19) Contrast with Wald’s approach where experimentation is carried out in multiple stages.

65 In explaining how the two incompatible approaches became one single approach in practice, Goodman explains that even though the story is complex, he says:

[T]he basic theme is that therapeutic reformers in academic medicine and in government, along with medical researchers and journal editors, found it enormously useful to have a quantitative methodology that ostensibly generated conclusions independent of the persons performing the experiment. It was believed that because the methods were “objective,” they necessarily produced reliable, “scientific” conclusions that could serve as the bases for therapeutic decisions and
Lang et al (1998) protests that “there has not been vigor [from the part of journal editors] in disentangling the components of a p-value (...) it continues to be used as a measure of the importance and credibility of study results” and with the editors of the journal Epidemiology adding that “however regrettable, the practice of calculating and reporting p-values is nearly ubiquitous” but that “it would be too dogmatic simply to ban the reporting of all p-values” from the journal. (1998, 8)

Good and Hardin (2006), on the other hand, while reporting that there are over 300 references of warning of misuse of Fisherian and NP approaches to testing hypotheses, contend that “the majority of these warnings are ill informed,” stressing errors that will not arise if practitioners were to proceed as the authors of those approaches recommended researchers doing in practice. “Placing the emphasis on the why, not the what of statistical procedures. Use statistics as a guide to decision-making rather than a mandate.” (2006, 25)

I think there is an important difficulty with conventional statistical tests, especially for RCTs, that is mostly ignored in discussions of limitations or ambiguities involving such methods. Consider the Neyman-Pearson paradigm of hypothesis testing. The fundamental idea expressed by a “minimum clinically significant difference” between treatments conflates a few important issues. First, there is the question of expectation, namely, what the effect is to be expected from the new treatment—e.g. 10% or 25% efficacy. When the government policy. This method thus facilitated a subtle change in the balance of medical authority from those with knowledge of the biological basis of medicine toward those with knowledge of quantitative methods, or toward the quantitative results alone, as though the numbers somehow spoke for themselves. This is manifest today in the rise of the evidence-based medicine paradigm, which occasionally raises hackles by suggesting that information about biological mechanisms does not merit the label “evidence” when medical interventions are evaluated. (ibid 1001)
experiment is first designed, and the choice of the alternative hypothesis is made, the researcher must choose it according to what he (or the DMC) thinks the effect is. They have an expectation about an aspect of the world—true state of nature—i.e. an expectation regarding the reality of the treatment effect—e.g., how often, how big of an effect. This expectation is most likely informed by previous studies, including phase I and phase II trials. And the expectation plays a critical role in Neyman-Pearson tests.

A second issue is the question of what to stipulate, namely, what effect to demand from the treatment in order to consider the treatment worthwhile. This is a question of practical or ‘scientific’ significance. A small effect or just a tiny bit of improvement over the standard treatment (or placebo in case of a placebo-control trial) may not merit attention from researchers and policy makers after all. In this case, when the experiment is initially designed, researchers must choose the alternative hypothesis according to what they think is important, as worth investigating, worth spending time and effort to test. They must bring some valuation up-front regarding the treatment effect. This valuation is likely to be informed by a number of contextual factors, e.g. which treatments are currently available to patients, what sort of population they have in mind for the drug, what sort of urgency the situation involves, how much money researchers have at their disposal, and other factors.

A third set of considerations the researcher must have, when designing the RCT with the choice of the alternative hypothesis, concerns the particular choice of probabilities of type I and type II error. Not so much the particular level of significance alone—although the particular level is of relevance—but the ratio of probabilities between type I and type II errors. For instance, why 1 to 4 (0.05/0.20)? The choice reflects considerations of expected
losses, utility, or costs attendant upon type I and II errors in some form or another—but which are never fully explicit nor explained in the justification of such a test. Why should RCT stakeholders fear rejecting a true hypothesis more than failing to reject a false one—thereby rejecting a true alternative—by a ratio of 1:4? This is a normative judgment in which trial participants should be given a chance to weigh in.

These considerations are all buried within an NP RCT design. Within a decision theoretic framework however, at least along the lines the one I propose in my next chapter, these ideas are better served when made separate and explicit. To the extent that their demarcation and specification are possible, there would be a point of advantage over conventional statistical methods that conflate these considerations for early stopping decisions.

To sum up, there is a plurality of statistical approaches available to researchers. The same interim data can mean different things to researchers. The case of toxoplastic encephalitis in HIV-infected individuals illustrates this point. Moreover, even when sharing the same school of statistics (e.g. conventional statistics), the demands of different regulatory agencies can reflect different preferences over statistical approaches in addition to differences in monitoring policies—e.g. the U.S FDA’s preference for an NP approach, compared to the UK’s CTSU with a preference for a Fisherian one.

In my next case study I build a stronger case for the need of a decision theoretic framework for interim monitoring. By looking at another consideration for early stopping, the balancing between long-term versus short-term drug effects in RCTs, I discuss a case of
disagreement over an early stop due to harm case. The interim decision involved in the trial focused on an increased risk of HIV transmission from mother-to-child through breastfeeding.

2.4. Vitamin A as prevention of HIV transmission from mother-to-child through breastfeeding

The decision to stop a trial early is difficult. The decision to stop a trial early due to harm is typically the most difficult of early stop scenarios. That is because it involves a trade-off between an apparent yet possibly spurious harm and a potential yet undemonstrated benefit. This case illustrates the sort of dilemma that investigators face when dealing with a case of early stop due to harm. In this particular instance there was a tension between two conflicting undesirable outcomes to be avoided. The first was stopping too early on modest evidence of a potential increase of HIV-infection mother-to-child. That meant the abandonment of vitamin A as a supplement that might have proven beneficial had more data been allowed to accumulate in the trial. The second was permitting the trial to continue in the face of growing evidence of harm, thereby risking greater exposure of HIV from mother-to-child over a longer period of time.

Cases of early stop due to harm share similarities with cases of early stop for efficacy. The decision to stop also involves a trade-off between a potential yet undemonstrated effect and an early apparent effect. In harm cases the beneficial effect is the yet undemonstrated effect, whereas the apparent—but possibly spurious—effect is the
adverse effect. One essential difference between the trade-offs involved in harm cases and efficacy cases is that the amount of evidence for establishing harmful effect is significantly different than the amount demanded by early stop cases for benefit. Because it is appropriate to demand less evidence for harm to justify early termination than would be appropriate for cases of early stop for efficacy—a point which is explained in section 3.2.3—the nature of the dilemma faced by DMCs in harm cases is different than in cases for efficacy. I examine the fundamental dilemma underlying cases of early stop due to harm in more detail in chapter 3—see section 3.2.3.

HIV infection was—and continues to be—a major public health problem in many developing countries, particularly in sub-Saharan Africa where 20 to 45% of children born to HIV-infected women become infected during the gestation or breastfeeding periods. Because of this, there was a need to know whether vitamin supplements could reduce HIV transmission from mother-to-child. From a public health perspective this seemed like a good idea, given that vitamin supplements are considered a low-cost intervention in countries where antiretroviral drugs are cost-prohibitive and consequently unavailable.

Researchers from the NIID and the Harvard School of Public Health, in collaboration with the Ministry of Health in Tanzania, set-up an RCT to test whether vitamin A could reduce vertical transmission. This was a double-blind, placebo-control trial examining the effects of vitamin A supplementation versus multivitamins—excluding vitamin A—on the transmission of HIV from mother-to-child. HIV transmission or child death within the first
2 years \(^{66}\) were the primary endpoints of the study. This was a study with 1078 HIV-infected pregnant women, and a sample size calculated with 90% power to detect a 30% efficacy on the risk of transmission, 0.05 type I error, and a background rate of transmission assumed as 30%.\(^{67}\) "Likelihood ratio tests were used to assess the statistical significance of the interactions." (Fawzi et al 2002, 1937)

In September of 2000, the DMC recommended that the trial with the group given vitamin A in the study be stopped, because vitamin A had a negative, harmful, impact on one of the primary endpoints. Specifically, it increased the risk of HIV transmission from mother-to-child through breastfeeding—contrary to what conventional wisdom had suggested was going to happen. Of the children whose mothers did not receive vitamin A, 15.4% were infected by 6 weeks, compared with 21.2% of those whose mothers received vitamin A. (ibid 1938)\(^{68}\)

Little is known about how common early stop due to harm studies are. (Mills et al 2006) This trial is a typical example of how RCT reporting is deficient in several important ways. Despite CONSORT guidelines explicitly stating that RCTs “should report any plan for monitoring for harms and rules for stopping the trial because of harms,” in this particular instance, neither the stopping rule nor the monitoring plan—e.g., the number of interim analyses and a list of predefined adverse events—were reported. This makes it very

\(^{66}\) Including infection or death during the intrauterine period.
\(^{68}\) RR = 1.38, 95% CI 1.09-1.76; p-value=0.01
difficult for a decision analyst to understand what the DMC’s rationale was when deciding to stop.

In the background of Fawzi et al, there were results from another important study about vitamin A deficiency in an undernourished population. It was a known fact that vitamin A deficiency was common among women in developing countries, and widely prevalent in south Asia. Vitamin A deficiency was associated with increased risks of urinary and reproductive infections, and predisposed women to increased mortality. Until the mid-1990’s, however, concern about maternal vitamin A deficiency had focused on its effects on fetal and infant survival. Little attention had been paid to its effects on the health consequences for women, until a double-blind RCT was conducted to assess whether supplementation of pregnant women with vitamin A could reduce pregnancy-related mortality.69 “The results of the trial had shown that taking vitamin A on a weekly basis reduced mortality related to pregnancy by 40%, with effect estimates being similar during pregnancy and post partum.” (West et al 1999, 575) That is, preventing maternal vitamin A deficiency in undernourished populations (e.g. Nepal) lowered the risk of mortality during and after pregnancy. This was a very encouraging result for public health policy.

Other studies regarding vitamin A supplementation had also shown that vitamin A reduced mortality and morbidity rates among African children suffering from HIV. While there were observational studies that had suggested that vitamin A concentrations in plasma were inversely associated with the risk of transmission, other similar studies

suggested no relationship. In the discussion section of the Fawzi et al publication, the authors report:

We observed more cases of HIV-1 infection among children of mothers in the vitamin A arm compared with those who did not receive vitamin A. The increased occurrence of infection in the vitamin A group was unexpected. Data from two other randomized, placebo-controlled trials from Malawi [29] and South Africa [30] found no effect of vitamin A on transmission. There are, however, differences between our trial and the latter two trials. In both trials, the supplements were given during the antenatal period only, whereas in the Tanzania trial supplementation continued during the antenatal and breastfeeding periods. (Fawzi et al 2002, 1942)

The authors added the following disclaimer in their final report regarding what might have caused the increase in HIV transmission:

An alternative explanation to the findings of vitamin A supplements and HIV transmission may be that the supplements prolonged the survival of HIV-uninfected children who were then exposed to HIV-infected breast milk for a longer period. (Fawzi et al 2002, 1942)

With this last quote in mind, it is here that a decision analyst might want to understand the rationale for the DMC's early stopping decision. That is, given (1) that two other randomized, placebo-controlled trials (one from Malawi and one from South Africa) had found no effect of vitamin A on transmission, and given (2) that vitamin A supplements had proven beneficial to pregnant HIV-uninfected women by showing “significant reduction in maternal mortality among women from Nepal” (ibid 1942)—as acknowledged by the authors of the study—how did the DMC balance the short-term harmful effects of
HIV transmission against the possible long-term beneficial effects of vitamin A in the face of these uncertainties?  

In the next section, I begin to explore this question by examining how a decision analyst might want to go about representing this case of early stopping due to harm, as a means to understand the DMC’s rationale during the trial. A more extensive discussion will follow in chapter 4.

2.4.1. Differences in judgments about the balance of long-term versus short-term treatment effects

The study illustrates several interesting points about RCTs: conventional wisdom can be wrong, alternative explanation is difficult to rule out in the face of uncertainty, similar studies can have contradictory results, and researchers make implicit use of multiple sources of evidence. Prior to this RCT, vitamin A was considered as providing added benefit to individuals with HIV. Given vitamin A’s benefit history, the DMC had to weigh in some fashion the balance between obtaining more evidence and safeguarding the interests of mothers and children in the trial. If the trial had continued beyond the point of the observed evidence, it would have placed children in the trial at unnecessary risk. On the other hand, it may be difficult to reverse the decision to abandon vitamin A—as a matter of public policy—if early stopping is later found to be an erroneous decision, hindering the

70 “This may not be cost-effective to pursue unless such a counseling program is already in place as part of a strategy to reduce mother-to-child transmission.” (ibid 1942)
adoption of vitamin A supplementation as a means of reducing mortality and morbidity rates among African mothers suffering from HIV. Although it is difficult to explicitly model such potential losses with a high degree of plausibility, an attempt—although limited—might still prove illuminating for understanding the DMC’s rationale.

The choice between vitamin A supplementation and multivitamins without vitamin A is not a choice between two supplements that can be decided lightly and exclusively on the interim data. That is because given the background of the circumstances we find in the RCT, a justification is required before any change in vitamin A supplements is adopted as a matter of public policy. Even though understanding the statistical monitoring plan and statistical evidence can be helpful for understanding the reason for the DMC’s decision, the totality of evidence must be considered in light of earlier studies.

The experience of studying vitamin A as prevention of HIV transmission from mother-to-child illustrates how the DMC’s decision to stop the trial early due to harm requires more insight into the interim data, their interpretation, and the possible consequences of their actions than just following what could have been an implicit statistical stopping rule. In chapter 4, section 4.4 I return to this case by proposing a formalization of the balancing between short-term versus late-term effects of vitamin A as an intervention.
2.5. Conclusion

In summary, early stopping of RCTs involving HIV/AIDS treatments are difficult. The reporting of early stopping decisions in such trials is deficient in several ways. Proper disclosure of the statistical methods and interim monitoring policies are often lacking, making it difficult for anyone interested in understanding the results and recommendations that follow from such trials. There is a need for a clear framework for representing and evaluating early stopping decisions given the disagreements seen during the monitoring of such trials.

In this chapter, we have explored three case studies and identified four key factors that shape early stopping decisions: scientific uncertainties, disagreements about the statistical monitoring plan, theoretical differences about statistical inference and what constitutes a relevant standard of proof, and the need to balance long- and short-term treatment effects. A decision theoretic framework that has the potential to account for these factors remains, so far, unavailable. Without such a framework, it is likely that early stopping decisions will remain unclear and poorly justified; presently and at best, DMC reports address just one or two of these points in a cursory manner. In chapter 4, I propose a decision theoretic framework for dealing with this predicament.
Chapter 3. Qualitative framework

In this chapter I articulate a comprehensive, and mostly qualitative, approach to the evaluation of early stopping decisions of RCTs. I shall refer to this non-technical treatment as the qualitative framework.

What does it take for an early stopping decision of an RCT to be considered a good decision? This question is important not only to philosophers or DMC members, but also to anyone who wants to evaluate early stopping decisions of RCTs. Yet as we have seen (in part in chapter 2, and in part in chapter 1), most approaches to stopping rules do not discriminate effectively between good and bad decisions. While statistical approaches tend to focus on the epistemic aspects of stopping rules (cf. Proschan, Lan, and Wittes 2006), often overlooking ethical considerations, ethical approaches to RCTs tend to neglect the epistemic dimension (cf. Freedman 1987, 1996). In this chapter, I shall answer the question by developing a comprehensive, but mostly qualitative, framework that incorporates both ethical and epistemic considerations.

---

71 Proschan, Lan, and Wittes (2006) is a good example of this disregard. Their unified approach aims at modeling the mechanism of accumulating data in RCTs as following a single motion, which they call “Brownian motion”. Throughout their approach, they “stress the need for statistical rigor in creating an upper boundary.” While the “upper boundary” of monitoring rules aims at demonstrating the efficacy of new treatments, “lower boundaries” deal with harm, or “unacceptable risk”. Having said that, “most of [their approach] deals with the upper boundary [problem] since it reflects the statistical goals of the study and allows formal statistical inference. But the reader needs to recognize that considerations for building the lower boundary (or for monitoring safety in a study without a boundary) differ importantly from the approaches to the upper boundary.” (2006 p. 6) We find a similar approach also in Moyé (2006). Neither study devotes serious attention to safety, despite acknowledging its significance.
3.1 Preliminaries

As seen in section 1.5, the following summarizes our current predicament with RCTs: (1) a growing number of RCTs stopped early; (2) many researchers not adhering to standards of conduct for RCTs; (3) frequent omission of important information from RCT publications; and (4) public mistrust of RCTs. To address these problems, and to enable readers both to better understand a trial’s conduct and to be in a better position to assess the validity and legitimacy of its results, we need a way to represent (or ‘reconstruct’), in a plausible fashion, the interim monitoring decisions of RCTs.

The qualitative framework to be developed below will address a fundamental and common question about RCTs: what does it take for an early stopping decision to be a good decision? To answer this question, the two most basic tasks of the framework are to deliver:

1. A reconstructed justification for a given interim monitoring decision; and
2. A verdict as to whether the decision is ethically justifiable (permissible).

The framework should also, if possible, provide a basis for making comparative judgments about the permissibility of different and at times conflicting early stopping decisions.
Based on the HIV/AIDS case studies introduced in the preceding chapter, I propose that the qualitative framework should meet certain requirements. This preliminary section sets out these basic requirements and indicates some important restrictions.

### 3.1.1. Basic requirements

1. **Clarity.** The framework must provide clear standards of representation and clear criteria of evaluation. That is, we need to be clear both about how an early stopping decision is represented and about the principles used for its evaluation.

2. **Scope.** The guidelines for representation must be flexible enough to accommodate the wide variety of early stopping decisions. That is because different types of early stopping decisions reflect different types of dilemma, which consequently refer to contextual factors differently. In this manner, the framework must be wide enough to accommodate: (1) the different types of early stopping decisions found in RCTs; (2) common disagreements about decisions based on them; and (3) common types of statistical monitoring procedures—whether based on frequentist or Bayesian principles of inference—and their respective stopping rules as often implemented and guiding these decisions in RCTs.

3. **Flexibility.** The framework should yield answers that vary with changes in problem representation. The framework assists with the identification of a set of relevant decision criteria and with representation of the monitoring rules that govern stopping
decisions. Having identified these criteria, the framework tries to represent decisions as optimal, in the following ‘weak sense’ of optimality: no alternative does better on every criterion.

4. *Explanatory power.* Our framework should provide criteria amenable to justification. An adequate framework should show how good early stopping decisions actually contribute to the permissibility of these decisions.

5. *Balancing epistemic and ethical factors.* Frameworks currently available either tend to focus on the epistemic performance of monitoring rules—such as those found in the statistical literature—or to focus on meeting the specific demands of medical ethics applied to clinical research, such as those proposed by the notion of “equipoise disturbed” (cf. Freedman 1987), or “therapeutic obligation” (cf. Marquis 1983). Such approaches lead to the view that early stopping decisions should be assessed on the basis of either ‘scientific’ or ‘medical’ objectives. My framework, by contrast, represents early stopping decisions by combining both objectives on the basis of factors contextualized in a particular decision situation.

In order to meet these requirements, the first step is to introduce a basic classification of early stopping decisions into three main types (presented in sections 3.2 and 3.3). This preliminary classification allows us to identify relevant criteria that might influence the decision. The second step, given our identification of a set of criteria for representing and evaluating early stopping decisions, and given the existence of a
philosophical theory of rational decision-making, is to develop a specialized decision-theoretic model. This means defining the components of the decision-theoretic model:

(1) A set of available actions (at a given interim point during the trial);

(2) The foreseeable consequences of each action;

(3) The set of possible states of the world (i.e. a set of hypotheses about the effect of interest);

(4) A set of alternative statistical monitoring rules; and

(5) Decision criteria.

I call this a complete DMC decision specification.

A fundamental constraint on the decision specification is the *weak optimality* requirement: having identified a set of relevant decision criteria, concentrate on representations under which actual DMC decisions are *weakly optimal*, in the sense that no alternative under consideration fares better on every criterion. Conflicts of opinion can then be analyzed by exploring the distinct ranges of assumptions and scenarios under which conflicting stopping strategies emerge as optimal choices.

This broadly Bayesian approach offers the hope of meeting all five requirements listed above. Because we have the components of decision theory *within* the model, the framework allows us to explicitly account for the epistemic and ethical components that should influence the interim monitoring decision of RCTs. For instance, by being able to represent interim decision options (either to stop or continue) as decisions under
uncertainty, each interim decision will in turn, depend on epistemic elements (e.g., population frequencies, predictive probability of the test statistic given the stopping rule) and ethical elements (e.g., costs and benefits of possible treatment outcomes). In this light, the decision model can meet the balancing requirement of epistemic and ethical factors.

Our modeling approach is also able to meet scope by being able to account for the different types of early stopping decisions we find in practice (via the classification of section 3.3), while also meeting clarity by setting explicit standards for representing and evaluating interim monitoring decisions (sections 3.5 and 3.6).

3.1.2. Restrictions

To conclude this preliminary discussion, I note two important restrictions to the decision-theoretic model. The first restriction is unavoidable: the decision-theoretic model is capable of representing only some of the criteria that might, in real life, influence data monitoring decisions. The second restriction is that I do not aim to provide an algorithm for making early stopping decisions. I am not proposing a theory that takes a description of an RCT decision as input and delivers a sharp verdict about whether or not the trial should be stopped. Instead, I offer a broad general framework that allows us to explore and analyze the justification for stopping decisions. The qualitative version of the framework is presented in this chapter; the next chapter develops a simplified formal model together with detailed application to the three historical examples of early stopping decision described in chapter 2.
3.2 Overview of the framework

As stated in the preliminary section, the main objective of the proposed framework is to address the following question: What does it take for an early stopping decision to be a good decision? My approach to this question, as noted, will be to reconstruct a justification for any given interim monitoring decision, and then to evaluate how well that justification succeeds, either in isolation or (more commonly) in comparison to a rival monitoring decision. Inasmuch as my framework utilizes the resources of decision theory, it counts as broadly Bayesian and allows us to combine the epistemic and ethical components that should influence the decision. We do not need to rely on a purely statistical model that is either error-statistical (E-S) or narrowly Bayesian (i.e., just Bayesian statistics).

Before we proceed with the overview of the framework, it will be helpful to become clearer on what the framework is supposed to do and what it is not supposed to do. The decision framework is not intended as a re-construction of what DMCs actually have done, or often do, when faced with the complexity of whether to stop or continue their RCTs. Deliberations from DMCs are unavailable. If they exist, they do not exist in the public sphere. When they do exist they are often relegated to “Appendix 16.1.9” of study reports prepared and submitted to regulatory authorities such as the U.S. FDA. In rare circumstances, there is an actual case study identifying and discussing “lessons learned”

72 According to ICH E3, “Any operating instructions or procedures used for interim analyses should be described. The minutes of meetings of any data monitoring group and any data reports reviewed at those meetings, particularly a meeting that led to a change in the protocol or early termination of the study, may be helpful and should be provided in Appendix 16.1.9. Data monitoring without code-breaking should also be described, even if this kind of monitoring is considered to cause no increase in type I error.” (ibid p. 21)
specific to the RCT (cf. DeMets, Furberg, and Friedman 2006) and even then, relevant factors of interim decisions are often missing, making the public transcript an unreliable basis for anyone trying to appraise DMCs’ deliberations.

Here is a bird’s-eye view of how the framework does what it is supposed to do. The application of the framework for approaching an early stopping decision involves four general steps: classification, identification, reconstruction, and evaluation of interim monitoring decisions— which includes early stopping decisions. First, there is the job of classifying the type of early stopping decision in the case (sections 3.2.1 and 3.3), which leads us to the type of dilemma that the DMC had to face. The second step is the identification of salient factors, i.e. the “material facts” of the case, which is assisted by the identification of the particular DMC’s dilemma (sections 3.2.2 and 3.3). The third step is reconstruction (sections 3.2.3 and 3.4). This task reconstructs the DMC’s decision as ‘optimal,’ using the weak sense of optimality noted in the preceding section. And fourth, there is the task of evaluating, via a two-stage process, whether the DMC’s early stopping decision is in fact permissible (sections 3.2.4 and 3.5).

The exposition of the framework begins with concrete examples of RCTs. At times there will be clear disagreements between experts about the specific early stopping decision— e.g. between the primary investigator of the study and its DMC (as seen in section 2.3) or between two famous biostatisticians. At times the disagreements are between different DMCs conducting similar trials (as seen in section 2.2), or between different public health policies. At other times the disagreements are found simply “within” (you) the reader— the decision analyst— as when we find ourselves in a dilemma and
conflicted about a hard-case clinical trial (cf. Stanev 2011, Stanev 2012). So how does the framework work with a disagreement and how do we reason through it? Figure 3.1 below gives an overall picture of the proposed framework with its several components and stages necessary for representing and evaluating early stopping decisions of RCTs. I introduce each stage with its essential elements in sections 3.2.1 – 3.2.4.
Figure 3.1 Representation of the framework’s main elements and stages used for the classification, identification, reconstruction, and evaluation of early stopping decisions.
3.2.1. Classification

The framework begins with a concrete RCT case—or a pair of concrete and comparable RCT cases—stripped of some of its complex features. The simplification allows the decision analyst to focus on a limited set of questions about the early stopping decision at stake. The particular factors and details that are not stripped vary with the specific type of early stopping decision. An alternative way of putting this point is that different factors are held fixed depending on the type of decision. This means that the first step of the framework is to provide a simplified description of a concrete example of early stopping, organized around one of the three models that correspond to the three most common types of early stopping decision: efficacy, harm or futility (referring here to the classification step of Figure 3.1). I shall say a little here about each of these three concepts.

3.2.1.1. Efficacy

Although the results of RCTs can be expressed in a number of ways, by far the most common way is in terms of efficacy. Efficacy (following (Gordis 2009)) is the extent of reduction in the rate of occurrence of the primary event of interest—e.g., death or the development of a disease or complication—by means of a new intervention. For any new drug, we say that efficacy = \[ ((\text{rate in those who received placebo}) - (\text{rate in those who received the drug})) / (\text{rate in those who received the placebo}) \]. The rate, or risk, of death, disease or complication is calculated for each group, and the \textit{reduction in risk} is what we refer to as efficacy (Gordis 2009, 152). I revisit the term in more details in section 3.3.1.
should note, however, that there are other approaches to expressing RCT results, such as calculating the ratio of the risks in the two treatment groups, i.e. the relative risk (RR), defined as the ratio of probabilities or of relative frequencies: that of the event (developing a disease) occurring in the exposed group divided by that of the event occurring in the non-exposed group (Gordis 2009, Chap 11). Odds ratios (OR) are also a common way of expressing effect size, and provide a good estimate of RR when a disease is infrequent. Formally, OR = (odds that a control participant develops disease) / (odds that a placebo participant develops disease). Nothing about my framework requires the use of efficacy when representing RCTs, but efficacy gives us a simple way of expressing RCT results. I examine the situation for cases of early stop due to efficacy in section 3.3.1.

3.2.1.2. Harm

The term 'harm' is often ambiguous in practice. A harmful event is commonly taken to be either an unfavorable trend, such as when the data favor the greater efficacy of placebo, or an independent unfavorable event such as toxicity or other adverse effect. Practitioners in their publications are often not as careful as one would like to see. At times the term harm is left only partly specified; at other times, it is completely unspecified. I revisit the notion and its difficulty in more details in section 3.3.3 when I examine the general dilemma underlying early stop due to harm cases.
3.2.1.3. Futility

By futility, I mean either the lack of any evidence of efficacy or the inability to answer the study's primary question of interest. These will be further explained in section 3.3.2 when I examine the general dilemma underlying early stop due to futility scenarios.

I shall extend these brief analyses of efficacy, harm and futility to characterizations of the three corresponding types of stopping decisions shortly, in explaining the classification stage of the framework in section 3.3.

3.2.2. Identification

Once we have identified the type of early stopping decision in a concrete case, we are in a position to identify (and therefore to analyze) the corresponding dilemma that the DMC had to face. For each type of early stopping decision, there is a characteristic dilemma that accompanies it. This dilemma can be stated in a general form. Below, I list all three dilemmas in their respective general forms.

1. For cases of early stop due to efficacy, the general dilemma faced by the DMC is modeled as the following question: How much longer should the trial continue given that early benefit is observed?
2. For cases of early stop due to futility, the general dilemma faced by the DMC is modeled as the following question: *What is the likelihood for a change in data trend?*

3. For cases of early stop due to harm, the general dilemma faced by the DMC is modeled as the following question: *How much longer should the trial continue, so that evidence for harm sufficiently rules out the possibility of a spurious effect?*

Each of these general dilemmas will in turn refer to typical factors appealed to by DMCs. These typical factors are what I refer to as the *salient factors* (or *contextual factors*) of the case. Table 3.1 below displays the link between type of early stop and general dilemma, as well as a list of the typical factors corresponding to each of the three general dilemmas.

The typical factors are contextualized to the circumstances of the case. This contextualization, in turn, permits a link to what might be deemed decision criteria, represented in purple in Figure 3.1, connecting the identification step with the reconstruction stage. For instance, the degree of conservatism of the decision criterion should correlate positively with the availability of affordable, safe and effective alternatives to the new treatment under evaluation. The urgency of the situation should justify the stringency of the decision criterion. The more effective and safe the available treatments, the weaker the patients’ need for a new drug, the less urgent the public interest in alternatives, and consequently, the more conservative the decision criterion can afford to be.
By way of quick illustration: in the mid-1980's when HIV/AIDS was fatal, and well-tested, safe, and effective treatments were not available, less conservative criteria were justified. This contrasts with the situation of the second half of the mid 1990's when drug cocktails containing protease inhibitors were relatively effective and available. The situations before and after the availability of effective drugs for treating HIV/AIDS warrant different decision criteria—with different levels of conservatism.

With the dilemma and its contextual factors identified, we are in a position to move on to the next stage of the framework, namely, reconstruction. I return to the identification stage of the framework in section 3.3 (Classification and Identification).
<table>
<thead>
<tr>
<th>Type of early stop decision</th>
<th>Interim	extsuperscript{73}</th>
<th>Dilemma	extsuperscript{74}</th>
<th>Contextual factors</th>
<th>Desideratum</th>
<th>Conditions for early stopping</th>
</tr>
</thead>
<tbody>
<tr>
<td>For efficacy</td>
<td>Statistically significant</td>
<td>How much longer should the trial continue given that early benefit is observed?</td>
<td>Patient health status (healthy, unhealthy)</td>
<td>Weighing early favorable effect vs. uncertainty of late unfavorable effects</td>
<td>If unhealthy, stat. significance is (often) sufficient but not necessary. If healthy, stat. significance is (often) necessary but not sufficient.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Treatment alternative (available, unavailable)</td>
<td>Conservatism of decision criterion (greater availability, greater stringency)</td>
<td>If unavailable, stat. significance is (often) sufficient but not necessary. If available, stat. significance is (often) necessary but not sufficient.</td>
</tr>
<tr>
<td>For futility</td>
<td>Not statistically significant</td>
<td>What is the likelihood for a change in data trend?</td>
<td>Range of reasonable treatment size assumptions</td>
<td>Lower the conditional power, weaker the permissibility to continue.</td>
<td>If conditional power is dismally low (compared to original power), stop trial.</td>
</tr>
<tr>
<td>For harm</td>
<td>Not statistically significant</td>
<td>How much longer should the trial continue, so that evidence for harm sufficiently rules out the possibility of a spurious effect?</td>
<td>Type of trial (placebo-control, active-control)</td>
<td>Minimal clinical importance (mci)</td>
<td>If placebo-control, then if mci is unlikely, stop trial. If active-control, then if mci is unlikely, then consider treatment current usage.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Treatment current usage (widespread, not in use)</td>
<td>Secondary outcomes</td>
<td>If widespread use, then if no secondary benefits, stop trial.</td>
</tr>
</tbody>
</table>

Table 3.1 A summary of the main relevant factors, desiderata, and respective early stopping conditions for the different types of early stopping decisions.

	extsuperscript{73} This refers to the amount of statistical evidence \textit{revealed} for the primary outcome at the pertinent interim point in the trial.

	extsuperscript{74} Although there is a common underlying dilemma for all types of early stop cases—“\textit{how much longer should the trial continue?}”—each type of early stop decision has its own ‘follow up’ question.
3.2.3. Reconstruction

The reconstruction of the DMC decision includes several elements. It may include the actual stopping rule—if available—or some reasonable reconstruction of the actual stopping rule. In my framework, the reconstruction step is also meant to capture the reasoning that led to the selection of that (actual or reconstructed) rule. This requires the representation of a set of alternative decisions (e.g. continuing with the trial at 1st interim or stopping), a set of expected losses, and one or more alternative statistical monitoring plans. The statistical monitoring plan (see Figure 3.1—including the stopping rule, the number of interim analyses, and the efficacy in which the study was based, as reported in the study’s original publication (or as specified in its protocol).

I cover the task of reconstructing early stopping decisions in detail in section 3.4. For now it suffices to say that the proposed schema represents the ‘bare bones of the case.’ The representation is considered incomplete until the early stopping decision is justified as permissible. The standard for permissibility is the existence of at least one decision criterion on which the stopping decision is ‘optimal’, i.e., it maximizes expected utility. I hinted at the meaning of optimality earlier; I now say a little more about this idea.

According to the framework, a decision criterion prescribes a way in which, for a given representation of an early stopping decision, the criterion picks an “optimal stopping decision”. Because criteria are given by the type of stopping decision, dilemma, and salient factors of the case, the decision criteria is relative to a set of pre-defined criteria. We say
that an early stopping decision is permissible if it is ‘optimal’ according to at least one decision criterion. If we succeed in giving the RCT case a representation, and show that the stopping decision is ‘optimal’ on at least one decision criterion, we then say that the early stopping decision is permissible, or justifiable. Once again, the reconstructionist line of thinking is this: the decision analyst knows the early stopping decision that the DMC came up with given the RCT. The representation then shows—by reverse engineering—what epistemic and ethical values are consistent with the early stopping decision. If the representation seems indefensible, then the DMC’s early stopping decision is also indefensible. This approach to representation allows for evaluation criteria based on contextual factors. That proves to be crucial in meeting the normative requirements of the framework.

3.2.4. **Evaluation**

The last step of the framework is evaluation. Because evaluation is the most complex step, it requires a longer explanation than the other three. Evaluation is composed of two stages: a policy stage and a running stage. The first stage of evaluation reflects a policy decision which is logically prior to the running stage (or conduct) of the RCT. The first stage is concerned with the monitoring rule per se, i.e. with the justification of a choice of monitoring rule. The second stage is concerned with the justification of a particular action, i.e. decision, falling under its stopping rule. This second stage may include the ‘optimal’ decision criterion if one is identified during the reconstruction of the early stopping decision.
The need for a two-stage evaluation is supported by appeal to an intuitive understanding of what a just or fair assessment of decisions requires. An example serves to illustrate the intuition. We seem to give more weight to harm when it is engendered by a rule which is officially adopted by a DMC than when it is due to insufficiently enforced rules or insufficiently protective DMCs. For instance, burdens on trial participants due to an authorized restriction (e.g. a highly stringent and unrealistic stopping condition) seem morally more serious than burdens engendered by poorly monitored RCTs (e.g. ignoring unforeseen effects) or by DMC skeptics (e.g. skepticisms about the clinical importance of certain results), even if the harms on trial participants are exactly the same—or if the probability of harm faced by trial participants is the same. Unless our criteria of judging the permissibility of a case save this intuitive distinction, an evaluation of early stopping decisions is incomplete. Thus, we need to make a distinction between the choice of monitoring rule and the decision about whether to stop or continue the RCT.

To accommodate this intuitive distinction between the choice of a monitoring rule and the choice of a particular action falling under it, our evaluations of monitoring decisions must not simply compute or ‘tally up’ in some way the foreseeable effects each available action would have on trial participants. When judging permissibility, a more complete evaluation of the early stopping decision must weigh these effects differently for distinct types of causal link (and hence for distinct stopping rules) connecting the candidate decision to benefits and harms for trial participants. And this can be accomplished with the two-stage process outlined above.
Consider the claim: “The DMC stopped its trial early due to efficacy.” Until the reader knows the type of trial concerned and its governing rules—rules which defined the act of early stopping—she does not fully understand what the DMC did. The two-stage evaluation is necessary not only to account for the fact that exceptions to stopping rules are not uncommon in practice, but also because a breach of the stopping rule by the DMC, which may appear to maximize expected utility according to some representation, may not in fact have an adequate justification. The fact that a decision to continue a trial—increasing the chances of harm to trial participants—appears to be justified as maximizing expected utility given the choice of stopping rule, should not be considered an adequate justification—unless the rule under which the action falls is also appropriate to the given RCT context.\(^75\)

In this sense, we say that the first stage of evaluation is a policy decision, since it pertains to the prior choice of a statistical monitoring rule, and more specifically, the choice of the stopping rule. This stage of evaluation is covered in section 3.5 including its list of factors, and later exemplified in section 3.5.1.

Another important point about the need for a two-stage evaluation is in order. Even though there is a commitment always to represent the DMC decision as ‘optimal’ (permissible), the reconstruction of an early stopping decision, by itself, does not imply having a complete evaluation of the early stopping decision. The reason is this: in cases where there are rival opinions about the appropriate stopping rule, it may be

\(^75\) This distinction is motivated by Rawls’ “two concepts of rules”. (see Rawls 1955) The distinction is further motivated by an “intuitive understanding of justice” as seen in Pogge (2007) when critically examining Rawls’ *A Theory of Justice* (1971).
demonstrated by comparison that a rival stopping rule is better suited or more appropriate given the RCT context. This form of evaluation, however, is available only in cases where disagreements occur—see chapter 2 for disagreements over early stopping.

The second stage of evaluation is the running stage, which restricts attention to the particular representation of the case and its decision criterion, given the study’s stopping rule. The evaluation is that of the particular action falling under its stopping rule. This stage has two aspects to it. The first aspect aims at complementing the policy stage in the following way. It deals with the fact that a proper evaluation of early stopping decisions should embrace time-dependent factors as well, that is, factors that are relevant at the moment the interim assessment took place. Time-sensitive factors consider outside evidence (for or against the new intervention via some other RCT) that might have become available at the moment of the analysis. Another example is whether there is any significant change in the rate of the spread of the disease since the start of the trial. Yet another factor is whether the DMC realizes its assumptions about the background rate of the event of interest (e.g. virus transmission) were wide of the mark. These are factors whose values are difficult to account for at the policy stage, representing the fact that something new and relevant might emerge in the middle of the trial, since the values are chronologically contingent.
3.2.4.1. Comparing DMC decisions

By using the framework, readers are able to compare alternative DMC decisions. Why did the DMC choose to continue the trial rather than to stop it? Given the adoption of a particular statistical monitoring rule, had the foreseeable consequences of its actions been different, would the DMC have taken a different course of action? Under what realization of expected losses would the DMC have chosen a different course of action? These counterfactual “what-if” questions reveal a conscious attempt on the part of the reader (decision analyst or ‘justificatory’ agent) to provide reasons for the DMC decision. This point about counterfactual questions indicates a Woodward-style intuition about explanations, namely, that we have explanations if we can answer counterfactual questions. The suggestion is that what-if-things-had-been-different information is intimately connected to our judgments of explanatory relevance.

Without getting into details, the idea worth pointing out is that the analyst gains understanding of the DMC’s action in the ability to grasp a pattern of counterfactual dependence associated with relationships that are potentially exploitable for purposes of manipulation. For instance, we might come to understand how changes in patient health status will change the conservatism of early stopping rules: as health increases, greater efficacy is needed in order to outweigh possible long-term side effects. To grasp such relationships is the essence of what is required for the analyst to be able to explain the

76 “We see whether and how some factor or event is causally or explanatorily relevant to another when we see whether (and if so, how) changes in the former are associated with changes in the latter.” (Woodward 2003, 14)
DMC’s decision to continue the trial despite early evidence for efficacy. (See Figure 3.2 for details)

Similarly, this explanation can tell the analyst how, if she were to find an equivalent RCT in a context where no alternative treatments were available, the conservatism of early stopping rules would have been different. And she may also see, by way of contrast, that changing the type of statistical inference in the stopping rule (e.g. from a frequentist to a Bayesian rule) will make no difference to the DMC’s interim decision.77

By focusing on a reconstruction of the DMC’s decision, we attempt to give reasons—reasons that can turn out either for or against the DMC’s decision—that are now made public. If justification for an action consists in giving reasons for that action, then the framework should account for the fact that not every reason is a good reason. Good reasons should, first, help readers understand why the DMC chose its action instead of an alternative action. In order to do this, we should have a framework that enables readers to raise further questions about the DMC’s decision, by allowing comparisons among alternative DMC decisions.

77 Similarly to Woodward-style intuitions about manipulations, these manipulations do not require that humans be able to carry them out. Manipulation—intervention of the right sort—is hypothetical, and may not be within the technological abilities of humans.
3.3 Classification and identification

According to the U.S. FDA, the most fundamental role of the DMC is “to enhance the safety of RCT participants.” (GCTS, 9) According to its *Guidance for Clinical Trial Sponsors (Establishment and Operation of Clinical DMCs)*, the agency recommends that sponsors consider using a DMC when “the study endpoint is such that a highly favorable or unfavorable result, or even finding of futility, at an interim analysis might ethically require termination of the study before its planned completion”. (ibid 4) In other words, the DMC is expected to follow interim data analyses and to provide recommendations to investigators and sponsors on whether to stop or continue the trial. While the FDA document stresses safety, the DMC’s responsibility is better characterized as determining the ethical and scientific appropriateness of continuing the RCT.

A recommendation to stop early should be made if the DMC judges the interim results to be sufficient or compelling enough in one of the three cases of early evidence that we have identified: efficacy, harm or futility. As already discussed, each of these three types involves a different set of concerns and a different dilemma for the DMC. But before I proceed with a detailed model for each of these three types, I should say a bit more about the general ethical principles (and their source) guiding the selection of factors in the

---

78 This is the U.S. FDA *Guidance for Clinical Trial Sponsors (Establishment and Operation of Clinical DMCs)*. I shall leave aside for the time being other important responsibilities often expected and carried out by DMCs such as the need to preserve the trial integrity, data quality and credibility, as well as making available pertinent information about the trial to a wider clinical community in an expedient fashion.

79 The DMC’s responsibilities should include safeguarding the interests of trial participants by assessing the safety and efficacy of interventions, and keeping vigilance over the conduct of the trial. These responsibilities imply providing recommendations about stopping, modifying or continuing the trial on the basis of interim data analyses (Ellenberg, Fleming and DeMets, 2003).
qualitative framework. This will help us understand how DMC mandates relate to the selection of contextual factors when reconstructing early stopping decisions.

According to *International Conference on Harmonization* documents such as the ICH E6, general DMC mandates include the “public assurance that the rights, safety, and well-being of trial subjects are protected, consistent with the principles in the Declaration of Helsinki”. (ICH E6, 1) The DMC mandates are not only morally binding on DMC but they are binding on principal investigators too. The mandates are in addition to any local regulatory and legal demands where the study is conducted.

Fundamental ethical principles according to the Declaration of Helsinki include the respect for the individual trial participants, the right to make informed decisions regarding participation in the study—both initially and during the course of the trial—and the careful assessment of predictable risks and burdens on trial participants and communities affected by the condition under investigation. Together with actual case-studies, it is these principles which inform the DMC mandates that function as sources for the identification of contextual ethical factors in the decision framework. Thus, my primary objective is to ensure that these factors are represented in the framework.

Factors that are not included at the “identification step” in the analysis of early stopping decisions are factors that are not necessarily part of the general DMC mandate. But here an important disclaimer is in order. Because the qualitative framework is flexible

---

80 Article 9 of the Declaration of Helsinki.
81 Article 22 of the Declaration of Helsinki.
82 Article 18 of the Declaration of Helsinki.
(hopefully, one of its virtues), other factors not identified in DMC mandates can nevertheless be added in a reconstruction of an early stopping decision. Indeed, it is important to acknowledge that any reconstruction of any early stopping decision is, in principle, defeasible according to the proposed framework. Despite the fact that DMC mandates guide the selection of ethical factors for the identification and reconstruction of early stopping decisions, the mandates *per se* cannot guarantee that all critical relevant factors are included in the reconstruction of early stopping cases. Additional factors may derive, for example, from ethical norms shared by the wider community.

The rationale for using DMC mandates as the primary authority to guide the identification of ethically relevant factors can be explained by the need to make early stopping decisions public. I construe the ‘publicity’ standard in the following way. If the DMC mandates are intended to guide the DMC’s behavior, then the underlying principles guiding DMC decisions must be communicated so that they can serve as a public standard for evaluating early stopping decisions, and early stopping rules. Any early stopping decision (or monitoring rule) that is not ‘open’ to reconstruction automatically violates the publicity standard. In other words, early stopping principles should be *public* and in *accordance* with DMC mandates. In the next section, I give a detailed model for each of the three types of early stopping decisions that accord with DMC mandates.
3.3.1. Early stop due to efficacy

In this class of early stopping scenarios, the general dilemma faced by DMCs is how to balance the early evidence for efficacy against the uncertainty about long-term information concerning toxicity or side-effects. The difficulty that DMCs face in these early stop cases might be expressed by the following question: *How long should the trial continue after early benefit is observed?* For treatments like those for HIV/AIDS, which are presumably intended to be administered for months or years (indeed, in most cases for the remainder of the patient’s life), understanding the long-term effects, which include toxicity as well as long-run benefit, should clearly be significant. The problem, however, is that keeping trial participants in the control group off treatments shown to be beneficial—at least in the interim—has critical ethical and epistemic implications that DMCs have to struggle with, given that DMCs carry responsibilities towards securing the safety of trial participants.

From an ethical point of view, what should the allowed course of action be? Should the trial be stopped early and a beneficial treatment effect stated publicly? Or should the trial be continued to find out whether there is in fact a late-developing adverse effect? On the one hand, the first course of action would be considered inappropriate, and most likely very costly, if there is a late side-effect that clearly outweighs the beneficial effect observed early in the trial.\(^8\) On the other hand, the alternative course of action would be considered inappropriate, and most likely costly, if no late side effects turn out to exist, because then

---

\(^8\) Costly here means not only costly to patients, but also costly to public health care systems where applicable.
the superior treatment would be withheld from both trial participants and those patients outside the study that still await new treatment—possibly for what may turn out to be yet several years of further delay.\(^{84}\)

From an ‘epistemic’ point of view, if the DMC recommends the trial to continue after an early statistical monitoring boundary is crossed—when for instance a statistically significant result is observed early in the trial according to its statistical monitoring rule for the primary event of interest—then it becomes difficult to make sense of what purpose subsequent statistical boundaries will serve and whether there is any appropriate subsequent interim stopping point for the RCT. If the stopping rules of the study are disregarded because they are considered only as “guidelines,” then for the remainder of the RCT, there will be a suspicion about the real importance of such stopping rules. After all, what would the guidelines be for stopping the trial, given early evidence for efficacy? As suggested by Ellenberg, Fleming and DeMets (2003), perhaps, from that moment on, there will no longer be a reason to follow the early stop guideline for efficacy. Perhaps the most that could be said is that the decision to stop early would have to follow “sustained” evidence of efficacy or beneficial effect such as we might see if, for example, a pre-specified statistical boundary is crossed not only once but at two consecutive interim analyses.

\(^{84}\) I should also note that a second related issue that DMCs face, when considering early stop for efficacy, is whether or not early results are realistic, that is, whether or not the estimates of benefit might be too high to be true, or at least overly optimistic. Stuart Pocock—a leading U.K. biostatistician—has been a vocal critic of stopping trials early due to this reason. According to him (and others as well) trials that stop early for efficacy tend to be on a “random high” of observed benefit, and when further data on the primary outcome of interest is collected (by either the same or other RCTs), most often some “regression to the truth” to a more modest effect estimate occurs. It is primarily due to the problem of overly optimistic estimates that Pocock and colleagues recommend results “requiring proof beyond reasonable doubt that a treatment is sufficient to affect future clinical practice” before stopping a trial. (see Pocock 1992, 1999, 2005 for such views) My example of the evaluation stage of the framework in section 3.4 deals with an instance of this concern and possible problem for the monitoring and early stopping of RCT.
I should note, however, that one might be tempted to think that a way out of this dilemma is simply to allow for protocol change during the course of the study. That is, we can change the study in such a way that those patients in the control group are given the option to take advantage of the early benefit effects detected, if they so choose.\textsuperscript{85} By allowing patients to switch from control to the treatment group, the protocol change would permit trial participants to potentially benefit from an observed favorable early treatment effect.\textsuperscript{86} This change in the conduct of the study, however, will not change the horns of the dilemma that DMCs face. An early protocol change would give ‘untreated’ eligible patients—both in and outside the study—exposure to an immediate beneficial effect, but it might also subject them to a possible risk (harm) regarding late side effects and complications that might outweigh (or reverse) the early beneficial effect. Continuing the trial, on the other hand, without giving pertinent patients the option to switch treatments, although it might safeguard them against unnecessary exposure to possible toxicity and long-term adverse effects, deprives them of the choice of a potential immediate benefit. Essentially, the dilemma faced by the DMC does not change.

In order to cope better with the representation of the dilemma for efficacy, I suggest that the decision analyst might benefit from distinguishing between two sub-species of early stopping for efficacy, which I shall call E-A and E-B cases. E-A cases involve patients who face life-threatening stages of the disease,\textsuperscript{87} whereas E-B cases involve mostly healthy

\textsuperscript{85} This might be justified if one takes the principle of autonomy to be more important than other ethical principles.
\textsuperscript{86} It could also imply that such protocol change would be accompanied by a DMC’s recommendation to treat patients outside the study under similar conditions, to take advantage of a possible early beneficial effect, if they so choose.
\textsuperscript{87} They also include RCTs with high-risk patients with poor prognosis, such as those diagnosed with chronic heart failure conditions, where life expectancy is expected to be fewer than four years and less than one year for those in severe conditions.
patients or patients facing early stages of the disease. I make this distinction because different stages of the disease, and different health status or prognoses of trial participants, should warrant different stopping rules. In these two categories, the DMC must look for different ways to balance early evidence for efficacy against the uncertainty about long-term toxicity or side-effects. Let me expand on this distinction.

### 3.3.1.1. E-A cases

For patients experiencing a life-threatening illness such as advanced stages of HIV/AIDS, early evidence of efficacy may be sufficient to warrant stopping the trial even if it is unknown whether early benefits are sustained over the long-run. In this case, it makes good sense to imagine that early stopping should be considered permissible precisely in order to take advantage of the treatment’s short-term benefit. The mere possibility of a late treatment effect that is contrary to an early beneficial effect is not a sufficient reason to continue an RCT. E-A cases are represented by the 1\textsuperscript{st} quadrant (unhealthy, no alternatives) in Figure 3.2 below.

### 3.3.1.2. E-B cases

If the RCT involves mostly healthy patients, such as asymptomatic HIV-positive individuals or patients in early stages of AIDS, then long-term treatment effects can be considered of greater importance when balancing against early evidence of efficacy. In this
type of case, giving greater weight to long-term outcomes over apparent early benefit makes good sense, and the decision to continue the trial, despite early evidence for efficacy, can be more easily justified. E-B cases are represented by the 3rd and 4th quadrants in Figure 3.2 below.

According to the suggested distinction, then, a decision criterion for early stopping for efficacy that might be considered sufficient for the early stopping of an RCT involving patients in the advanced stages of a deadly disease may provide necessary, but not sufficient, conditions for a similar early stop decision in an RCT where trial participants are mostly healthy with good prognosis. This distinction is represented by the x-axis in Figure 3.2 below.

Another dimension that can strengthen differences between E-A cases and E-B cases (represented by the y axis in Figure 3.2), insofar as their evaluations for early stopping are concerned, is the availability of alternative treatments. This dilemma will be referred to as the “urgency” of the situation. The less effective and safe the available treatments, the stronger the patient’s need for a drug or treatment and the more urgent the public interest in alternatives; consequently, the less conservative the decision criterion to stop early for efficacy can afford to be. For instance, in the mid-1980’s when HIV/AIDS was considered fatal, and well-tested, safe, and effective treatments were not available, less conservative criteria for early benefit were justified. This situation can be contrasted with the circumstances in the second half of the mid-1990’s when drug cocktails containing protease inhibitors were relatively effective and available, warranting more conservative

88 More implications of the conservatism of the criteria for decision making are discussed in section 3.3.1.
decision criteria to stop early for efficacy. The situations before and after the availability of effective drugs for treating HIV/AIDS warrant different early stopping decision criteria.

With these distinctions in mind, we might say that when the potential losses of events occurring predominantly in the long-term are considered high compared to that of events occurring in the short-term, as happens in E-B cases, then the case for continuing the trial, despite early evidence of efficacy, can be made stronger than the case for continuing in E-A trials.

Of course, it must be acknowledged that the division into E-A and E-B cases is a simplification. In reality, both patient health and urgency occupy a spectrum, as exemplified by the RCT case involving the interim analyses of zidovudine in asymptomatic HIV-infected patients (ACTG 019).89 This case will be discussed in chapter 4.

89 There certainly exist RCT cases involving either mostly healthy trial participants with good prognosis but no alternative treatments, or patients with poor prognosis but for whom some alternatives exist. Such cases fall between the E-A and E-B spectrum (and this is clearly consistent with Figure 3.2).
Figure 3.2 Representation of the two weighing dimensions—and their corresponding factors—for cases of early stopping due to efficacy. The two dimensions are: (1) weighing the stringency (amount of certainty) of early evidence for efficacy, and (2) weighing the efficacy level for early stop.

In sum (see Table 3.2), the urgency of the situation warrants different early evidence for efficacy in the early stopping decision, and the health status of trial participants warrants different weights to efficacy levels when comparing early benefits vs. the uncertainty of long-term side effects. Both considerations assist us in representing and thus justifying early stopping decisions of RCTs. Combining all of these contextual
considerations (the conservatism of the decision criterion along with the two weighing mechanisms depicted in Figure 3.2. for comparing early favorable effect against the uncertainty of unfavorable long-term effects) will be important when meeting the framework's requirements.

3.3.2. Early stop due to futility

In futility cases, the DMC dilemma centers around the question of whether the current trial will provide sufficient evidence for investigators to answer the question the trial was originally designed to address. More specifically, in these cases, the DMC is mostly concerned with whether the interim data trend is low enough that there is very little chance of establishing a significant beneficial effect (vis-à-vis statistical significance) by the end of the trial.

For the purpose of answering this question, the statistical method of conditional power (CP) is frequently used by the DMC to assess whether an early trend (e.g. an unfavorable early trend) can still reverse itself, i.e. whether a statistically significant favorable trend can still emerge by the end of the trial. CP is often computed under the original assumptions of the trial, but at times, under varying assumptions of treatment effect. CP allows the DMC to assess how likely it is that the early trend might reverse itself, given the interim data, and whether the trial should be extended beyond its planned termination.
When considering whether to stop a trial due to futility, DMCs typically consider the probability of type II error: false negative error, i.e., failures to reject the hypothesis that there is no treatment difference, when in fact there is a difference. The concern is whether, at interim, they can detect the effect that they are hoping to see. An important related issue that can complicate matters arises from the employment of CP itself. The issue is that multiple applications of CP increase the chances of missing a real benefit effect by increasing the probability of false negative results. An early stopping decision that occurs after multiple applications of CP can eliminate the chances of an improvement that might yet be seen if the intervention is continued.

This has implications for a decision analyst who is trying to represent the DMC’s early stopping decision. It means, when assessing the permissibility of early stop due to futility, that cases that reflect multiple applications of CP methods should be considered with care, and specifically with an eye to controlling type II errors. In order to appreciate the difficulty involved in evaluating such cases, I remind the reader of two important trade-offs that DMCs face, particularly relevant for cases of stopping due to futility.

The first trade-off is the one inherent in experiments that involve sampling, namely the well-known trade-off between false positive error and false negative error. The more stringent the DMC’s requirement for detecting a certain efficacy—i.e. the smaller the chances of erroneously declaring the existence of an efficacy—the greater the chance of not detecting the efficacy when the efficacy is really there. The situation is analogous to the trade-off between “having an alarm without a fire” vs. “a fire without an alarm.” The trade-
off always exists with sampling. This much is well understood. But there are important restrictions by which things could go wide of the mark.

There is the important issue of whether the DMC might have been mistaken about the efficacy size. Their initial assumptions about how big of an effect they were going to see could have been misplaced. Perhaps they were too optimistic about the effect size when designing the RCT, which explains their inability to detect a smaller yet worthwhile clinically significant effect. (General rule: it is easier to detect big effects than small ones.)

In order to deal better with the representation of dilemmas for cases of futility, I once again suggest that the decision analyst can helpfully distinguish between two sub-species of cases about early stop due to futility, which I shall call F-A and F-B cases. F-A cases represent disagreements over whether efficacy is smaller than what the DMC has initially thought it was, and whether the current RCT can still detect it given the current trial size, whereas F-B cases represent disagreement about whether a smaller efficacy is clinically significant and worth detecting after all. I make this distinction because different disagreements over cases of futility call for different early stopping principles, since they permit DMCs to weigh differently the use of CP methods in deciding whether to stop or continue the trial. Let me expand on them.
3.3.2.1. F-A cases

In these cases, the disagreement is over the actual value of the parameter of interest, the efficacy size. Here, the decision analyst has ways of representing and evaluating disagreements by taking into account the interim data. How big of an effect was the DMC seeing at interim? With this information a DMC should be able to revise its opinions about efficacy and assess via CP whether there is still a chance of detecting the effect with the degree of stringency originally demanded in the trial. Bayesian CP methods are well suited for such task since they provide a natural way of updating priors about efficacy given interim data.

One example of a trial, introduced in section 2.3 (and to be examined in detail in chapter 4), illustrates a central issue motivating early stopping due to futility, with the focus of disagreement over the treatment effect size. The case study reflects the multi-center RCT conducted in 1994 evaluating an intervention that could prevent a brain-infection in HIV+ individuals, which is a good example of how CP can be important in assisting the DMC.

The study was designed with a target of 600 patients followed for a period of 2+½ years, with an estimate that “a 50% reduction of TE with pyrimethamine could be detected with a power of .80 at a two-sided significance level of .05 if 30% of patients given placebo developed TE during follow-up.” (Jacobson et al 1994, 385) Survival was the primary end point of the study. By March 1992, at the time of the fourth interim analysis, while patients were still being enrolled, the investigators found unexpectedly low toxoplasmic encephalitis
(TE) event rates, which compromised the original power of the study, and together with an early unfavorable trend in survival, the DMC recommended the early stopping of the trial.

The decision to early stop was not without disagreement. In a subsequent article authored by the same researchers, they explain that “a Haybitte-Peto interim analysis monitoring plan for early termination” was used by the DMC. (Neaton et al 2006, 321) Despite the DMC recommendation that the trial cease, “the chair advocated continuing the study due to the uncertainties about the future TE event rate and about the association of pyrimethamine with increased mortality” while “the DMC reaffirmed their recommendation to stop the trial.” (Neaton et al 2006, 324)

As we will see in chapter 4 when representing this trial, assuming that the chair and the DMC used the same statistical monitoring plan in arriving at their respective conflicting recommendations, including similar judgments about the consequences of continuing or stopping the trial (to be explained in section 3.4.1), the difference between their recommendations, can be seen as based on “the uncertainties about the future TE event rate”. Understanding the CP and a range of reasonable treatment effect sizes in proper context will prove to be important factors in identifying the early stopping principle (ratio decidendi) and meeting the framework’s requirements.
3.3.2.2. F-B cases

In these cases, the disagreement is over the appropriate size of the trial and the need for a power adjustment. Disputes can get a bit more complicated in this context. Disagreements prove to be the most challenging, since ethical and epistemic considerations are closely connected. That is because of the expected matching between clinical significance and statistical significance. Researchers often design their trials by first specifying a minimum clinical efficacy—i.e. minimum efficacy worth detecting and declaring. This is considered the DMC’s design efficacy: the efficacy on which the RCT’s sample size was originally based. This observation has an important implication for the decision analyst.

Because sample size constrains the evaluation of efficacy by changing power—as sample size increases, it becomes easier to detect a difference—90—the analyst should consider representing such cases as disagreements over whether the trial should be extended beyond its planned termination. By way of contrast with F-A cases, in F-B cases the focus of disagreement shifts from a dispute about the size of the effect to a dispute about the need to increase the trial size in order to detect small yet clinically important efficacy.

90 That is because the precision of the test statistic increases as sample size increases; which consequently increases power. By increasing sample size you decrease variability. By decreasing variability you increase the chances of detecting real differences. Since researchers do not have control over the actual value of efficacy, since it is a parameter and its value is unknown, researchers manipulate the sample size in order to control power.
Consider the following illustration. Suppose a study is designed with 900 healthy patients, providing adequate power (e.g., 80%) to detect 40% reduction in the transmission rate of HIV at the 0.05 significance level, under the assumption that the rate in the control group would be 15%. Suppose that at the first interim analysis, the transmission rates of the virus were way off target. They are observed as only 5% in the control group and 3.5% in the treatment group. The DMC sees that the rates are 1/3 of the expected transmission rates on which the study was based.

With such low transmission rates, the DMC wonders whether it has the ability to answer its original question, given the current rates and the size of the trial. Detecting such an advantage (i.e., 40% efficacy) would be very difficult given the background rates. Even if the DMC decreases the power of the study to 60%, and holds fixed the original significance level, the sample size would have to quadruple, assuming the transmission rates continue as they stand. The trial would have to increase to a size of 3800 patients.

Leaving aside the practical issues of whether this increase in accrual might be feasible—e.g., too difficult to find such number of participants, time or financial constraints—the DMC may have to wrestle with the question of whether detecting 1.5% difference in transmission rates is of clinical importance. Even though it might be seen by many as good public health policy to lower infection rates—after all the 1.5% difference in transmission rates still shows that those treated were 30% less likely to pass on infection than those who were not treated—according to others, effect size may be of little or no clinical importance. That is because such efficacy would not outweigh potential (yet unknown) long term effects due to such treatment. For doctors it might be too difficult to
convince healthy individuals, e.g. as soon as they test HIV+, to go on such treatment. Once on treatment, they are on it for life.

In sum, the decision to stop or continue in cases of futility is usually well represented either as a disagreement over a range of reasonable assumptions about efficacy values or as a disagreement over whether there should be an increase in trial size. Regardless of its representation, the DMC dilemma hinges on whether initial assumptions about the treatment effects were realistic and how they need to be revised in midcourse.

When evaluating F-A and F-B cases, the rationale for evaluating stopping rules is the following: if CP is dismally low compared to the unconditional power assumed in the study protocol, for a range of reasonable treatment effect sizes, then the DMC should have no grounds to continue the RCT since the study is unlikely to demonstrate the efficacy of the new treatment. What is deemed dismally low—the power threshold—will vary according to contextual factors illustrated in Figure 3.2 when exemplifying cases of early stop due to efficacy.

Before moving to cases of harm I should note that there is another dilemma, for cases involving futility. It is the problem of a late non-stopping decision rather than an early stopping decision. I will not chase this flip side of the coin here, but simply note that prima facie there seems to be nothing in principle about framing such cases that the framework would not be able to represent.91

91 Another consideration for futility cases is circumstances in which demonstrating the equivalence of the two treatments might nonetheless be clinically important. For instance, two treatments used in practice where one is a
3.3.3. Early stop due to harm

Of all types of early stop scenarios, the most complex cases are those that involve a decision to stop due to harm. There are a number of reasons for this complexity.

The first of these is simply the lack of clarity and specificity in what investigators and DMCs mean by “harm” in their respective trials. Different types of harm make different stopping and monitoring rules appropriate. While the notion of harm should allow for some flexibility, and it is not feasible to plan for every contingency, some prior thinking and clarity on what ‘harm’ means in the particular context of the RCT are helpful to everyone involved and interested in the particular study.

Second, and closely related to this initial point, is the prevalence of ad hoc responses to harm: “while many trials have monitoring rules for the primary efficacy outcomes, often there is reliance on ad hoc judgments about adverse risks rather than having formal monitoring rules.” (Redmond et al 2006, 133) When the situation involves unanticipated adverse events, it is true that statistical methods are less helpful. It is simply unfeasible to specify an a priori statistical monitoring plan that could control adequately for all possible factors warranting early stop due to safety.
But let us set aside these two admittedly intractable problems—lack of clarity and unforeseeability—and concentrate, in the remainder of this section, on (partially) foreseeable types of harm for which reliance on *ad hoc* judgment is not appropriate. As noted earlier, one such type of harm involves the appearance of an opposite effect in the primary event of interest—i.e., a “negative” trend where the direction of the effect is opposite to the one expected. Another partially foreseeable type of harm consists in a higher than predicted number of adverse effects. In both of these cases, statistical methods can be helpful—although still limited—and reliance on *ad hoc* judgment is much less clearly justified. Statistical methods used for monitoring efficacy, once they are properly adjusted, can be helpful in early stop for harm cases.

In particular: because an early negative trend is not that uncommon, statistical monitoring plans should take this possibility into account. As recognized by several experts, the statistical monitoring boundaries and the DMC’s attitudes to trial termination do not need to be symmetrical. DMCs generally do not—and should not—require the same level of evidence to stop an RCT for harm as they do for benefit, despite the fact that “boundaries still need to be sufficiently conservative to control the type I error in either direction” (DeMets et al 1999, 1987). The concern here is that if it is too easy for a study to claim harm, “potentially promising drugs with other useful attributes might be dismissed too readily.” (ibid)

The class of early stop cases for harm shares notable similarities with cases of early stop for efficacy. In efficacy cases, the fundamental question is whether the trial should
continue after early benefit is observed. The dilemma in harm cases, which must also grapple with the issue of how much longer to continue, is best represented by the question: *Is the evidence for harm sufficient to rule out the possibility that the early effect is not spurious?* Because the evidence demanded for demonstrating harm is less than that demanded for demonstrating efficacy, the focus of the dilemma shifts from being primarily an ‘ethical’ issue to an epistemic issue.

Let me clarify the distinction. In early stop cases for efficacy, an epistemic question about evidence becomes an ethical issue once there is sufficient (statistically significant) evidence for the primary effect of interest (i.e., efficacy). In early stop for harm scenarios, by contrast, an originally ethical problem about harm to trial participants turns into an epistemic issue, i.e. whether the suggestive but inconclusive (not yet statistically significant) evidence for harm is sufficient to rule it out as due to chance.92

We can make some headway in evaluating cases of early stop due to harm by making a distinction between *placebo-control* RCTs and *treatment-control* RCTs. This distinction, to some extent, corresponds to the distinction about the availability of alternative treatments (the ‘urgency’ dimension of Figure 3.2) suggested in our discussion of early stop cases for efficacy, and it plays a similar role when evaluating harm cases.

The distinction is helpful in the following way. If an early negative (but not statistically significant) trend is observed in a placebo-control RCT, then the criterion for

---

92 For the time being, while discussing cases of early stop for harm, and the permissibility rationale for their stopping, I assume that early evidence for efficacy is not statistically significant in such cases—otherwise the case would qualify as a case of early stop for efficacy instead of harm.
whether to stop due to harm should be contingent on whether the trial still has a chance of demonstrating a minimal yet beneficial effect that is of clinical importance. If there is a reasonable chance that the remaining trial can do so, then continuing is permissible. If however the chances are low, i.e. the demonstration of a minimal benefit effect of clinical importance is ruled out by the interim results, then continuing is not permissible. The question of whether or not the trial might still have a chance of demonstrating a minimal yet beneficial effect that is of clinical importance might be addressed (as in efficacy cases) by employing conditional power methods on what might be considered reasonable effect sizes of clinical importance.

The permissibility rationale in such cases of early stop for harm is that, if a new treatment is not showing to be and not promising to be any better than placebo, then not only it is unlikely to be of much clinical interest, but trial participants are better off by going on to other potentially promising RCTs, if those exist. When the treatment is no longer considered clinically important and trial participants are better off leaving, then early termination is justified.

If, instead, the situation pertains to a standard-control RCT—where there is an alternative treatment instead of placebo in the control group—then the permissibility of continuing the trial given early evidence of harm should be addressed differently. We should consider whether the new treatment is already in use for some other purpose, together with interim secondary outcomes, using the following criterion: assuming the evidence for harm is not statistically significant, and assuming that the demonstration of a minimal benefit effect of clinical importance is deemed unlikely by the interim results, the
trial can be allowed to continue despite the possibility of a truly harmful effect being present, if secondary benefits are observed during interim analyses. Let me expand.

The permissibility of continuing such trials would be justified for circumstances where it is useful for the DMC to rule out cases of truly harmful effect vs. neutral effects, in the following way. Assuming the treatment under evaluation is already widespread, and assuming that secondary outcomes during interim look promising in that trial, it might be significant to identify what sort of treatment the new treatment really is. That is, it might be helpful to distinguish between a new treatment that has no efficacy at all (e.g. no beneficial effect on mortality) but yet has some improvement on secondary outcomes (e.g. quality of life, disease progression, clinical deterioration), and a new treatment that increases mortality (true harmful effect) yet has some improvement on an important secondary outcome.

For treatments like HIV/AIDS ones, it is not uncommon that a newly developed drug may become commonly used before it has been evaluated in a large and ‘definitive’ phase-III RCT study. This was the case during the early years of AZT, which became widely available off-label and was used in patients with HIV+ that were not under nearly so severe health conditions as those in terminal stages of AIDS.
3.4 Reconstruction

This section discusses the reconstruction of an early stopping decision, at the moment of the interim decision. Section 3.5 discusses the earlier decision about selecting a monitoring rule.

To reconstruct a stopping decision, we must begin by listing the possible actions available to the DMC together with the range of possible true states of nature. Both of these components of the model are relatively clear. In the simplest case where we have just one interim analysis, the four possible actions (represented as \( a_1 \), \( a_2 \), \( a_1.f \) and \( a_2.f \) in Table 3.2 below) combine exactly two binary choices: a decision to stop at interim or to continue, and a decision to choose the standard treatment or the new treatment. The possible states are conveniently given in terms of the hypotheses used in the RCT statistical test, e.g., as differences between recovery rates. In simplest terms, suppose the set of hypotheses under consideration is \( \{ H_0, H_1 \} \). \( H_0 \) is the null-hypothesis—the hypothesis that the new treatment has no effect of interest, whereas \( H_1 \) is the alternative hypothesis in the study—the hypothesis that the new treatment has a certain efficacy. The relevant possible states are then just \( H_0 \) and \( H_1 \) (referred to as \( s_1 \) and \( s_2 \) in Table 3.2 below).

The basic idea in my decision model is that the DMC observes interim data and, guided by its decision monitoring rule (the function mapping observed data into the set of possible actions), decides which action to take. As we know, however, interim data and the monitoring rule are not sufficient for the proper ‘reconstruction’ of DMC decisions. Many DMC decisions (whether to stop or continue) often violate the prescription of the
monitoring rule. That is because statistical monitoring rules generally function only as guidelines for the DMC. There are numerous cases in which the DMC acted in violation of their stopping rule when, for instance, deciding to continue the trial despite early evidence for efficacy (e.g. despite evidence deemed statistically significant according to its monitoring rule). At other times, the DMC decided to stop when there was some evidence for efficacy, even though the evidence was not deemed statistically significant according to its monitoring rule.

One lesson from such cases is that an early stopping rule (or principle) is not the same as the design monitoring rule. There is a difference between the a priori stopping rule (the design stopping rule adopted by the DMC) and the possibly revised stopping rule that is adopted as new information emerges in the trial.

Still, our reconstruction of early stopping decisions begins with the basic tools from decision theory. Accordingly, to each pair \((a_i, s_j)\), consisting of an action and a state of nature, there is an outcome given by \(o_{ij}\) (see Table 3.2). The decision matrix of Table 3.2 describes in abstract form early stopping decisions (and their outcomes).

<table>
<thead>
<tr>
<th>Action</th>
<th>(s_1 (H_0))</th>
<th>(s_2 (H_1))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>@interim</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a_1): choose new treatment</td>
<td>(o_{11})</td>
<td>(o_{12})</td>
</tr>
<tr>
<td>(a_2): choose standard treatment</td>
<td>(o_{21})</td>
<td>(o_{22})</td>
</tr>
<tr>
<td><strong>@final</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a_1.f): choose new treatment</td>
<td>(o_{41})</td>
<td>(o_{42})</td>
</tr>
<tr>
<td>(a_2.f): choose standard treatment</td>
<td>(o_{51})</td>
<td>(o_{52})</td>
</tr>
</tbody>
</table>

**Table 3.2** RCT monitoring decisions (in abstract).
Each possible outcome, however, is associated with two essential elements that should be represented for each act-state pair \((a_i, s_j)\): a utility or loss (i.e. ‘payoff’) and a relevant probability to be explained below (see Table 3.3). Our task is the following: given a set of options for the DMC, attempt to represent the DMC’s actual decision as ‘optimal’ in some sense of optimality, that is, under some decision criterion. Section 3.4.2 will offer an example of decision criterion.

<table>
<thead>
<tr>
<th>State of Nature</th>
<th>Action</th>
<th>Verdict</th>
<th>Description</th>
<th>Outcome</th>
<th>Payoff</th>
<th>Relevant Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s_1 [H_0])</td>
<td>(a_1)</td>
<td>Stop</td>
<td>Accept (H_1)</td>
<td>(o_{11})</td>
<td>(L(a_1, s_1))</td>
<td>(p_0 = p_\delta(a_1, s_1))</td>
</tr>
<tr>
<td></td>
<td>(a_2)</td>
<td>Stop</td>
<td>Accept (H_0)</td>
<td>(o_{21})</td>
<td>(L(a_2, s_1))</td>
<td>(p_0' = p_\delta(a_2, s_1))</td>
</tr>
<tr>
<td>(s_2 [H_1])</td>
<td>(a_1)</td>
<td>Stop</td>
<td>Accept (H_1)</td>
<td>(o_{12})</td>
<td>(L(a_1, s_2))</td>
<td>(p_1 = p_\delta(a_1, s_2))</td>
</tr>
<tr>
<td></td>
<td>(a_2)</td>
<td>Stop</td>
<td>Accept (H_0)</td>
<td>(o_{22})</td>
<td>(L(a_2, s_2))</td>
<td>(p_1' = p_\delta(a_2, s_2))</td>
</tr>
</tbody>
</table>

Table 3.3 Representation of decision outcomes and payoffs at interim point.

For our purposes, the ‘payoff’ of each potential action, is represented by a loss (or utility) function \(L(a_i, s_j)\). This function maps the set of all possible combinations of actions and true states of nature into a particular utility scale. The value \(L(a_i, s_j)\) is supposed to be a measure of the foreseeable benefit (or loss) of taking action \(a_i\) when a state \(s_j\) obtains. Here, the value \(L(a_i, s_j)\) represents disutility, so having a high value is bad.

---

93 These are probabilities engendered from a particular decision stopping rule (e.g. the stopping rule adopted in the RCT). Given a particular decision stopping rule (\(\delta\)), the probability of reaching decision \(a_i\) when \(s_j\) is true has a specific value, and is represented for our purposes as \(p_\delta(a_i, s_j)\). On the one hand, when dealing with a Bayesian stopping rule, this probability \(p_\delta(a_i, s_j)\) is multiplied by a prior probability which is attached to the particular state of nature \(s_j\). This product of probabilities generates the relevant probability referred to in the table. On the other hand, if we are dealing with a frequentist stopping rule, there are no probabilities attached to states of nature, in which case the relevant probability reduces to \(p_\delta(a_i, s_j)\)—which might be better represented as \(p_\delta(a_i | s_j)\) as opposed to \(p_\delta(a_i, s_j)\).

94 New treatment is not superior to standard treatment. E.g. there is no clinically significant difference between the treatments.

95 New treatment is superior (e.g. more effective, safer) to standard treatment.
3.4.1. Loss function

To derive a loss function for the purposes of representing an early stop decision, I suggest beginning with the following simple assumptions, which are derived from Heitjan, Houts and Harvey (1992):

A1. We assume that the sample size of the RCT is relatively small compared to the number of potential patients (patient horizon) that might benefit from the treatment chosen at the conclusion of the RCT.

A2. We assume that, there is a fraction of patients from the patient horizon, beyond those participating in the RCT that are going to be treated with a standard treatment during each interim,\(^{96}\) i.e. during each monitoring portion of the RCT. Further, this fraction is an increasing function of time: the longer the trial takes to complete, the greater the number of patients from the patient horizon that are treated with the standard treatment.

\(^{96}\) When no standard treatment is available, and treatment comparison is made against placebo, the assumption is that the longer the trial takes to complete, greater is the number of patients from the patient horizon that are left untreated.
A3. We assume that at the end of the trial, all patients will receive the treatment that is chosen as the superior of the two, unless no treatment is declared superior, in which case, patients will receive the standard treatment.

These are the basic and minimum assumptions for considering a potential loss function $L(a_i, s_j)$. Other assumptions will be added to these depending on the assessment of the foreseeable consequences of making mistakes when deciding (choosing) between the two treatments. Let me now suggest two types of loss functions that might serve as guiding-examples for our reasoning: the ‘ethical’ loss function, and the ‘scientific’ loss function.

For the ‘ethical’ loss function, one possibility is to base the loss at $(a_i, s_j)$ on the expected proportion of ‘responses’ in the patient horizon (e.g. positive events such as recoveries, of which the primary end point of the study is based) when the DMC takes action $a_i$ and the true state of nature is $s_j$. Thus, for instance, if we are modeling a case of early stop for efficacy, we might reason: the best of outcomes is when the superior treatment (e.g. new treatment) is correctly chosen as the superior treatment early in the trial (e.g. at first interim), followed by the second best outcome when the same conclusion is made but at the next interim (which could be at the end of trial, if we model the trial with a single interim analysis halfway, for instance). The worst decision is to conclude that the new treatment is superior early in the trial when in fact the new treatment is not; the second worst outcome is when a similar conclusion is made but at the next interim. By applying similar reasoning and filling out the remaining possibilities, we may obtain a scale. Just to complete the example here, the third best outcome could be correctly
choosing the standard treatment early in the trial, when it is in fact superior to the new treatment. Table 3.4 exemplifies this suggestion.

When ranking outcomes, the decision analyst should contrast both the pair of possible mistakes, ‘false-positive’ RCT conclusions against ‘false-negative’ RCT conclusions, and the pair where the correct conclusion is drawn. Given the type of early stop decision at hand and its corresponding list of relevant factors (see Table 3.1), the analyst considers the seriousness of the foreseeable consequences for each decision. Contextual factors play a role in assessing the importance of each possible mistake (or correct conclusion). For instance, if the RCT case is an example of an early stop due to efficacy scenario, with healthy trial participants and treatment alternatives available, the analyst reasons whether the loss associated with a false-positive decision \( (L(a_{1,f}, s_1) \) or \( L(a_1, s_1) ) \) might reflect exposing patients to unnecessary toxicity, and whether the loss for a false-negative decision \( (L(a_{2,f}, s_2) \) or \( L(a_2, s_2) ) \) reflects discouraging the pursuit of what might be considered only a marginally superior treatment.

How do the two decisions compare? If, due to toxicity of the new treatment, the loss expected from a false-positive is considered to be more serious than the loss expected from a false-negative, given that there are a number of other ongoing promising trials, then \( (L(a_{1,f}, s_1) \) or \( L(a_1, s_1) ) > (L(a_{2,f}, s_2) \) or \( L(a_2, s_2) ) \). If however, the case pertains to patients that are mostly unhealthy, then a false-positive reflects exposing patients to no greater efficacy while a false-negative reflects discouraging pursuit of a truly promising (yet rare) treatment. In this case, \( (L(a_{1,f}, s_1) \) or \( L(a_1, s_1) ) < (L(a_{2,f}, s_2) \) or \( L(a_2, s_2) ) \).
Let us turn now to the idea of a ‘scientific’ loss function. One possibility is to base the loss at \((a_i, s_j)\) on the principle that the main purpose of the RCT is to find out whether or not the new treatment is superior to the standard treatment regardless of the consequences. Correctly concluding that the new treatment is more effective than standard would have the same utility (same value) as correctly concluding that the new treatment is less effective. With this loss function, for instance, the maximum utility might be realized when the DMC makes the correct decision early during the trial, which includes choosing the superior treatment at an early interim when the drugs are clinically different, or not declaring the superiority of either treatment when the drugs are deemed clinically equivalent (no superiority between them). Similarly, making an incorrect choice early on the trial, when the drugs are not clinically equivalent, will have the minimum utility (or maximal loss). Making either of these choices (correct or incorrect) at the final analysis of the trial would lessen its maximal utility (if correct), or lessen its loss (if incorrect). For the decisions at final analysis, utility values could be represented by multiplying the maximal and minimal losses by a factor based on the number of patients in the horizon, i.e., the estimated number of patients that might benefit from the new treatment. Table 3.5 exemplifies this suggestion.
Table 3.4 Example of the ‘ethical’ loss function.

<table>
<thead>
<tr>
<th>Action</th>
<th>State of Nature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$s_1 (H_0)$</td>
</tr>
<tr>
<td>@interim</td>
<td></td>
</tr>
<tr>
<td>$a_1$: choose new treatment</td>
<td>$L(a_1, s_1) = 1^{st}$ worst</td>
</tr>
<tr>
<td>$a_2$: choose standard treatment</td>
<td>$L(a_2, s_1) = 3^{rd}$ best</td>
</tr>
<tr>
<td>@final</td>
<td></td>
</tr>
<tr>
<td>$a_1, f$: choose new treatment</td>
<td>$L(a_{1, f}, s_1) = 2^{nd}$ worst</td>
</tr>
<tr>
<td>$a_2, f$: choose standard treatment</td>
<td>$L(a_{2, f}, s_1) = 4^{th}$ best</td>
</tr>
</tbody>
</table>

Summary: $L(a_1, s_2) < L(a_{1, f}, s_2) < L(a_2, s_1) < L(a_{2, f}, s_1) < L(a_2, s_2) < L(a_{1, f}, s_1) < L(a_1, s_1)$

Table 3.5 Example of the ‘scientific’ loss function.

<table>
<thead>
<tr>
<th>Action</th>
<th>State of Nature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$s_1 (H_0)$</td>
</tr>
<tr>
<td>@interim</td>
<td></td>
</tr>
<tr>
<td>$a_1$: choose new treatment</td>
<td>$L(a_1, s_1) = \text{worst}$</td>
</tr>
<tr>
<td>$a_2$: choose standard treatment</td>
<td>$L(a_2, s_1) = \text{best}$</td>
</tr>
<tr>
<td>@final</td>
<td></td>
</tr>
<tr>
<td>$a_1, f$: choose new treatment</td>
<td>$L(a_{1, f}, s_1) = 2^{nd}$ worst</td>
</tr>
<tr>
<td>$a_2, f$: choose standard treatment</td>
<td>$L(a_{2, f}, s_1) = 2^{nd}$ best</td>
</tr>
</tbody>
</table>

Summary: $L(a_1, s_2) = L(a_2, s_1) < L(a_{1, f}, s_2) = L(a_{2, f}, s_1) < L(a_1, s_1) = L(a_2, s_1) < L(a_2, s_2) = L(a_{1, f}, s_2) < L(a_1, s_1) = L(a_2, s_2)$

At this point, we must recognize that interim data, monitoring rules, and the foreseeable consequences for each action by themselves are not sufficient for a plausible reconstruction of early stopping decisions. Thus, in addition to the possible ‘payoffs’ (relative losses), we need two other elements in order to assess the permissibility of early stopping decisions. We need to link the expected losses to contextual factors relevant to the dilemma the DMC face, in such way that the links can be manipulated counterfactually, and
we also need a decision criterion. In the next section I introduce one decision criterion that can assist the decision analyst when reconstructing her case. The decision criterion is: the maximization of expected utility (MEU).

3.4.2. Decision criteria

In this section I discuss one candidate decision criterion for early stopping decisions. A decision criterion is a principle that takes a given representation of an early stopping decision and picks a stopping decision which is ‘optimal’. A necessary condition for the permissibility of an early stopping decision is to have a decision criterion that (a) makes the stopping rule ‘optimal,’ and (b) is itself justifiable. The most natural candidate is the maximization of expected utility (MEU) criterion: choose the action that maximizes expected utility (or minimizes expected loss). This is commonly considered a natural and rational criterion for decision making.

As we have already noted, in order to properly assess the permissibility of an action, the decision analyst requires a two-stage evaluation, namely, the policy level and the running level. Simple maximization of an action will not count as a full justification of an early stopping decision. Nevertheless, in this section, we continue to focus on the “running” stage and ignore the policy component.
3.4.2.1. Maximum expected utility (MEU)

According to MEU, the decision analyst chooses the available action with maximal expected utility. Given the utilities of the outcomes and probabilities of the states of nature, the decision analyst computes the expected utility of the various alternatives, as follows:

\[
EU(a_i) = P(s_1)u(o_{i1}) + P(s_2)u(o_{i2}) + \ldots + P(s_m)u(o_{im})
\]

Note: \(u(o_{ij}) = -L(a_i, s_j)\).

The decision analyst then picks the action (or an action) with maximal EU (or minimal expected loss).

The rationale for considering this decision criterion as a first candidate is clear: in the long run, where the DMC faces similar RCTs with similar choices repeatedly, no other criterion can be expected to do better than MEU. For contexts in which multiple repetitions of similar RCTs are unlikely, the decision criterion is less relevant and thus more problematic. For a unique and rare RCT, which involves a single early stopping decision, we would be hard pressed to justify the use of MEU.

Leaving aside for the moment the problem of justifying the decision criterion, let us assume that for a given stopping rule \(r\), the probability of reaching decision \(a_i\) when \(s_j\) is

---

97 The probabilities referred here are unconditional probabilities, because states are independent of action.
98 Note: MEU in the single case can be justified via a representation theorem, but that is only if the utilities have no independent meaning, and we introduced solely to represent preferences.
true has a specific value \( p_i(a_i \mid s_j) \). Suppose that the utility of action \( a_i \) when \( s_j \) is true is \( u(a_i, s_j) \). The utility is obtained by averaging over all possible outcomes.

And the expected utility of stopping rule \( r \) when \( s_j \) is true is:

\[
EU(r \mid s_j) = \sum_i p_i(a_i \mid s_j)u(a_i \mid s_j)
\]

If the prior probability distribution on the values of \( s_j \) can be given by some means, represented by \( f(s_j) \), as is the case for a Bayesian stopping rule, then the \textit{a priori} expected utility of stopping rule \( r \) is:

\[
PEU(r) = \sum_j f(s_j)EU(r \mid s_j)
\]

Notice that for cases requiring \( f(s_j) \), if the stopping rule is a frequentist stopping rule, an adjustment is necessary because the notion of a probability attached to a state of nature is inapplicable. This adjustment can be made by using the principle of insufficient reason (or indifference). The decision analyst assigns equal probability value to each state of nature. For instance, if there are three possible states of nature and the decision analyst has no more reason to believe any one of them to be the true state than any other, then she assigns each state an equal probability value—i.e. \( 1/3 \). This is a reasonable approach, since it gives the decision analyst a ‘reinterpretation’ of a frequentist stopping decision under what is essentially a Bayesian decision criterion.
3.4.3. Loss function, factors and decision criteria: a paradigm case

Using the two weighing mechanisms and contextual factors identified for understanding early stop due to efficacy cases, our paradigm case (see Figure 3.2), the relationship between losses incurred by a type I error and a type II error are represented below by quadrant (relevant dimension). Accordingly, the importance of efficacy depends not only on the health status of trial participants, but also on the availability of alternative treatments—the weighting principles are expressed within each quadrant.
Table 3.6 The contextualization of a loss function to the case of early stop due to efficacy scenarios. Different quadrants represent different permissibility of early stopping rules.

<table>
<thead>
<tr>
<th>Quadrant 2 (unhealthy, several alternatives)</th>
<th>Quadrant 3 (healthy, several alternatives)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ L(\text{False } +) &gt; L(\text{False } -) ]</td>
<td>[ L(\text{False } +) \gg L(\text{False } -) ]</td>
</tr>
</tbody>
</table>

Threshold for harm with unhealthy participants is **higher than** with healthy participants. Level of efficacy required to offset possible late unfavorable effects should be relatively low.

Strength of evidence for efficacy with acceptable alternative treatments is **higher than** with no alternative treatments.

**Upshot:** loss incurred from a false negative (missed opportunity to claim a real yet small efficacy) is **outweighed** by the potential loss incurred from a false positive (declaring incorrectly efficacy where several alternatives already exist and late side effects have limited importance).

<table>
<thead>
<tr>
<th>Quadrant 4 (healthy, no alternatives)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ L(\text{False } +) &gt; L(\text{False } -) ]</td>
</tr>
</tbody>
</table>

Threshold for harm with healthy participants is **lower than** with unhealthy participants. Level of efficacy required to offset possible late unfavorable effects must be relatively high.

Strength of evidence for efficacy with acceptable alternative treatments is **higher than** with no alternative treatments.

**Upshot:** loss incurred from a false negative (missed opportunity to claim a real yet small efficacy) should not outweigh the potential loss incurred from a false positive (declaring incorrectly efficacy where several alternatives exist and late side effects have great importance).

<table>
<thead>
<tr>
<th>Quadrant 1 (unhealthy, no alternatives)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ L(\text{False } +) \ll L(\text{False } -) ]</td>
</tr>
</tbody>
</table>

Threshold for harm with unhealthy participants is **higher than** with healthy participants. Level of efficacy required to offset the possibility of late unfavorable effects should be smaller than with mostly healthy patients.

**Upshot:** loss incurred from a false negative (missing an opportunity to claim real efficacy) **outweighs** the potential loss incurred from a false + (wrongly declaring efficacy in a context where no alternative exists and the uncertainty of late side effects is of low importance).

<table>
<thead>
<tr>
<th>Quadrant 3 (healthy, several alternatives)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ L(\text{False } +) \gg L(\text{False } -) ]</td>
</tr>
</tbody>
</table>

Threshold for harm with healthy participants is **lower than** with unhealthy participants. Level of efficacy required to offset possible late unfavorable effects must be relatively high.

Strength of evidence for efficacy with acceptable alternative treatments is **higher than** with no alternative treatments.

**Upshot:** loss incurred from a false negative (missed opportunity to claim a real yet small efficacy) should not outweigh the potential loss incurred from a false positive (declaring incorrectly efficacy where several alternatives exist and late side effects have great importance).

<table>
<thead>
<tr>
<th>Quadrant 4 (healthy, no alternatives)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ L(\text{False } +) &gt; L(\text{False } -) ]</td>
</tr>
</tbody>
</table>

Threshold for harm with healthy participants is **lower than** with unhealthy participants. Level of efficacy required to offset the possibility of late unfavorable effects must be substantial, not small—compared to a trial with mostly unhealthy patients.

**Upshot:** If efficacy level observed early in the trial is small, then the loss incurred from a false negative (missing an opportunity to claim a real yet small efficacy) is **outweighed** by the potential loss incurred from a false positive (wrongly declaring the existence of an efficacy which is insufficient to offset possible late side effects).
By linking contextual factors to a loss function, the decision analyst is now in a position to assess whether the stopping rule is ‘optimal.’ In the next section, evaluation stage, I explain the weak sense of optimality used for evaluating stopping rules. I also explain the two-stages of evaluation: the policy stage and the running stage.

### 3.5 Evaluation

The evaluation of stopping decisions is governed by the following notion of comparative (pairwise) optimality: as long as the stopping decision does no worse than its alternative along its relevant dimension or dimensions (as specified in the relevant quadrant from Table 3.6), it counts as ‘optimal’. To say that one decision does no worse than another is to say that it has equal or greater expected utility. The ‘relevant dimension’ is the key parameter in the expected utility calculation.

In short: if the decision analyst succeeds in giving the RCT case a reconstruction, and shows that the stopping decision does no worse than its rivals along the relevant dimension, then the DMC's early stopping decision was permissible (i.e., justifiable).

---

99 Alternative decision principles, such as maximin, might be used in place of expected utility maximization (MEU). The framework could accommodate such alternatives. For simplicity, I restrict my attention to MEU.
In the next section I explain the policy stage which pertains to the prior choice of a statistical stopping rule, i.e. the choice of stopping rule by the DMC for the purposes of monitoring its RCT (the design stopping rule).

### 3.5.1. Policy stage and the choice of stopping rule

In this section I explain how the decision analyst should approach the policy-level selection of stopping rules. Since the selection of the stopping rule is prior to the conduct and monitoring of RCTs, we may regard the choice of stopping rule as a policy decision. But in order to understand such policy, it is important that we first clarify the most basic duty of the DMC. This is important because our choice of stopping rule should accord with this duty. I begin by explaining what this primary duty is, and then proceed to explain how selection of the stopping rule should accord with this duty.

The U.S. FDA Guidance for Clinical Trial Sponsors (GCTS) states: "a fundamental reason to establish a DMC is to enhance the safety of trial participants in situations in which safety concerns may be unusually high" (GCTS, section 2.1 p.4). If by “a fundamental reason” we mean that a DMC’s first priority is to protect—and when possible to increase—the well-being of trial participants, then any reasonable DMC procedure for selecting stopping rules should reflect this duty. This means choosing stopping rules that are safe to trial participants. The relevant notion of ‘safety’ in this context must be complex, since trials of new treatments cannot rule out exposure to harm. But if we assume that the DMC, together with the primary investigator of the trial and trial participants, can agree on a
threshold of acceptable harm for their RCT—e.g., on what constitutes undue and unnecessary harm to trial participants—then we can propose ways for choosing safe stopping rules. The problem, however, is to clarify this basic duty of the DMC to protect trial participants.

There are at least two ways of clarifying the basic duty of the DMC, both of which involve examining the justification for this duty. One way is on utilitarian grounds. A version of utilitarianism, more specifically a version of rule-utilitarianism, may see choosing safe stopping rules as a means of promoting the overall benefits of every person in the long run. That is, as long as in the long run everyone benefits from the choices of these safe rules, the rules are good choices. ‘Everyone’ could mean different groups of individuals. It could mean every participant in the trial, every person in the population who might benefit from the new treatment, every tax payer or voter under the trial’s jurisdiction, or every person on the planet. The point is that the DMC needs to weigh the short-term interest of a larger population, in ignoring rules that enhance the safety of trial participants, against the long-term interests of maintaining the very institution of RCTs from which everyone benefits. The hope is that we might get a clearer sense of an acceptable threshold of harm by developing this rule-utilitarian approach.

The second way of justifying the basic duty of the DMC is by appealing to a form of the Rawlsian ‘difference principle’, i.e., by adopting a ‘maximin’ strategy for choosing our stopping rules. Stopping rules should be chosen in such a way that they give greatest weight (or highest priority) to the interests of patients who are most in need of health protection. The selected stopping rule should aim at safeguarding trial participants who are
gravely ill (‘less well off’) from becoming tools as DMCs and investigators pursue strong
evidence for or against a treatment—e.g. when pursuing overwhelming evidence sufficient
to change the minds of skeptic clinical practitioners.

These two approaches to the “basic duty” to promote safety, proceeding from very
general ethical principles, only take us so far. At this point, I switch to a practical approach
that starts with concrete guidelines for trials, modeled on Salbu (1999). My intention is
that these guidelines will be broadly in accord with both the utilitarian and Rawlsian
approaches sketched above. Salbu’s approach does not target DMCs directly, or the
problem of choosing a stopping rule; rather, his focus is federal-level policies. In this
respect, my ‘policy stage’ evaluation of stopping differs from that of Salbu. But in many
respects, our approaches overlap. As for Salbu, my approach is based on recent case
studies in which “the dilemma of opposing risks” was salient. My ‘summary’ rules, like
Salbu’s, are low-level inductive generalizations from experience. Like Salbu’s framework,
my framework proposes “extrapolating guidelines” for establishing the appropriate level of
rigor for stopping rules. Like his, my framework aims at serving “the public interest.”

100 In a law study review that examines the U.S. FDA’s drug approval process, Salbu (1999) presents an
interesting characterization of the “fundamental dilemma” that the FDA faces. When looking at the conflicting
objectives of caution and expedience that confront the Agency during each new drug application, Salbu
explains how while fulfilling its mission to monitor and control the safety and efficacy of drugs, “the Agency
continually walks a razor’s edge between two opposing risks—premature approval of dangerous drugs and
undue delay in making safe, effective, and medically useful drugs available to the public.” (ibid p. 96) This
predicament is characterized as “the FDA’s fundamental dilemma,” which “concerns the unavoidable
tradeoffs between under-conservatism and over-conservatism in the drug approval process.” (ibid 97)
Even though Salbu’s examination is directed originally at the level of federal policy, with the chief aim of
“proposing congressional and administrative guidelines for reducing the dissonance between the two policy
goals that the Agency face,” with some qualifications, much of the “FDA’s fundamental dilemma” is also
applicable to the type of dilemma and the level of scrutiny that DMCs often have to face when monitoring
RCTs and have to decide on whether to stop or continue them.
But I use the term ‘summary’ rules with some hesitation. That is because the generalizations I suggest at the policy level, as guides for selecting safe stopping rules, are not simply generalization from past DMC behavior. It is not as if we can look back at past cases of HIV/AIDS trials, and find what decisions have made ‘good’ choices of stopping rules. They are supposed to describe the appropriate behavior DMCs are allowed to engage in when choosing stopping rules that enhance the safety of trial participants, given that we do not rule out the need for some exposure to harm. In this, they go beyond past cases.

At the heart of my proposed approach to policy-level selection of stopping rules lie three 'heuristics'. These three considerations are not independent of one another, nor is the list supposed to be exhaustive. All three principles, however, illustrate the appropriate way of choosing safe stopping rules in the interest of trial participants. The three relevant generalizations are:

1. **Availability.** The conservatism of the stopping rule should correlate positively with the availability of safe and effective alternatives to the drug treatment under review. That is because, roughly speaking, the more effective and safe the available alternatives, including the standard treatment, the less urgent the need for early stopping due to either efficacy or futility.

2. **Type of intervention.** The conservatism of the stopping rule should correlate positively with any dependency of treatment effectiveness on indefinite medication and any increased potential for health hazards associated with long-term use. If the situation pertains to patients experiencing a life-
threatening illness such as advanced stages of HIV/AIDS, early evidence of efficacy may be sufficient—compelling enough—to warrant stopping the trial even if it is unknown whether early benefits are sustained over the long-run. In this case, it makes good sense to imagine that stopping rules should be less conservative precisely for taking advantage of the treatment’s short-term benefit. The mere possibility of a late treatment effect that is contrary to an early beneficial effect is not a sufficient reason to continue an RCT. If, on the other hand, the situation pertains to patients who are mostly healthy, such as asymptomatic HIV individuals (or patients in the early stages of AIDS), then long-term effects of the treatment can be considered of greater importance and balanced against early evidence of efficacy. In this type of intervention, giving greater importance to longer-term outcomes makes good sense, and thus warrants more conservative stopping rules.

3. **Prediction.** The conservatism of the stopping rule should correlate negatively with the degree of reliability indicated by the primary end point as a predictor of efficacy. The capacity of the primary endpoint to predict the primary event of interest (e.g. survival) is also a relevant factor. Simply put, surrogate markers (e.g. CD4 cell-count or viral load), due to their lower degree of reliability in predicting clinical responses as compared with more reliable clinical endpoints (e.g. mortality or morbidity), warrant a different weight for similar responses early in the trial. It seems reasonable to expect differences in the amount of evidence to compensate for the difference in the reliability of evidence. Thus, for surrogate markers, it seems reasonable to
demand more stringent stopping rules than if the primary endpoint is not a surrogate.

The decision analyst should take into consideration these three heuristics when comparing the appropriateness of each stopping rule for the context in which we find the RCT. Once a specification of the case is at hand, the decision analyst assesses whether or not the choice of stopping rule corresponds to the degree of conservatism the situation permits.
Chapter 4. Quantitative framework

In this chapter I introduce a decision-theoretic framework (DTF) for two types of decisions faced by a DMC: the initial choice of a data monitoring plan, and subsequent decisions about whether or not to stop a trial early. Section 4.1 presents the formal elements constituting the DTF: the basic observation data model, a set of possible states of nature, a set of available actions, loss functions, and one or more potential stopping rules. This section also provides examples of computer simulations implementing the framework for comparing decision stopping rules. The examples rely upon some simplifying restrictions (to be explained below): discrete loss functions, dichotomous observations, a single pair of treatments, two hypotheses, two loss functions, and four types of available action. Even with these restrictions, the examples show how the DTF provides a useful means of modeling both the choice of a monitoring plan and early stopping decisions.

Sections 4.2 through 4.4 expand the basic framework through application to the three principal types of stopping decisions discussed in the preceding chapters: stopping due to efficacy, stopping due to futility, and stopping due to harm.
4.1 Decision-theoretic framework (DTF)

To choose a clinical trial design is to make a decision. The primary objective of the DTF to be introduced here is the evaluation of statistical monitoring plans. The choice of such a plan is treated as a decision problem under uncertainty, represented as a 4-tuple \((\Theta, \Delta, Y, L)\). \(\Theta\) is our parameter space (the set of possible true states), \(\Delta\) is a set of possible monitoring plans (or monitoring rules), \(Y\) is the observation (data) model, and \(L\) is a specific loss function. We shall also refer to a set \(A\) of basic (or allowable) actions, e.g., to stop and declare a preferred treatment. The members of \(\Delta\) are functions from observed data into this set of basic actions; we write \(\delta(y): Y \rightarrow A\) for a statistical decision monitoring rule.

In order to illustrate DTF for comparing statistical decision monitoring rules, we keep our example as simple as possible. An approximate analysis considering relevant features will suffice for our purposes. We model the event of interest (e.g. the progression to AIDS) as a binary event (e.g., whether or not HIV+ progressed to AIDS for a given HIV+ patient). The example in this section involves dichotomous observations, a single pair of possible treatments, two hypotheses, two loss functions, two statistical decision rules (in this case employed by U.S. and European researchers), one “overarching” maxim, three types of basic actions, and easy to compute numbers.

Suppose \(N\) individuals suffer from HIV+. We suppose that there are just two possible treatments: the standard treatment \(T_1\) and the experimental treatment \(T_2\). In order to find which of the two treatments is more effective, we conduct a randomized controlled trial on
2n of the total N patients, with n patients assigned to each treatment. Suppose the remaining \( N - 2n \) patients receive the treatment selected as the more effective of the two when the trial ends, unless no treatment is declared superior, in which case the remaining patients will be treated with the standard treatment \( T_1 \).

We restrict our attention to statistical monitoring rules that permit termination after \( n/2 \) (half the number of patients assigned to each treatment group) have been treated. That is, the RCT has a single interim analysis halfway through the intended trial. Keeping calculations relatively simple, we treat the monitoring of 10 pairs of HIV+ patients (\( n=10 \)), i.e. 5 patients assigned to each treatment by interim analysis, and 10 patients assigned to each treatment by final analysis, with \( N=100 \). These restrictions entail some loss of generality. For instance, they do not allow a systematic representation of all statistical monitoring procedures that would be available in practice. On the other hand, the simplifications allow the basic structure of the framework to emerge clearly.

Before elaborating on the decision-theoretic model, a few remarks are in order concerning both the variables to be included and the particular values assigned to those variables.

Guided primarily by the DMC mandates—as noted in section 3.3 (Classification and identification)—the set of variables included in the reconstruction of early stopping decisions should include those having to do with scientific merit (e.g. control of type I and type II errors as expressed in the statistical monitoring rule \( \delta(y) \)) and the protection of trial participants (e.g., number of trial participants \( n \), the balance of foreseeable risks and
benefits). I shall also add variables related to the protection of the public at large (e.g., patient horizon N). Because people may choose to participate in a trial for the benefit of others, (as opposed to participation for their own benefit), the reconstruction of the decision should account for such facts whenever possible in its loss function. This consideration justifies the incorporation of N, since (strictly speaking) promoting the welfare of the public at large goes beyond the narrowest construal of the DMC mandate.

As to the particular values that I assign to these key variables in my initial example: they are not arbitrary. They are selected with documented DMC mandates in mind. But because I am not at a stage where the decision model is usable prospectively in a trial, the values assigned in the model are intended primarily for the purposes of illustrating the quantitative framework and the sort of considerations that the DMC should have were it to follow its mandates.

This leads to a highly contentious question: whose utility values are supposed to be represented in a given reconstruction? If we follow the DMC mandates, it is reasonable to imagine that the specific utility values should be those reflecting the values of trial participants first and foremost. I shall adopt this assumption, but with the caveat already noted that trial participants commonly attach significant value to the well-being of the wider community.

There is a further pressing issue: how are individual participant utilities to be aggregated? The model leaves this question open. Lacking a method for handling multiple

---

101 This generalizes on a point noted in section 3.3, that community-based norms may influence a DMC decision.
trial participants with different utility values, the decision framework is unsuitable for some reconstructions. For instance, where the difficulty of obtaining agreed utility values among all \( n \) trial participants is an important issue, the usefulness of my decision framework is limited. But for the purposes of illustrating how the framework functions, I shall assume agreed-upon aggregate utility values.

In general: in representing difficult decisions, it is helpful not to be dogmatic about any particular set of variable values. Rather, we should think instead about whether DMC decisions would stand up under a range of values; alternatively, we can employ the framework to answer ‘what-if’ questions that reflect different value assignments.

The next few sub-sections spell out the components of the DTF, identifying the simplifications.

### 4.1.1. Observation model

For each of our two treatments and each patient, we shall model our observation as yielding a single outcome: whether or not the HIV+ patient progressed to AIDS. Our observations are thus expressed by two integers, \( y_1 \) and \( y_2 \), which denote the number of cases of no progression among the \( n \) patients using treatments \( T_1 \) and \( T_2 \) respectively. Following with our example, at interim, \( n = n/2 = 5 \), thus \( y_1 \in \{0, 1, 2, 3, 4, 5\} \); whereas at the end of the trial, \( n = n = 10 \), thus \( y_1 \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \).
Suppose that there is a fixed probability for each patient of no progression to AIDS with treatment $T_i$, $\theta_i$ ($i = 1, 2$) (assumed independent and constant from patient to patient). We assume that our observations will follow the binomial distribution:

$$p(y_i | \theta_i) = \binom{n}{y_i} \theta_i^{y_i} (1 - \theta_i)^{n-y_i} \quad (i = 1, 2)$$

Here, $p$ is the probability of getting exactly $y_i$ “no progressions” ('successes') in $n$ patients ('trials'), for $y_i = 0, 1, 2, \ldots n$.

Since we are interested in the difference between $T_1$ and $T_2$, let $\theta$ stand for the difference in response rates between $T_1$ and $T_2$: $\theta = \theta_2 - \theta_1$. Thus, the joint distribution of $y_1$ and $y_2$ given $\theta_1$ and $\theta_2$ is:

$$f(y | \theta) = \prod_{i=1}^{2} \binom{n}{y_i} \theta_i^{y_i} (1 - \theta_i)^{n-y_i} \quad (i = 1, 2)$$

$y$ and $\theta$ denoting $(y_1, y_2)$ and $(\theta_1, \theta_2)$, respectively; $y \in Y$ and $\theta \in \Theta$. 
4.1.2. Parameter space $\Theta$

Let us assume that the no progression rate of $T_1 (\theta_1)$ is 0.5 and that of $T_2 (\theta_2)$ is unknown. Moreover, assume that $\theta$ takes one of only two possible values (so that we need only consider two hypotheses):

$$H_0: \theta = 0 \quad \text{(}T_2 \text{ is equal to } T_1 \text{ i.e., } \theta_2 = \theta_1 = 0.5\text{),}^{102} \text{ or}$$

$$H_1: \theta = 0.2 \quad \text{(}T_2 \text{ is more effective than } T_1 \text{ i.e., } \theta_2 = 0.7)$$

For our example, we can list the probability mass functions $p(y_2 | \theta_2)$ with $n = 10$ for the two possible values of $\theta_2$:

<table>
<thead>
<tr>
<th>$y_2$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_2 = 0.5$</td>
<td>0.001</td>
<td>0.010</td>
<td>0.044</td>
<td>0.117</td>
<td>0.205</td>
<td>0.246</td>
<td>0.205</td>
<td>0.117</td>
<td>0.044</td>
<td>0.010</td>
<td>0.001</td>
</tr>
<tr>
<td>$\theta_2 = 0.7$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.009</td>
<td>0.037</td>
<td>0.103</td>
<td>0.200</td>
<td>0.267</td>
<td>0.233</td>
<td>0.121</td>
<td>0.028</td>
</tr>
</tbody>
</table>

4.1.3. Set $A$ of basic (allowable) actions

Still keeping things simple, our set $A$ of basic actions will consist of only four possibilities. Given the above assumptions, at the interim analysis, we can either continue the trial or stop with one of two declarations:

---

$^{102}$ By ‘equal’ I do not mean that $T_1$ and $T_2$ are literally equal, or that they reduce to the exact same treatment effect. ‘Equal’ here means that $T_1$ and $T_2$ are practically indistinguishable given a sufficient measure of difference between treatment effects.
a₁: Stop and declare “T₂ is more effective than T₁”; or

a₂: Stop and declare “T₂ is equal to T₁”,

If we continue the trial at interim, then at the end of the trial we can act in one of two ways:

a₁.f: Stop and declare “T₂ is more effective than T₁”; or

a₂.f: Stop and declare “T₂ is equal to T₁”.

At either the interim analysis or the end of the trial, the choice of a basic action is made on the basis of sample data \( y_1 \) and \( y_2 \). Specific sample data expressed by the pair \( (y_1, y_2) \) leads to the choice of action to take. This choice depends ultimately on the earlier choice of decision monitoring rule. In concise terms: a decision monitoring rule is a function from sample data \( y \) into the set \( A \) of allowed actions \{a₁, a₂, a₁.f, a₂.f\}. It specifies how actions are chosen, given observation(s) \( y \). The next sub-section elaborates.

4.1.4. Decision monitoring rule \( \delta(y):Y \rightarrow A \) and the set \( \Delta \) of such rules

We first consider how we might represent candidate monitoring rules by using simplifications of two such rules that emerged in the AZT trials of the 1980s: the U.S. decision monitoring rule illustrated in Figure 4.1 below, and the European Concorde rule.
(Figure 4.2). The U.S. rule is based on frequentist properties (Neyman-Pearson model), where the focus is on controlling error probabilities. We use an instance of a group sequential monitoring procedure much like the one adopted by U.S. researchers. Like the U.S. monitoring plan, this rule aims at controlling the overall type I error by having it be no more than 0.05 (approx.). For this, consider U.S. researchers adopting the following rule:

**At interim (when n = 5):**

a₁: Reject $H_0$ (assume $H_0$ is false) for values $(y_2 - y_1) \geq 4$, i.e., for $(y_2 - y_1) = \{4, 5\}$; stop and declare “$T_2$ is more effective than $T_1$”;

a₂: Accept $H_0$ (assume $H_0$ is true) for values $(y_2 - y_1) \leq -4$, i.e., for $(y_2 - y_1) = \{-4, -5\}$; stop and declare “$T_2$ is equal to $T_1$”.

This means that for $(y_2 - y_1) = \{-3, -2, -1, 0, 1, 2, 3\}$, researchers continue with the trial.

**At the end of the trial (when n = 10):**

a₁.f: Reject $H_0$ (assume $H_0$ is false) for values $(y_2 - y_1) \geq 4$, i.e., for $(y_2 - y_1) = \{4, 5, 6, 7, 8\}$; stop and declare “$T_2$ is more effective than $T_1$”, otherwise

a₂.f: Do not reject $H_0$ for $(y_2 - y_1) = \{-8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3\}$; declare “$T_2$ is equal to $T_1$”.

---

103 Notice that values of $(y_2 - y_1) = \{9, 10\}$ are not available at the end of the trial, because the trial would have stopped at interim according to a₁.
Why would the U.S. researchers choose such a rule? Perhaps because their aim is to control the overall type I error. The adopted rule controls the overall probability of type I error, keeping it to 0.058. That is, at the end of the trial the probability $P((y_2 - y_1) \geq 4 \mid H_0) = 0.058$. The rule controls type I error by having the probability of type I error at interim set to 0.01, which is a more stringent value (threshold) than the probability of type I error at the end of the trial, which is 0.05.

Figure 4.1 shows how this decision monitoring rule works as a function of possible values ($y_2 - y_1$), both at interim and at the end of the trial. Specifically, the figure plots the range of ($y_2 - y_1$) values at the end of the trial—indicated by the x axis, i.e. final (treatment − control)—conditional on values ($y_2 - y_1$) at interim—indicated by the y axis, i.e. interim (treatment − control). $a_1$ is represented by the set of black circles, $a_2$ is represented by the set of red squares, $a_1.f$ is represented by the set of purple triangles, and $a_2.f$ is represented by the set of white squares. Grey circles represent impossible values for ($y_2 - y_1$). For example, coordinate (X=6, Y=0) is not an option, given that ($y_2 - y_1$) = 0 at interim eliminates the possibility of ($y_2 - y_1$) = 6 at the end of the trial.

---

104 Similarly to previous footnote, values of ($y_2 - y_1$) = {-9, -10} are not available at the end of the trial, because the trial would have stopped at interim according to $a_2$.
105 $P((y_2 - y_1) \geq 4 \mid H_0) = 0.058$, computed according to the joint probability distribution. E.g., $p((y_2=5, y_1=0) + p(y_2=5, y_1=1) + p(y_2=4, y_1=0) \mid H_0) = 0.014$.
106 The probability of type I error at interim is calculated by $P((y_2 - y_1) \geq 4 \mid H_0) = [p(y_2=5, y_1=0) + p(y_2=5, y_1=1) + p(y_2=4, y_1=0) \mid H_0] = 0.01$.
107 The probability of type I error at the end of the trial is calculated by $P(8 \geq (y_2 - y_1) \geq 4 \mid H_0) = 0.05$, represented by $a_{1.f}$ (purple triangles) in Figure 4.1.
Figure 4.1 U.S. decision monitoring rule as a function of possible values of \((y_2 - y_1)\), with \((y_2 - y_1)\) at the end of the trial plotted as a function of \((y_2 - y_1)\) at interim. Actions: \(a_1\) (black circles), \(a_2\) (red squares), \(a_1.f\) (purple triangles), and \(a_2.f\) (white squares). See appendix table B.1 for additional details.

For the Concorde decision monitoring rule, we shall assume they used a Bayesian monitoring rule for reasons we provide later. This rule requires the specification of prior information about \(\theta\) expressed in terms of a prior probability function \(p(\theta)\). For simplicity, we consider neither an optimistic nor a pessimistic set of priors, but ‘flat’ priors: \(p(\theta = 0) = \)
\[ p(\theta = 0.2) = 0.5. \] When using a Bayesian monitoring rule, the evidence provided by data \( y \) is contained in the likelihood ratio which is multiplied by a factor, the ratio of prior probabilities, producing the ratio of posterior probabilities. Therefore, when choosing between \( H_0 \) and \( H_1 \) on the basis of \( y \), the rule chooses the hypothesis with larger posterior probability. For instance, putting it in terms of rejecting \( H_0 \), the rule may reject \( H_0 \) when the likelihood ratio is less than 1. If so, with flat priors, at the end of the trial, the rule may reject \( H_0 \) as long as \( y_2 \geq 6 \); otherwise it does not reject \( H_0 \).

<table>
<thead>
<tr>
<th>( y )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(y</td>
<td>H_0) )</td>
<td>165.382</td>
<td>70.878</td>
<td>30.376</td>
<td>13.018</td>
<td>5.579</td>
<td>2.391</td>
<td>1.025</td>
<td>0.439</td>
<td>0.188</td>
<td>0.081</td>
</tr>
<tr>
<td>( p(y</td>
<td>H_1) )</td>
<td>70.878</td>
<td>30.376</td>
<td>13.018</td>
<td>5.579</td>
<td>2.391</td>
<td>1.025</td>
<td>0.439</td>
<td>0.188</td>
<td>0.081</td>
<td>0.035</td>
</tr>
</tbody>
</table>

How are we to choose between these two candidate monitoring rules? That’s where the DTF comes in. However, as a preliminary step to making the two monitoring rules comparable, despite their different statistical philosophies, we need to make an adjustment. We need to choose a ‘cutoff’ point, at the end of the trial, for the Concorde rule to be as close as possible to the U.S. one, while keeping with our choice of easy to compute numbers. What we are looking for is a factor (a corresponding likelihood ratio) such that the range of \((y_2 - y_1)\) values whose sum of probabilities is the closest the rule can get to the type I error value of 0.05 used by U.S. researchers.

Given our choice of easy to compute numbers (\( n = 5 \) at interim, and \( n = 10 \) at the end of the trial), the answer is: a ‘cutoff’ of 0.25/0.75. That is, it is not sufficient that the rule rejects \( H_0 \) when the likelihood ratio is < 1. Rather, the rule must reject \( H_0 \) when the
likelihood ratio < 0.25/0.75, which means for values of \((y2 - y1) = \{8, 9, 10\}\). Figure 4.2 shows such a rule.

One important difference to notice between Figure 4.2 and Figure 4.1 is that, because the evidence provided by data \(y\) is contained in the likelihood ratio, the probability values of the control group are irrelevant. They cancel out in the ratio because \(\theta_1 = 0.5\), regardless of whether \(H_0\) is true or \(H_1\) is true. Only \(\theta_2\) varies, namely, the treatment’s effect.

\[
\frac{p(H_0 \mid y)}{p(H_1 \mid y)} = \frac{p(H_0)p(y \mid H_0)}{p(H_1)p(y \mid H_1)} = \frac{p(H_0)p(y_1 \mid H_0)\cancel{p(y_2 \mid H_0)}}{p(H_1)p(y_1 \mid H_1)\cancel{p(y_2 \mid H_1)}} = \frac{p(y_2 \mid H_0)}{p(y_2 \mid H_1)},
\]

when priors \(p(H_0) = p(H_1)\).
Figure 4.2 Concorde decision monitoring rule as a function of possible values of \((y_2 - y_1)\), with \((y_2 - y_1)\) at the end of the trial plotted as a function of \((y_2 - y_1)\) at interim. Actions: \(a_1\) (black circles), \(a_2\) (red squares), \(a_1.f\) (purple triangles), and \(a_2.f\) (white squares).

4.1.5. Loss function

The loss function is meant to incorporate the ethical and epistemic values attached to possible outcomes of the trial. We start with the ‘ethical’ loss function, \(L_E(\theta, a)\). This function can be used to compare the two treatments by paying a penalty for each patient
assigned the inferior treatment. One penalty point is assessed for each patient assigned the inferior treatment.

One way to think of \( L(\theta, a) \) is to approach it with respect to a particular patient. If \( H_0 \) is true, and if the patient is given \( T_1 \) the loss incurred with such treatment is 0 (since \( T_1 \) and \( T_2 \) are considered equivalent); but if the patient is given \( T_2 \) in this situation, then there is a cost, which I shall refer to as \( cc \).\(^{108}\) The other possibility is that \( H_1 \) is true. In this case, if the patient is given the inferior treatment \( T_1 \), the loss is \( d \) (which we assume for now is 1 unit); if given \( T_2 \) (superior treatment) the loss is 0. The table below summarizes this situation for the individual patient:

<table>
<thead>
<tr>
<th></th>
<th>( H_0 )</th>
<th>( H_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_1 )</td>
<td>0</td>
<td>( d )</td>
</tr>
<tr>
<td>( T_2 )</td>
<td>( cc )</td>
<td>0</td>
</tr>
</tbody>
</table>

Now let us make \( L(\theta, a) \) sensitive to the “effect size” \(|\theta_2 - \theta_1|\). By “effect size” we mean the percentage difference of no progressions between treatments. Thus, assuming \( H_1 \) is true (\( \theta = 0.2 \)), for every \( n=5 \) patients (every segment of interim analysis), the single loss unit is the loss expected (by association) from one fewer patient having a positive no progression. That is 0.2 of five patients, which equals 1; it is the result that researchers should have obtained (or could have expected) had researchers continued with the trial

\(^{108}\) cc is a loss that can be conceptualized in different ways. For my purposes, it can be understood as the “cost” of having a patient subjected to a new treatment, when that treatment is no better than standard treatment.
until the next interim. Below are sample losses according to this new version of $L_E(\theta, a)$, accounting for effect size, $n$, and $N$, for values of $d = 1$ and $cc = 0.01$:

<table>
<thead>
<tr>
<th>Action</th>
<th>True state of nature</th>
<th>$\theta_2 - \theta_1 = 0$</th>
<th>$\theta_2 - \theta_1 = 0.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>@interim ($n = 5$)</td>
<td>a$_1$: choose $T_2$</td>
<td>$(N - n)cc = 0.95$</td>
<td>$(\theta_2 - \theta_1)(n d + (N - 2n)\theta) = 1$</td>
</tr>
<tr>
<td></td>
<td>a$_2$: choose $T_1$</td>
<td>$0$</td>
<td>$(\theta_2 - \theta_1)(n d + (N - 2n)\theta) = 19$</td>
</tr>
<tr>
<td>@final ($n = 10$)</td>
<td>a$_1$.f: choose $T_2$</td>
<td>$(N - n)cc = 0.9$</td>
<td>$(\theta_2 - \theta_1)(n d + (N - 2n)\theta) = 2$</td>
</tr>
<tr>
<td></td>
<td>a$_2$.f: choose $T_1$</td>
<td>$0$</td>
<td>$(\theta_2 - \theta_1)(n d + (N - 2n)\theta) = 18$</td>
</tr>
</tbody>
</table>

4.1.6. Decision criteria

Average loss is obtained by averaging the loss function over all possible observations:

$$E_y[L(\theta, \delta(y))| \theta] = \sum_{y, \delta(y)} L(\theta, \delta(y)) f_y(y_i| \theta)$$
If we have prior information about $\theta$ which can be expressed in terms of a prior probability mass function $p(\theta)$, then the Bayes risk of decision rule $\delta(y)$ is the expectation of the average loss over possible values of $\theta$:

$$r(\delta(y)) = \sum_{\theta \in \Theta} \sum_{y \in Y} L(\theta, \delta(y)) f_y(y | \theta) p(\theta)$$

In our example of early stopping due to efficacy, I use Bayes risk as the means (i.e. the rationale) for comparing the U.S. monitoring rule against the Concorde monitoring rule. With the components of the DTF in place, we can now turn to this comparison.

### 4.2 Early stop due to efficacy

#### 4.2.1. Simulation

The essential idea of the DTF is that the decision to stop or continue a trial is made by weighing the consequences (losses) of possible actions, averaging over the distribution of future observations. Crucially, the future (as yet unobserved) results are given a probability distribution. Consider the situation of the U.S. researchers at the end of their first interim analysis, with 0.05 p-value in favor of $T_2$ using their sequential monitoring
rule. How does it compare with Concorde’s decision monitoring procedure given our ethical loss function?¹⁰⁹

To compare both DMC decisions, an assessment of the associated losses expected by their respective decisions is needed. Figure 4.3 shows, for a set of different effect sizes, the weighted-losses ¹¹⁰ for stopping and continuing, according to U.S. and Concorde monitoring procedures, when $H_1$ is true.

¹⁰⁹ The treatment of my early stop due to efficacy case can also be found in my (2011) paper “Statistical decisions and the interim analyses of clinical trials” Theoretical Medicine and Bioethics, Vol. 32:61-74.
¹¹⁰ Weighted-loss = (loss)*(probability of taking that action).
Figure 4.3 Weighted-losses of stopping (acting a1) and continuing (averaging actions a1.f and a2.f) for each procedure, when $H_1$ is true.

We shall assume that the decision to stop the US trial was based on whether ending it had a lower expected loss than continuing. The expectation here is with respect to the weighted-losses of continuing in light of the fixed set of future actions, a1.f or a2.f. By averaging the weighted-losses of a1.f and a2.f, one can compare the weighted-losses of stopping versus continuing for a whole range of effect sizes. Notice that for every effect size in the considered range, the US decision to stop the trial has a lower expected loss than continuing. The decision to stop is therefore in agreement with every effect size in the
range. This, however, is not the case with Concorde’s decision monitoring rule. Given my ethical loss function, and assuming that Concorde’s decision to stop was also based on whether ending the trial had a lower expected loss than continuing, the only effect size that could have warranted a decision to continue was a high expected treatment size, i.e., $\theta_2 = 0.9$. Otherwise, the trial should have stopped since stopping at interim had a lower expected loss than continuing.

On the other hand, if one focuses attention on a particular effect size $\theta_2 = 0.7$, the weighted-losses for Concorde’s actions $a_1$, $a_1.f$, and $a_2.f$ are 0.17, 0.48, and 10.06, respectively. Once one averages $a_1.f$ and $a_2.f$, one notices that in order to account for Concorde’s decision to continue—based on a lower expected loss than stopping—one needs the loss incurred by a mistreatment for a given patient early in the trial to have been approximately 31 times greater than the penalty of $d = 1$ unit per mistreatment assumed in the ethical loss function $L_E(\theta, a)$. In other words, for Concorde’s monitoring procedure to support continuation, under my DTF principle of ending the trial as soon as doing so has a lower expected loss than continuing, the loss incurred from the inferior treatment would need to have been 31 times greater than the one adopted by the US trial (assuming US researchers made their decision based on the ethical loss function $L_E(\theta, a)$).

Let us take stock of where we are in our comparison of monitoring rules and stopping decisions. A guiding principle is that we should try to represent both DMC decisions as reasonable and to justify both decisions as ‘equally’ ethical, under the very

\[\frac{(0.48 + 10.06)}{2} = 5.27,\] and once divided by 0.17, we reach the multiplicative loss factor of 31.
same maxim. In order to adhere to this principle, we must consider the range of effect sizes under different loss functions.

One contender is the ‘scientific’ loss function, which represents the idea that value attaches only to finding the true state of nature and ignores all other consequences. In particular, this loss function ignores the effects of mistreatment. From the point of view of the scientific loss function, correctly declaring that “T₂ is more effective than T₁” has the same utility as correctly declaring that “T₂ is equal to T₁”, even though the two states of nature may have quite distinct consequences for present and future patients. In other words, the function assigns the same penalty for any incorrect conclusion. To be definite, let us represent the penalty as a 10 unit loss. The table below summarizes this loss function:

<table>
<thead>
<tr>
<th></th>
<th>H₀</th>
<th>H₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>T₂</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

Let us return to our comparison of the U.S. and Concorde rules, evaluating the Bayes risk but using the scientific loss function in place of the ethical loss function. Figure 4. shows how the two monitoring rules perform under Bayes risk, for the range of effect sizes, under each loss function. By assuming an equal set of priors for every pair of hypotheses (each pair having a different effect size), the U.S. monitoring rule is outperformed by the Concorde’s rule with respect to the Bayes risk except when the effect size is small, θ₂=0.6. This is shown by the left graph in Figure 4.4.
With respect to the Bayes risk, the U.S. rule is therefore ‘less’ ethical (has a greater Bayes risk) than Concorde's when the effect sizes are not small, i.e., for values $\theta_2 = \{0.7, 0.8, 0.9\}$. That is, for all effect sizes, except when $\theta_2=0.6$, the Bayes risk for Concorde's rule is smaller than the U.S. rule. That is no surprise, given the fact that Concorde’s rule is a Bayesian monitoring rule.

Obviously, a different set of priors would produce different Bayes risks. Had the Concorde researchers been optimistic regarding AZT’s benefits (e.g. $P(H_1)=0.8$, $P(H_0)=0.2$), Concorde’s Bayes risk curve would have resulted in a shift upwards in the left plot of Figure 4. In that case, for all effect sizes except $\theta_2=0.9$ (when the effect is largest), the US decision monitoring would have outperformed Concorde’s rule with respect to Bayes risk.

Notice here how the situation is reversed if we shift from the ‘ethical’ to the ‘scientific’ loss function. The comparison between the U.S. and Concorde monitoring rules is illustrated by the graph in Figure 4.4. In the graph on the right, the U.S. monitoring rule outperforms the Concorde’s rule with respect to the Bayes risk in all effect sizes, except when effect is large, $\theta_2=0.9$. With respect to this function, and using Bayes risk as a comparative rationale for their adopted rules, the U.S. rule is therefore overall ‘more’ ethical then Concorde’s except when the effect is large, i.e., $\theta_2=0.9$. 
4.2.2. Discussion

The above analysis shows how the proposed decision-theoretic framework allows disparate statistical decision rules to be compared in their appropriate context, while providing important considerations for deciding on trial conduct based on a range of epistemic and ethical factors.

In hard-case clinical trials like the AZT example, the formalization of the different DMC decision rules assists us in reconstructing the complicated decisions that took place in those trials; in particular, it makes the pertinent interim statistical decisions more precise. This is a desirable result, since it facilitates a possible dialogue regarding the two
apparently conflicting DMC decisions. By explicitly stating the decision rule, decision criteria, and potential loss functions, the framework permits those interested in these particular AZT trials to make an informed judgment regarding the two DMCs’ different early stopping decisions. If a member of the lay public is uncomfortable with a particular rule or statistical criterion chosen for early stopping, then she or he is given a means to compare and thus object to either DMC’s monitoring decision, whether on ethical or epistemic grounds.

Notice: in the first instance, the framework operates under a Principle of Charity, namely, a requirement to represent each DMC decision as rational. As just pointed out, however, this important initial step provides the resources for raising objections to DMC. By introducing precise parameters of cost and probability, we are able to construct ‘what-if’ scenarios.

We compared the rules under three circumstances: varying effect size (but holding fixed the ethical loss function), varying the loss function, and using Bayes risk. A virtue of the present account is that it lets us make such comparisons, and it makes explicit what assumptions would have to be in play for a particular stopping decision to be justified.

In the first comparison varying effect size, we saw that the U.S. decision to stop the trial was justified for every effect size considered. That was not the case with the Concorde decision to continue its trial. When varying effect size we saw that the only way we could make sense of Concorde decision to continue its trial was with a large effect size. For all other effect sizes Concorde would have had to stop its trial. In the second comparison, we
held fixed a particular effect size ($\theta_2=0.7$), using the same ethical loss function, and we saw that by increasing the loss incurred by a mistreatment for a given patient early in the trial by a factor of 31 rather than the penalty of 1 unit assumed in the U.S. rule, we could then make sense of the Concorde decision to continue its trial.

In our last pair of comparisons, we saw how using the ethical loss function, overall, the U.S. rule was outperformed by the Concorde rule (except when the effect size was small) using Bayes risk as the rationale for comparison. The situation was reversed if we used the scientific loss function. There the U.S. monitoring rule outperformed the Concorde’s rule with respect to the Bayes risk in all effect sizes, except when effect was large, $\theta_2=0.9$. With respect to the scientific loss function, and using Bayes risk as a comparative rationale, the U.S. rule was found to be overall ‘more’ ethical than Concorde’s rule.

Are the results generalizable? Even though the framework is limited, it attempts to reconstruct (in a plausible fashion) the arguments given for the decisions in each trial. The fact that neither trial provided justifications for its interim decision in the respective publications is already an indictment of these early stopping decisions. If the framework succeeds in giving plausible reconstructions, then justifications for the competing decisions exist. Without justifications, the generalizability of the DMCs’ decisions and the subsequent analysis of their decisions are problematic. By providing a common framework that structures justifications in the individual cases and comparisons between cases, we do indeed have the beginnings of a general model for interim decision-making.
But the framework leaves us with further questions. For instance, does the framework open some new way of seeing why the U.S. and European decisions were more than simply reasonable ones, suggesting some further social or political motivation behind their decisions? Is there some avenue for adjudicating statistical decision rules? In the next section, I shed further light on these questions by bringing the DTF approach to bear on a different case: a case of early stop due to futility. It is another case in which we have a clash of opinions about stopping, this time between the DMC and the principal investigator (PI). As in the AZT case, we will see that both perspectives can be rationalized; it will perhaps be more surprising to learn that the DMC decision to stop the trial and the PI’s decision to continue the trial can be justified as ethical under the very same maxim.

4.3 Early stop due to futility

4.3.1. Simulation

According to DTF, for cases of early stop due to futility, the decision to stop or continue is made by weighing the consequences of possible actions, given the interim results, and using the prospective probability for eventual rejection (or acceptance) of H according to conditional power (CP) calculations.

Consider the disagreement between the DMC and the principal investigator (PI) of the study described in section 2.2, *Toxoplasemic Encephalitis* (TE) in HIV+ individuals.
Earlier we saw that given the unexpectedly low TE event rates seen during interim, the DMC recommended to stop the trial. Despite the DMC recommendation, the PI recommended “continuing the study due to the uncertainties about the future TE event rate” (Neaton et al 2006, 321). Thus, the question we face is: how can we account for the difference between their recommendations, assuming both recommendations were reasonable?

When approaching a case with DTF, certain factors are best held fixed, while others emerge as crucial in explaining disagreements. For instance, how does the DMC recommendation that the trial cease compare with the PI’s recommendation that the trial continue?

As in the original study by Jacobsen et al (1994), suppose the proportion of Toxoplasmic Encephalitis (TE) under treatment $T_1$ is expected to be 30%, and that investigators estimate that a 50% reduction of TE (i.e. from 0.3 to 0.15) in treatment $T_2$ can be detected with 80% power at a two-sided significance level of 0.05.  

The first two components of the decision-theoretic model are: (1) for each treatment we assume observation as yielding a single outcome, namely, whether or not the HIV+ individual developed TE; and (2) we assume statistical monitoring rules that permit termination after $n/2$ trial participants have been treated. That is, the RCT has a single

interim analysis halfway through the trial—as opposed to four interim analyses in the original Jacobsen et al (1994) study.

In our model, we treat the monitoring of 120 pairs of patients (n=120), 60 patients assigned to each treatment by interim, and 120 patients to each treatment by final analysis, with N=1000. (Obs.: Given our assumptions to achieve 80% unconditional power we need n to be approximately 120.)

Suppose that at the first and only interim analysis, there were 3 patients lost to follow-up in the experimental group (60 – 3 = 57) and 1 lost to follow-up in the standard group (60 – 1 = 59). Furthermore, 3 patients under standard treatment and 3 patients under experimental treatment developed TE, as in Table 4.1.

<table>
<thead>
<tr>
<th></th>
<th>TE</th>
<th>No TE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard (T1)</td>
<td>3</td>
<td>56</td>
<td>59</td>
</tr>
<tr>
<td>Experimental (T2)</td>
<td>3</td>
<td>54</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>110</td>
<td>116</td>
</tr>
</tbody>
</table>

Table 4.1 Summarized interim data

---

113 This is the probability at the start of the trial of achieving a statistically significant result at a pre-specified significance level and a pre-specified alternative treatment size.
114 Computed according to two-proportion z-test.
115 By lost to follow-up I mean patients who have either become lost by being unreachable or opted to withdraw from the RCT at the interim point.
116 Similar to the original trial where the TE event rates for experimental and standard treatments were both low, i.e. 5.5% and 5.3% respectively.
In futility cases, because the dilemma of whether to stop or continue is modeled around the concern of whether interim results can still change, as to allow a favorable (statistically significant) result at the end of the trial, we need to compute conditional power (CP) even before bringing in a loss function to the model. That is because CP gives us the conditional probability of a significant result (rejecting $H_0$ under the overall type I error) at the end of the trial, given the interim data.

We compute conditional power (CP) according to Proschan et al (2006).\textsuperscript{117} With information fraction \( t = \frac{p(1-p)(2/n)}{p(1-p)(2/n+2/n)} = 0.48 \), where \( p = (0.3 + 0.15)/2 = 0.225 \), \( n = 120 \), with 0.05 alpha and 80% original power (\( z_\beta = 2.8 \); \( z_{\alpha/2} = 1.96 \)), substituting in \( CP_\theta(t) = 1 - \Phi\left\{\left[ Z_{\alpha/2} - t^{1/2} Z(t) - \theta(1 - t)\right]/(1 - t)^{1/2}\right\} \) we obtain \( CP(t) = 0.22 \).

Although not dismally low, the conditional power of 22% is much lower than the original power to detect the planned 50% difference TE rates between treatments. Since considerably low power suggests that the trial is unlikely to reach statistical significance even if the alternative hypothesis is true, this means, given the observed TE rates at interim, it is reasonable to consider the trial futile.

However, due to the uncertainties about the future TE event rate expressed by the PI, we need a way of computing CP that permits accounting for the possibility of uncertainties about TE event rate. Because predictive power (Bayesian CP) permits a greater freedom of choice for the alternative treatment effect $\theta$, we model the PI's

\textsuperscript{117} See Chapter 3 (section 3.2 Conditional Power for Futility) for details p.45-52.
computation as following predictive power. Given a prior distribution over possible values of \( \theta \), \( \pi(\theta) \), the PI had a wider range of values of efficacy to consider, each associated with a prior. In this fashion, the PI might have considered the probability that the test statistic reached the rejection region at the end of the trial, given the results at interim, for each value of \( \theta \).

Incorporating prior information into CP is relatively straightforward. It means computing CP for each value of \( \theta \), and then averaging them using values from \( \pi(\theta) \) as weights, thus producing a predictive power (Table 4.2).

\[
\begin{array}{cccc}
\theta & \pi(\theta) & CP & Weighted-CP \\
0.1 & 0.1 & 0.01 & 0.001 \\
0.25 & 0.15 & 0.04 & 0.006 \\
0.5 & 0.3 & 0.22 & 0.066 \\
0.75 & 0.25 & 0.59 & 0.148 \\
1 & 0.2 & 0.89 & 0.178 \\
\end{array}
\]

Predictive power = 0.40

Table 4.2 Predictive power at interim \((t=0.48)\)

The other component of the decision-theoretic model is the loss function. We start with an ‘ethical’ loss function, \( L_E(\theta, a) \). The main goal of using such a function is to compare both treatments by paying a penalty for each patient assigned the inferior treatment, while rewarding the assignments of superior treatment. One loss unit is counted for each patient assigned the inferior treatment.
Let me remind the reader of $L_E(\theta, a)$. One way to think of $L_E(\theta, a)$ is to take the perspective of a particular patient. If $H_0$ is true, whether the patient is given $T_1$ or $T_2$ the loss incurred with the treatment is 0 (since the treatments are equivalent), but when $H_1$ is true, if the patient is given the inferior treatment $T_1$ the loss is $d$ (-1); otherwise, if given $T_2$ (superior treatment) the loss is 0.

In our model we also make $L_E(\theta, a)$ sensitive to the “effect size”. By “effect size” I mean the reduction rate of TE between treatments. Thus, assuming $H_1$ is true, (50% reduction in TE from an original 0.3 rate) for every segment of interim analysis ($n=60$), the single unit ($u$) is the utility incurred from each of the expected 9 additional patients (from 18 to 9) having a reduction in TE, the result that researchers should (or could) have expected had they continued with the trial. Below is $L_E(\theta,a)$ with $d=-1$, $u=1$, $dd=-0.1$, and $cc=-0.01$.

<table>
<thead>
<tr>
<th>Action</th>
<th>True_state_of_nature</th>
<th>$H_0;\theta=0$</th>
<th>$H_1;\theta=0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>@interim($n=60$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_1$:choose $T_2$</td>
<td>$(n \cdot dd)+(N \cdot cc)=-16$</td>
<td>$n(d+dd) + \theta(N-n)u=404$</td>
<td></td>
</tr>
<tr>
<td>$a_2$:choose $T_1$</td>
<td>$(n \cdot dd)=-6$</td>
<td>$n(d+dd) + \theta(n)u=-36$</td>
<td></td>
</tr>
<tr>
<td>@final($n=120$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_1.f$:choose $T_2$</td>
<td>$(n \cdot dd)+(N \cdot cc)=-22$</td>
<td>$n(d+dd) + \theta(N-n)u=308$</td>
<td></td>
</tr>
<tr>
<td>$a_2.f$:choose $T_1$</td>
<td>$(n \cdot dd)=-12$</td>
<td>$n(d+dd) + \theta(n)u=-72$</td>
<td></td>
</tr>
</tbody>
</table>

$dd$ is the cost of having a patient switch treatments, after drug acceptance; while $cc$ is the loss for assigning a new drug that is non-superior to a patient.
We are now in a position to consider the DMC situation during interim analysis with CP=0.22. How does it compare with the PI's predictive power (0.40), given our loss function? Figure 4.5 shows cut-offs CPs for continuing trial as function of utility (u), given the study assumptions (80% power at a two-sided significance level of 0.05), interim data (Table 4.2), and $L_E(\theta, a)$.

**Figure 4.5** Cut-off CPs for continuing trial as a function of u, under original test assumptions, interim data, and $L_E(\theta, a)$. 

![Cut-off Conditional Power](image)
4.3.2. Discussion

Notice that with CP=0.22, the DMC decision to stop the trial has a lower expected loss than continuing only for utilities smaller than 0.66; otherwise, utilities greater than 0.66 would be required for continuing. Given our loss function, and assuming that the PI’s decision to stop was also based on whether ending it had a lower expected loss than continuing, then, with CP=0.40, the utilities that could have warranted the decision to stop are relatively small (u < 0.36). Otherwise the trial should have continued its course, since stopping at interim had higher expected loss than continuing.

The disagreement between the DMC and the principal investigator, given their respective CP values, can, therefore be explained for utilities between 0.36 and 0.66. For small utilities (between 0.2 and 0.36) both the DMC and the PI would have agreed on stopping the trial; they would also have agreed on continuing, had the utility values been high (u > 0.66).

Figure 4.5 also shows that if the utility increment incurred by each patient assigned the superior treatment is relatively small, then the minimum CP (cut-off) to warrant continuing is going to be very high for both the PI predictive power and the DMC CP. This minimum CP drops (on both approaches) as the utility increment becomes larger. The graph also shows that the frequentist minimum CP to continue is always slightly higher than the minimum for the PI CP. But overall, the difference between the two CP values is not much different, except when the utility increment is rather small. In these cases,
focusing our attention on methods, rather than decisions, it matters more which sort of CP monitoring rule one uses; much less so where there is a large utility increment.

The DTF analysis is helping to make clear the distinctive features of early stopping due to futility. While certain factors are best held fixed, e.g. CP calculations and expectations about future reversal, other factors that are relevant to interim decisions emerge as crucial. In this particular case, we saw how utility values—utility incurred from each of the expected additional patients having a positive recovery—help us explain the differences of judgments between the DMC and the PI decisions. The formal model helps us see how different approaches to the crucial parameters allow different choices of ‘best stopping decision’ to emerge.

In the next section, I bring the DTF approach to a different case: a case of early stop due to harm. It is another case in which we have a clash of opinions about stopping, this time between the DMC and a skeptic reviewer of the study. As in the cases of early stop due efficacy and early stop due to futility, we will see that both perspectives, i.e. DMC and skeptic, can also be rationalized and then compared. The upshot of the comparison is that the threshold of evidence that was demanded and adopted by the DMC is not permissible. The adopted stopping rule, once properly represented, is identified as being too lenient given the context in which we find vitamin A supplements.
4.4 Early stop due to harm

This was a study with 1078 HIV-infected pregnant women, and a sample size calculated with 90% power to detect a 30% efficacy on the risk of transmission, 0.05 type I error, and a background rate of transmission assumed as 30%.

In this RCT (introduced in section 2.4), the DMC recommended that the group given vitamin A be stopped, because vitamin A had a negative impact on one of the primary endpoints. Vitamin A increased the risk of HIV transmission from mother-to-child through breastfeeding. Of the children whose mothers did not receive vitamin A, 15.4% were infected by 6 weeks, compared with 21.2% of those whose mothers received vitamin A. (ibid 1938)

However, given that vitamin A supplements had proven beneficial to pregnant HIV-uninfected women in previous studies, by showing “significant reduction in maternal mortality among women from Nepal” (ibid 1942) the question we face is this: how did the DMC balance the possible short-term harmful effects of HIV transmission against the possible long-term beneficial effects of vitamin A in the face of uncertainties?

The dilemma faced by the DMC can be put this way: if the trial had continued beyond the point of the observed evidence, it would have placed children in the trial at unnecessary risk. On the other hand, it may be difficult to reverse the decision to abandon

\[ \text{RR}=1.38, \text{ 95\% CI 1.09-1.76}; \text{ p-value}=0.009 \]

\[ \text{“This may not be cost-effective to pursue unless such a counseling program is already in place as part of a strategy to reduce mother-to-child transmission.” (ibid 1942) } \]
vitamin A—as a matter of public policy—if later found to be an erroneous decision, hindering the adoption of vitamin A supplementation as a means of reducing mortality and morbidity rates among mothers suffering from HIV. Although it is difficult to explicitly model such potential losses with a high degree of plausibility, in this section, we attempt to understand the DMC’s rationale.\textsuperscript{121}

4.4.1. Simulation

\textit{Reconstructing the adopted stopping rule}

Let us begin with the reconstruction of the stopping rule. Because the stopping rule was neither explained nor mentioned in the report, we need to reconstruct it from certain facts found in the final report. We begin by noticing that “transmission through breastfeeding” (the primary event of interest) was defined as infection after 6 weeks of age among those who were not known to be infected at 6 weeks, and that “HIV diagnosis” (interim analysis) occurred at 3 months intervals thereafter. So it is reasonable to infer that:

(1) The DMC planned 8 interim analyses. That is because they “examined the effect of vitamin A on transmission during the period between 6 weeks and 24 months of age.”

\textsuperscript{121} The treatment of my early stop due to harm case can also be found in my (2012b) forthcoming paper “Modeling and Evaluating Statistical Decisions of Clinical Trials” in the \textit{Journal of Experimental & Theoretical Artificial Intelligence}. 

207
Presumably, the DMC did this at the same three-month intervals as the HIV diagnoses.

Two other relevant facts for the reconstruction of the stopping rule are:

(2) The DMC “recommended that the vitamin A arm of the study be dropped” because “a higher risk of HIV-1 transmission was observed in the vitamin A arm compared with the no vitamin A arm” with “p=0.009 (RR 1.38)”.\textsuperscript{122} (ibid 1938)

(3) “All p-values reported are two-sided; statistical significance in this study is defined as p <0.05.” (ibid 1937)

Because (1) tells us the number of interim analyses in the study, (2) the amount of evidence that was deemed sufficient to stop the trial, and (3) the nominal statistical significance in the study, it is reasonable to say that the adopted stopping rule was a Pocock stopping rule with 8 interim analyses planned and an overall $\alpha=0.05$. This is a reasonable deduction, because it would explain why the DMC found the evidence with p=0.009 at interim sufficient to stop the trial. That is, the reconstructed rule would only require a p-value of 0.01 as its cut-off for stopping at the pertinent interim, given the above assumptions.$^{123}$ (cf. Pocock 1977, table 1) A more conservative stopping rule—e.g. O’Brien-

\textsuperscript{122} 95% CI 1.09–1.76; “15.4% were infected by 6 weeks, compared with 21.2% of those whose mothers received vitamin A.” (ibid)

\textsuperscript{123} The Pocock rule as a spending function of the overall type I error is given by $\alpha(t) = \alpha \ln(1 + (e - 1)t)$ where $\alpha$ is the overall type I error, $\alpha(t)$ is the nominal type I error at t, and t is the information fraction of the trial. Note: I follow Proschan et al (2006, chapter 5) in their Brownian motion approach to the computation of spending functions.
Fleming—would not have allowed the DMC to stop early because of its extreme demands on early evidence in the trial.\textsuperscript{124}

Figure 4.6 compares our reconstructed stopping rule, the Pocock rule, with a more conservative stopping rule, an O'Brien-Fleming stopping rule [O-F], for illustration. The O-F stopping rule, by imposing more evidence early, namely, extreme statistical boundaries early in the trial, is a more demanding rule for early stopping than the Pocock rule. While the O-F rule is meant to resist stopping early due to expected high variability early on, the Pocock rule requires very similar—almost constant—levels of evidence for early and late interim analyses. Figure 4.6 compares the Pocock and O-F statistical boundaries vis-à-vis the cumulative probability of type I error at a particular interim point in the trial. The cumulative type I error is the probability of rejecting the null-hypothesis (when the null-hypothesis is true) at or before the pertinent interim—e.g. the cumulative type I error at 2\textsuperscript{nd} interim is 0.018 for the Pocock rule and 0.0001 for the O-F rule. While the Pocock rule increases the cumulative type I error at a constant rate, the O-F rule increases it very slowly early on the trial, followed by a speedy increase towards the end.

\textsuperscript{124} The O’Brien-Fleming rule as a spending function of the overall type I error is given by $\alpha(t) = \alpha 4(1 - \Phi(z_{\alpha/\delta}(t^{1/2})))$ where $\alpha$ is the overall type I error, $\alpha(t)$ is the nominal type I error at $t$, and $t$ is the information fraction of the trial. See Proschan et al (2006, chapter 5) as a method of computing spending functions.
Figure 4.6 Comparison of two candidate stopping rules—Pocock (blue diamonds) and O'Brien-Fleming (red squares)—based on a two-sided z-test of independent proportions, with 8 interim points, and an overall alpha of 0.05. The cumulative probability of type I error for each interim (as information fraction of the trial) is shown.

A second reconstruction of the stopping rule—the one I will concentrate my evaluation on later in the chapter—would be to keep fact (2), concerning the amount of evidence that was deemed sufficient to stop the trial, and fact (3), the nominal statistical significance in the study, but change assumption (1). That is, rather than model 8 interim analyses, we adopt a model of the stopping rule with 3 points only. That’s because the study reports the incidence of HIV mother-to-child transmission explicitly for 3 points only,
namely, at 6 weeks, 6 months and 24 months (end of the trial);\textsuperscript{125} the supposition being that data analyses were performed by the DMC \textit{at least} at these intervals. This is a reasonable conclusion, because it would explain why the DMC had found the evidence with $p=0.009$ at interim sufficient to stop the trial. The simplified reconstructed rule would only require a $p$-value of 0.022 as its cut-off for stopping at the pertinent interim, given the above assumptions. Figure 4. compares the Pocock and O-F rules supporting our reconstructed stopping rule.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Comparison of two candidate stopping rules—Pocock (blue diamonds) and O’Brien-Fleming (red squares)—based on a two-sided $z$-test of independent proportions, with 3 points, and an overall alpha of 0.05. The nominal statistical boundaries (statistical cut-off points) are shown on the left plot and the cumulative probability of type I error for each interim (as information fraction of the trial) is shown on the right plot.}
\end{figure}

\textsuperscript{125} “Of the children whose mothers did not receive vitamin A, 15.4% were infected by 6 weeks, compared with 21.2\% of those whose mothers received vitamin A. At 6 months, the cumulative incidences were 22.4 and 28.1\%, respectively, and at 24 months they were 33.8 and 42.4\%, respectively.” (ibid, 1938)
So far I have reconstructed the adopted stopping rule as a Pocock stopping rule, i.e. a non-conservative stopping rule. Holding fixed the reconstruction of the stopping rule (an implementation of the Pocock rule), in the next section (section 4.4.2) I consider factoring the interim results from the study into an analysis of whether stopping or continuing was permissible. Given that we have the incidence of HIV mother-to-child transmission explicitly for 3 points (and not for all 8 interim points as initially suggested in the first representation of stopping rule), I will now focus my evaluation on the reconstructed stopping rule with 3 points, thus leaving the reconstruction of stopping rule with 8 points without further analysis.

4.4.2. Evaluation

Consider a health policy advocate who knows that vitamin A supplementation in HIV+ pregnant women has been shown to reduce the incidence of adverse pregnancy outcomes, e.g. fetal loss and low birth weight. This decision analyst is someone who sees the results of Fawzi et al as corroborating (supporting) the public policy view of preventing the spread of HIV as paramount, and with the following consequence: either replace breastfeeding, given vitamin A supplementation, or stop vitamin A supplementation, given breastfeeding. This goes against a competing health policy view that he or she advocates: that promoting breastfeeding with vitamin A supplements is paramount—whether or not the mother is HIV+—due to its positive contributions to child nutrition, survival, and the mother’s wellbeing.
The decision analyst might want to know given observational studies showing the benefits of vitamin A, why the DMC decided to stop early and did not continue. Why did they think that the risks of continuing were higher than those of stopping, given the interim evidence? The authors of the study’s final recommendation was that "although it is difficult to base policy on the results of a single study, our data give little encouragement for the widespread use of vitamin A supplements in HIV+ pregnant women." (ibid)

This decision analyst would be skeptical about the DMC’s interim decision, because the final results (and recommendation) contradict previous observational studies that found that deficiency of vitamin A is associated with an increase of HIV transmission from mother-to-child.

Suppose the RCT was designed with a total of 539 participants per treatment, sufficient to detect a difference of 30% efficacy with \( \alpha=0.05 \) and power of 0.9\(^{126} \) according to reported assumptions about the trial.

While holding fixed the reconstructed stopping rule, the analyst computes what the difference in event rates would have to be, at the first interim point, i.e. one third of the way through the trial (\( n=180 \)), for the test to find evidence of \( p\)-value < 0.05 at the pre-specified \( \alpha=0.05 \). In other words she wonders how big an effect (i.e. how big of a difference in the mother-to-child transmission rates between groups) would have to be observed at the first interim in order to be statistically significant. She reasons, with 180 patients in each group,

\(^{126} \) These are based on a two-sided statistical test of proportions, i.e. the difference between two independent proportions (\( z\)-test).
that the DMC would have needed an efficacy of 41.6%\textsuperscript{127} for the two-sided test to produce evidence of p-value < 0.05 for α=0.05 and power=0.9, assuming symmetrical boundaries.\textsuperscript{128} She now wonders if such a difference between rates is likely to occur and whether it is reasonable to expect the same result, given the results of previous observational studies. Despite the methodological differences between RCTs and observational studies she compares the newly computed 41.6% value with results from observational trials, such as Greenberg et al 1997 which showed a much higher risk of mother-to-child HIV transmission among mothers who were vitamin A deficient compared to mothers who were not, i.e. 62.5% reduction rate.\textsuperscript{129}

With the comparison of effect sizes in mind, she finds that the criterion for early stopping at the first interim for reconstructed stopping rule (set at p=0.05) is “a lenient stopping criterion.” She reasons that to counter previous results in favor of vitamin A, a more stringent stopping criterion, namely, one that would have required more evidence vis-à-vis a smaller alpha (e.g. p=0.0001), is in order. Demanding a test result with p-value<0.0001 at first interim (instead of the lenient p-value<0.05 criterion) corresponds to mother-to-child HIV transmission rates with an efficacy of 55% against vitamin A.\textsuperscript{130} This efficacy value was neither seen at interim nor demanded by adopted stopping rule. She may now suspect—and reasonably so—that a decline in mother-to-child transmission rate this large is what should have been necessary to counter previous results in favor of

\begin{align*}
\text{Efficacy} = \frac{\text{(event rate in multivitamin group)} - \text{(event rate in vitamin A group)}}{\text{(event rate in multivitamin group)}} = \frac{.36 - .21}{.36} = .416
\end{align*}

\begin{align*}
\text{Based on a sensitivity power analysis for computing required effect size given } \alpha=0.05, \text{ power}=0.9 \text{ and n}=180 \text{ per group.}
\end{align*}

\begin{align*}
(0.16 - 0.06)/0.16 = 0.625
\end{align*}

\begin{align*}
\text{Based on a sensitivity power analysis for computing required effect size given } \alpha=0.001, \text{ power}=0.9 \text{ and n}=180 \text{ per group.}
\end{align*}
vitamin A. She then argues that the adopted stopping rule is unethical, since for all practical purposes it is the same as if the DMC had adopted a relatively easy stopping rule. She argues that given the conflicting results of vitamin A, a more appropriate stopping rule is a rule that reflects the view that “only the most extreme of evidence should stop the trial early”, such as a rule that demands p-value<0.0001 at interim (akin to O’Brien-Fleming rule, see Figure 2). She then concludes that the degree of conservatism (i.e. degree of evidence) that is found and demanded by the adopted (modeled) stopping rule, for a case of early stopping due to harm, is not permissible, because, once again, it is too lenient given the context in which we find vitamin A supplements.

4.4.2.1. Comparison of stopping rules using a loss function

Suppose the decision analyst compares her stopping rule to the reconstructed DMC stopping rule in the following manner. She makes \( L(\theta, a) \) sensitive to “efficacy” in the following way. The efficacy demanded by her stopping rule is that of 55%, whereas the efficacy demanded by the reconstructed stopping rule is that of 41.6% (according to her calculation above), both rules implemented with symmetrical boundaries. According to either stopping rule, if \( H_0 \) is true, efficacy is 0.

Let \( cc \) denote the cost of running the trial up to a given point, that is, at first interim \( cc = -10K \). At the end of the trial, the cost of running the trial is 3 times this value (i.e. -30K). Let \( p \) denote the probability that \( H_1 \) is true. Fixing attention to one stopping rule at a time,
she wants to compare (*a priori*, without any collected interim data) the following two possibilities: stopping early and declaring $H_1$ is true, and continuing the trial and declaring $H_1$ is true. Suppose a shorter trial (a trial stopped early) is not as convincing as a longer trial (one that goes to full completion), which is also reflected in the table below.\(^{131}\) She assumes the patient horizon $N = 1000$.

<table>
<thead>
<tr>
<th>Action</th>
<th>True_state_of_nature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$</td>
</tr>
<tr>
<td></td>
<td>(efficacy=0)</td>
</tr>
<tr>
<td>@interim(n=180)</td>
<td>NA</td>
</tr>
<tr>
<td>Stop early: choose $T_1$</td>
<td>$cc$</td>
</tr>
<tr>
<td>Stop early: choose $T_2$</td>
<td>$3 * cc$</td>
</tr>
<tr>
<td>@final(n=540)</td>
<td>NA</td>
</tr>
<tr>
<td>At the end: choose $T_1$</td>
<td>$cc$</td>
</tr>
<tr>
<td>At the end: choose $T_2$</td>
<td>$3 * cc$</td>
</tr>
</tbody>
</table>

We are now in a position to compare the DMC stopping rule versus the skeptic analyst stopping rule. How do the rules compare, given our loss function? The expected pay-off, $E(\delta, \theta)$, associated with decision $\delta = \{'stop early', 'at the end'\}$ and outcome $\theta = \{0, \text{efficacy}\}$ is calculated below.

\(^{131}\) The assumption here is that a trial that goes to completion is capable of changing clinical practice, whereas a shorter one is not, thus the difference in the loss function above.
4.4.3. Discussion

For the DMC stopping rule:

\[ E(\delta = \text{‘stop early’, } \theta = 0) = cc = -10000 \]
\[ E(\delta = \text{‘stop early’, } \theta = 41.6) = \theta \times (n) + cc = -2512 \]
\[ E(\delta = \text{‘at the end’, } \theta = 0) = 3 \times cc = -30000 \]
\[ E(\delta = \text{‘at the end’, } \theta = 41.6) = \theta \times (N) + cc = 11600 \]

It follows that the expected pay-off associated with decision ‘at the end’ \([11600p - 30000(1 - p)]\) is greater than the expected pay-off associated with decision ‘stop early’ \([-2512p - 10000(1 - p)]\) if and only if: \(p > 0.59\). That is, by following the maximization of expected utility, \textit{without any interim data}, the DMC would continue its trial when it believes that the probability that \(H_1\) is true, is greater than 0.59, otherwise the DMC would instead stop its trial early. This comparison is plotted in Figure 4.8 (DMC, graph in the left).

For the skeptic stopping rule:

\[ E(\delta = \text{‘stop early’, } \theta = 0) = cc = -10000 \]
\[ E(\delta = \text{‘stop early’, } \theta = 55) = \theta \times (n) + cc = -100 \]
\[ E(\delta = \text{‘at the end’, } \theta = 0) = 3 \times cc = -30000 \]
\[ E(\delta = \text{‘at the end’, } \theta = 55) = \theta \times (N) + cc = 25000 \]
It follows that the expected pay-off associated with decision ‘at the end’ \([25000p - 30000(1 - p)]\) is greater than the expected pay-off associated with decision ‘stop early’ \([-100p -10000(1 - p)]\) if and only if: \(p > 0.44\). That is, by following the maximization of expected utility, a priori, the skeptic would continue its trial when it believes that the probability that \(H_1\) is true is greater than 0.44, otherwise the skeptic would stop its trial early. This comparison is plotted in Figure 4.8 (DMC, graph in the right).

Notice that with this type of comparison, the decision analyst is required to assess the probability of an uncertain event even when she has no interim data at her disposal. If, however, we assume that subjective probability assignments can be made, a priori (i.e. before any interim data), such assignments do not need to be precise. In this particular comparison, all the decision analyst needs, once the DMC stopping rule has been reconstructed, is to assess whether her subjective probability value of \(p\), is less or greater than 0.59 for the efficacy demanded by the reconstructed DMC stopping rule.
Figure 4.8 Expected pay-offs associated with the decisions of ‘stopping early’ and ‘going until the end’ as a function of probability that $H_1$ is true, according to the DMC stopping rule (graph on the left), and the Skeptic stopping rule (graph on the right).

By following the maximization of expected utility decision criteria we see that the skeptic’s optimal decision rule is to stop the trial if she believes that the probability that $H_1$ is true is less than 0.44 and to continue the trial until completion otherwise. This is contrasted with the DMC’s optimal decision rule which says to stop the trial if the probability that $H_1$ is true is less than 0.59 and to continue the trial otherwise.
Chapter 5. Toward a pluralistic framework of early stopping decisions in RCTs

In this chapter it is argued—in light of previous discussions—that different families of statistical inference have different strengths and that they can coexist in serving the practical demands for designing and monitoring clinical trials. Section 5.1 provides a brief summary of my approach. Section 5.2 identifies some objections to the proposed framework. Section 5.3 identifies some achievements with the framework, including contributions to philosophy of medicine, and to philosophy of statistical inference. Section 5.4 identifies areas for further research.
5.1 Summary

In the preceding chapters I have set out the main elements of a systematic framework for modeling and evaluating early stopping decisions of RCTs. The framework utilizes and builds upon a plurality of statistical methods, loss functions and decision criteria for making sense of the different ways in which early stopping decisions of RCTs are made. But it does so in ways that depart from what is often associated with the two main competing approaches to statistical evidence and inference, namely, the schools of orthodox Bayesian and Error Statistics (or ES).

The two competing positions illustrated by their corresponding philosophies of science (Bayesianism and ES) have been insufficient in understanding what takes place during the design, monitoring and early stopping of RCTs. Because the first position supposes that stopping rules do not matter for an account of scientific evidence and inference, its traditional response has been mostly one of indifference to what takes place epistemically during the monitoring of RCTs. As a result of this indifference, orthodox Bayesians have separated themselves from having a meaningful discussion of the very instruments and strategies that practitioners have come to rely upon when designing and conducting RCTs.

The second position—illustrated by ES—sees the role of stopping rules as an essential part of an account of scientific evidence and inference. According to this position, without stopping rules, we lose the proper means for interpreting and evaluating experimental evidence. By making evidence contingent on stopping rules, ES requires
experiments to follow their pre-scheduled termination. This requirement, in turn, has the cost of not being able to accommodate the early stopping of experiments, at times in violation of the original stopping rule, while yet providing a useful or meaningful notion of scientific evidence.

What fundamentally distinguishes my framework from other ways of scrutinizing the early stopping of RCTs is that in order to determine whether a given early stopping decision is permissible, it is necessary to take into account not only the stopping rule upon which the experiment was originally based, but also the rationale for violating the stopping rule, given the context of the early stopping decision.

In referring to the representation and evaluations of early stopping decisions, I have in mind the specification of the rationale for both, the adoption of a designed stopping rule, and at appropriate times, the rationale for violating the stopping rule. Currently, the predicament one faces in the philosophy of statistics is that of having to adjudicate between competing families of statistical evidence and inference. My contention is that the proposed decision framework, when applied to the diverse contexts of RCTs, presents a viable alternative to a single and universal account of statistical evidence and inference. No universal adjudication between the competing accounts—Bayesianism vs. ES—is appropriate for addressing the needs of RCTs.
What I hope to have shown with this dissertation is that the two positions (Bayesianism and ES) by themselves, share the same unfortunate result in the context of RCTs: neither provides any basis for determining the permissibility of interim monitoring decisions of RCTs. What I have argued is that we need a new way of thinking of the role of stopping rules during the planning, conduct and monitoring of RCTs. By seeing RCTs in their particular contexts, with their particular ethical and epistemic constraints, we can explain the importance of both statistical families for a critical scrutiny of early stopping decisions. By using both families of statistical evidence and inference, we can make sense not only of the basic importance of stopping rules for designing experiments, but also of the proper grounds for violating stopping rules when appropriate, including the rationalization of such violations.

Because within the science of clinical research, we find ourselves with a plurality of statistical methods, it is important that a proper framework for specifying and evaluating early stopping decisions of RCTs take this pluralism seriously into account. To repeat myself, my interest has not been proposing yet another universal account of statistical evidence per se, but in the important noticeable variation in the ways in which practitioners justify the early stopping of their trials. That is, in pressing a systematic form of justification to which practitioners either explicitly or implicitly subscribe, we, lay experts, may have the means to regard whether a particular interim decision is either permissible or not permissible, according to a given representation of the decision. Because it is not surprising that people often disagree on what is of greatest value or harm for them and for others, it is reasonable to seek an account of early stopping decisions that saves disagreements, without having to necessarily solve them once and for all. Moreover, when
specifying a rationale for an early stopping decision, in a particular context, although different contexts may allow for comparisons between different interim decisions, judgments of superiority cannot be justified as always universally optimal—thus the need for a notion of weak optimality, where one decision looks optimal in one representation—one set of dimensions—but not so on another.

Because early stopping of RCTs is not self-justifying, there is an important sense in which these decisions must be able to meet legitimate challenges from trial participants and skeptical reviewers. On the one hand, one may argue that this ‘democratization’ of early stopping decisions may prove itself as an ‘unrealistic’ goal. After all, it might be suggested that having to put researchers in a position of opposition, that is, of fearing to lose their freedom due to lay interference, is too high of a demand to make, thus an unrealistic goal. If that is right, one question we might ask is whether we can still reconcile the public participation of skeptical reviewers of early stopping decisions with the autonomy of researchers and DMCs.

On the other hand, because RCTs carry an inherent tension between generating new knowledge and safeguarding the interests of trial participants, it is, perhaps, not too far-fetched to think that skeptical reviewers do share (to a large extent) the ethical and epistemic goals that DMCs have for their trial participants, thus suggesting the possibility of compromises on the ways early stopping decisions are justified. This compromise may require a translation of trial participants’ demands into notions of appropriate justification, i.e. permissible early stopping decisions, in such a way that the demands can be satisfied during the design and conduct of RCTs.
HIV/AIDS activists of the 1980s and 1990s were drawn deeply into the design and conduct of RCTs. As lay experts, activists were able to struggle to work out important participation in the very rationalization of the experiments their lives had come to depend on. If we think that that historical episode can be seen as an illustration of the possibility of further democratization of RCTs, then the emergence of a new technical demand for RCTs, namely, a public rationale for the early stopping of RCTs, may be likely.

Even though clinical research expertise has historically served mostly those in control of experiments (e.g. pharmaceutical companies, federal institutions) because the early stopping of RCTs lies at the important intersection of trial participant interests, skeptical reviewers of such experiments, and DMCs, this new technical demand, may be expected, although easier said than done. But one must start from somewhere. Examining the different types of early stopping decisions, and the difficulty of explaining their complexity and the democratic importance of their rationale, while trying to systematize their representation, is as good as any place to start such an important endeavor.

By focusing on the classification, identification, representation and evaluation of the rationale for early stopping, I have also examined how the epistemic and ethical aspects of data monitoring are typically intertwined. I have shown how a proper evaluation of interim statistical decisions of RCTs requires ways of modeling DMC decisions as answers to dilemmas. This approach, in turn, helped us see how the focus of a DMC’s dilemma—whether to stop or continue—can shift from being primarily an ethical issue to an epistemic issue (and vice versa). The upshot of modeling such dilemmas is that epistemic
aspects of RCTs are best seen as not standing apart from ethical concerns. Rather, when serving a public rationale for early stopping, the epistemic and ethical aspects of RCTs are best seen if modeled as balancing against one another.

Once this point about the balancing of epistemic and ethical factors has been appreciated, the following step is to consider what an early stopping principle of RCT might look like. If principles of early stopping are to help guide DMC’s behavior, these principles must be represented, and they must be capable of being taught so that they can serve as a public rationale for evaluating early stopping decisions. Any scientific approach whose principles of evidence and inference imply that it is not permissible to teach stopping rules, or early stopping principles, violates this publicity standard—i.e. it does not meet the public rationale for the early stopping of RCTs. In this fashion, it seems reasonable to say that in contrast to the two positions illustrated by Bayesianism and ES, the proposed framework seems capable of meeting this publicity standard.

### 5.2 Some objections

_Could this all be done as a Bayesian? What about as an error-statistician?_

The simple straightforward answer to both questions is no, and no. It cannot be done as a Bayesian, and it cannot be done as an error-statistician. As the answer is not so obvious, I will elaborate.
Let me begin with a Bayesian approach. If by a ‘Bayesian’ approach, we mean the simple approach of updating probabilities derived from a prior distribution for monitoring a trial, without any regards to a formal pre-specification of a stopping criterion, sample size, or sampling properties, then this Bayesian approach to RCTs will not be able to do what the proposed framework does, i.e., be able to properly represent and evaluate the diverse methods of RCT design and monitoring we find in practice.

There are two reasons for this. First, if by a Bayesian approach we mean the view subscribing to the likelihood principle—accepting the thesis of the irrelevancy of stopping rules to the interpretation of evidence—then, the approach fails to account for the plurality of methods we find RCTs with; including methods that are at odds with the likelihood principle. Indeed, the same point applies to any rule that promotes experimental aims that make stopping rules irrelevant. Thus, if by Bayesianism we mean having to accept an account of evidence and inference that is strictly Bayesian (in this narrow sense of accepting the implications of the likelihood principle), then Bayesianism loses grounds for judging methodological rules in RCTs that reflect different principles for interpreting evidence (e.g. error probabilities, predictions of the possible consequences of continuing, the need for a fixed sample size).

Second, once standard Bayesianism is supplemented with methods such as those for making predictions of the possible consequences of continuing or stopping, or having to assume, rather explicitly, the possible losses associated with consequences of stopping or continuing, then ‘Bayesianism’ will start to move away from standard Bayesianism, and start to look more like a hybrid ES-Bayesian approach. Despite the simplicity of strict
Bayesianism, a reduction of RCT monitoring to a dependence on prior probabilities and updating of posterior distributions falls short of accounting for the diversity of underlying rationale of pervasive methodological rules in RCTs. There is no direct implication of strict Bayesianism on rationales such as sampling properties, sample size calculation, and loss functions (rather implicit or explicit ones).

With regards to error-statistics, it will not do the job either. One reason is that an error-statistics approach to RCT will need to take into account the differences of opinions that we often find among DMC members. The decision-theoretic framework can account for such differences by using either different utilities or different utility functions altogether when representing differences of opinion. Moreover, an error-statistics approach to RCT will also need to account for the multiple sources of evidence when making an assessment of the interim evidence. The proposed framework allows for multiple sources of evidence whereas error-statistics does not.

In section 1.6 I have also raised two problems for error-statistics when it comes to explaining and justifying early stopping of RCT experiments. The first was error-statistics’ silence about the need for balancing between early and late-term effects in RCTs. There we saw how error-statistics is an incomplete guide to early stopping, and therefore the need to supplement it. The current framework makes sense of the method of conditional power functions as a necessary add-on to error-statistics. The second issue was how to prospectively control error rates in experiments requiring multiple interim analyses. Again, this suggests that error statistics is an incomplete account of the justification and early stopping of RCTs.
Where do you get the parameters in your model? If it is not clear, do we really have a helpful framework?

The parameters come from somewhat idealized circumstances of RCTs. The type of RCT and its context should inform the decision model. But the particular adoption of a simplified model (as seen in Chapter 4) that includes elements such as \( n \) pairs of patients equally randomized into two treatments, a patient horizon of size \( N \) (with \( N >> n \)), and a loss function that is proportional to both, the number of patients given the inferior treatment and to the “effect size” \(|\theta_2 - \theta_1|\), comes from my own adaptations to idealized models previously considered by epidemiologists and biostatisticians—(e.g. Anscombe 1963, Berry and Pearson 1985, Heitjan, Houts, and Harvey 1992).

Looking at the literatures of sequential analyses and RCT monitoring, for instance, it is not clear what general principles can be used for setting up potential decision models. Some might see this aspect of the framework as somewhat disturbing, since without a clear set of principles for selecting parameters to the model, it is not clear how one might approach representation of early stopping decisions, especially representations that must go beyond the idealized models used in chapter 4.

Even though this might prove a highly controversial aspect of the framework, my experience with the framework has been to approach the selection of parameters as if one were selecting expedient devices, that is, convenient ‘instruments’ to be replaced when necessary. Since it is not always clear when the representation of a particular early
stopping decision will work—i.e. capable of rationalization in any meaningful way—my recommendation is to start with a fairly idealized model first, before building in more realistic and complex decision models.

What value is added, over and above good judgment?

The framework allows the decision analyst to generate and quickly answer “what-if” questions (i.e., alternate scenarios). By being able to compare alternative scenarios the framework permits the decision analyst to provide reasons for the DMC decision. If what-if-things-had-been-different information is intimately connected to judgments about explanatory relevance, then the framework is a useful tool towards the understanding of DMC interim decisions.

With a particular representation of the DMC decision, the decision analyst gains understanding of the DMC’s action in the ability to reveal a pattern of counterfactual dependence associated with relationships that are potentially exploitable for purposes of manipulation. For instance, by holding fixed a subset of the parameters in the model (e.g. loss function, and a reconstruction of the stopping rule) the decision analyst might come to understand how changes in the predictive probability of the test statistic will change the DMC decision to stop the trial due to futility. Grasping such relationships is the essence of what is required for the decision analyst to be able to explain DMC decisions.
5.3 Some achievements

5.3.1. Contribution to philosophy of medicine

The proposed framework makes new contributions to the philosophy of medicine, most specifically, to the ethics of RCT design. Below I briefly explain two of these contributions: a better understanding of the methodological basis of RCTs, and a new need to revisit the current institution of informed consent.

Among the aspects of study design currently discussed in philosophy of medicine is the methodological basis of RCTs. While trying to “situate the challenges associated with RCTs within a systematic framework for ethical research,” (cf. Joffe and Truog 2010) current authors contend that “a rigorous methodology that helps maximize scientific validity” is needed. (2010 245) The problem, however, is that current philosophical work does not go deep enough in explaining the inherent complexities of having to balance epistemic and ethical constraints in RCT. Conflating several epistemological issues that are rarely, if ever, examined from the point of view of the ethics of clinical research—but that nonetheless have ethical implications for the study design—current philosophies of medicine can benefit from a careful analysis of the ethical issues raised by Neyman-Pearson testing. This is an area that the framework can make important contributions to.
Having a framework that can separate (formalize) essential elements of RCT design and monitoring (i.e., researcher’s expectation, explicit balance of type I and type II error, costs associated to each action at interim, rationing of type I error) can open new ways of seeing trial participants as potential contributors to trial design and monitoring. This particular point about trial participants contributing to trial design and monitoring is also relevant to my second contribution to philosophy of medicine, i.e. the literature of informed consent.

To my knowledge, the current institution of informed consent does nothing to accommodate the preferences of trial participants in the epistemological aspects with the sorts of ethical implications mentioned above. During the conduct and monitoring of a trial, researchers often learn new information that was not foreseeable during the design phase of the study (e.g., low event rate, unforeseen side effects), and this creates both an ethical and scientific need to revise the initial design of the study. This revision in turn should require a revisiting of the initial informed consent, reflecting the changes in the study, and consequently, requiring further input from trial participants. A well designed informed consent would reflect this need for an ongoing revision of study considerations. The current institution of informed consent does not. One reason for this failure in informed consent is that researchers are taught that “only toward the end of preparing a research project are the issues of informed consent relevant,” (2010, 5)\textsuperscript{132} which implies that by the time participants are asked to sign their informed consent, it has been already established that participants shall have no say in the design of the very experiments they are invited to

participate in. Having a framework that helps make the justification of early stopping decision public can shed light on the sorts of requirements that a well-designed informed consent policy may have to meet.

5.3.2. Contribution to philosophy of statistics

The principal contribution of the current work to philosophy of statistics is that it adds support to the view that statistics should not be governed by a single ideal of validity; rather, a pluralistic view of statistical evidence and inference is more appropriate for explaining and justifying early stopping of RCTs. This issue of pluralism is an important and live issue in the philosophy of statistics. According to the proposed framework, each of the competing statistical approaches is well-suited to address certain epidemiologic demands, but inappropriate as a universal strategy for assessing the efficacy and safety of medical intervention measures. This poses a challenge to philosophies of science advocating a single and unified principle of evidential relation for statistical reasoning.

The topic of early stopping of RCTs is also an important topic for philosophers of science interested in explaining and justifying what takes place in the conduct of experiments. So, in this respect, the current work makes contributions to the philosophy of scientific experiments by a doing careful examination of the complexities involved in the design and monitoring of RCTs.
My approach with respect to philosophy of statistics is also original in the sense that it brings in decision theory to resolve the controversy over stopping rules (applied to the field of medicine). The novelty of such an approach lies in its pragmatic orientation toward the representation and evaluation of statistical methodologies and their stopping rules that we find in the practice of RCTs. This contrasts sharply with most philosophical work, which has centered on theoretical issues.

5.4 Future projects

There are four interrelated areas to which one might want to expand the research here presented in the near future. They are: [A] the socio-politics of RCT reporting; [B] formal epistemology and the representation of monitoring decisions; [C] the permissibility of early stopping decisions; and [D] bioethics and early stopping for RCT. Let me explain each.

[A]: Despite efforts from regulatory agencies, recent systematic reviews of RCTs show that top medical journals continue to publish trials without requiring authors to report details for readers to evaluate early stopping decisions carefully. Yet differences in judgments about RCT monitoring policies are important in understanding early stopping decisions; indeed, statistical monitoring plans are mandated by official policy documents. But this mandate is rarely followed. Even when a DMC does follow the mandate, there are often disputes about the appropriate statistical monitoring policy. In my dissertation research, I provided preliminary examination of this problem of reporting on cases of early
stopping, but more research is needed on the way data monitoring committees deliberate, and on the way that top medical journals review and accept papers based on partial reporting.

[B] & [C]: Because a clinical trial is a public experiment affecting human subjects, a clear rationale is required for the decisions of the DMC associated with the trial. Early stopping decisions are not self-justifying. They must be able to meet legitimate challenges from skeptical reviewers. Both the formulation of challenges and the task of responding to challenges would be assisted if we had a systematic, formal (or quasi-formal) way of representing early stopping decisions. My dissertation introduces some useful apparatus to this end. Further work is needed in this area, formalizing the relation between the a priori (designated) stopping rule and the early stopping principle that is applied during the course of the experiment.

[D]: What do trial participants think about early stopping decisions of RCTs? For instance, do asymptomatic HIV/AIDS participants have an opinion about how to balance early evidence for efficacy against the possibility of late side effects? Because DMCs have responsibilities to trial participants, yet those participants have no vote in early stopping decisions, it seems that an important voice is missing. In my dissertation I propose a form of representing early stopping decisions that considers the expected losses of taking various actions at interim. At the very least, it seems that the opinions of trial participants should be taken into account in assessing these expected losses. Yet I am not aware of any systematic study of trial participants’ views about such matters.
Final Remarks

In this dissertation, I hope to have advanced our understanding of the role of stopping rules in RCTs and how practitioners make early stopping decisions. The complexities of monitoring and responding to interim data while upholding the two essential obligations researchers face—learning about the world and safeguarding the interest of trial participants—were examined. This is a rich and ongoing domain of knowledge, however, and there is much more work to do in order to probe questions like those raised above.
Bibliography


Ellenberg, et al. (2003), *Data Monitoring Committees in Clinical Trials*, John Wiley & Sons, LTD.


Appendix

Appendix A: Likelihood Principle Example

Suppose \( \theta \) is the probability of successfully converting a free-throw basketball.

Let us first consider \( e_1 \) which describes an experiment designed to have a fixed number of trials (free-throws) \( n = 190 \), with \( r = 102 \) successful conversions (modeled as a binomial).

According to Bayes’s Theorem:

\[
P(\theta | e) = \frac{P(e | \theta)P(\theta)}{P(e)} \tag{1}
\]

In the present instance, \( e = e_1 \).

\[
P(e_1 | \theta) = C(n,r)\theta^r (1 - \theta)^{n-r} \tag{2}
\]

Where \( n \) (is fixed) = 190 and \( r = 102 \) (the number of successes)

\( C(n,r) \) is a binomial factor: a function of \( n \) and \( r \) only.

\[
P(e_1) = \int C(n,r)\theta^r (1 - \theta)^{n-r} P(\theta)d(\theta) \tag{3}
\]

Substituting (2) and (3) in (1), we have

\[
P(\theta | e) = \frac{C(n,r)\theta^r (1 - \theta)^{n-r} P(\theta)}{\int C(n,r)\theta^r (1 - \theta)^{n-r} P(\theta)d(\theta)}
\]

And since \( C(n,r) \) is independent of \( \theta \), we can move it out of the integral in the denominator, thus canceling it out from top and bottom in Bayes’s Theorem.

\[
P(\theta | e) = \frac{C(n,r)\theta^r (1 - \theta)^{n-r} P(\theta)}{\int \theta^r (1 - \theta)^{n-r} P(\theta)d(\theta)}
\]

Thus, \( P(\theta | e_1) = \theta^r (1 - \theta)^{n-r} P(\theta) / \int \theta^r (1 - \theta)^{n-r} P(\theta)d(\theta) \) \tag{4}

Now, let’s consider, \( e = e_2 \).

\[
P(e_2 | \theta) = C(n-1,r-1)\theta^r (1 - \theta)^{n-r} \tag{5}
\]

Where \( n \) (is variable) = 190 and \( r = 102 \) (the number of successes)
\( C(n-1,r-1) \) is a negative binomial factor: a function of \( n \) and \( r \) only.

\[
P(e_1) = \int C(n-1,r-1) \theta^r (1 - \theta)^{n-r} P(\theta) d(\theta) \quad (6)
\]

Substituting (5) and (6) in (1), we have

\[
P(\theta | e) = \frac{C(n-1,r-1) \theta^r (1 - \theta)^{n-r} P(\theta)}{\int C(n-1,r-1) \theta^r (1 - \theta)^{n-r} P(\theta) d(\theta)}
\]

Similarly to the binomial case above, since \( C(n-1,r-1) \) is independent of \( \theta \), we can move it out of the integral in the denominator, thus canceling it out.

\[
P(\theta | e) = \frac{C(n-1,r-1) \theta^r (1 - \theta)^{n-r} P(\theta)}{\int \theta^r (1 - \theta)^{n-r} P(\theta) d(\theta)}
\]

Thus, \( P(\theta | e_2) = \frac{\theta^r (1 - \theta)^{n-r} P(\theta)}{\int \theta^r (1 - \theta)^{n-r} P(\theta) d(\theta)} \) (7)

Comparing (4) and (7), we notice that they are equal, i.e., \( P(\theta | e_1) = P(\theta | e_2) \).

Moreover, any other description of the trial for which \( P(e | \theta) = K \theta^r (1 - \theta)^{n-r} \), where \( K \) is independent of \( \theta \), yields the same posterior distribution for Bayesians.
Appendix B: Additional Graphs and Tables

<table>
<thead>
<tr>
<th>n</th>
<th>10</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>teta</td>
<td>0.5</td>
<td>0.001</td>
<td>0.010</td>
<td>0.044</td>
<td>0.117</td>
<td>0.205</td>
<td>0.246</td>
<td>0.205</td>
<td>0.117</td>
<td>0.044</td>
<td>0.010</td>
<td>0.001</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.01</td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.044</td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.002</td>
<td>0.005</td>
<td>0.009</td>
<td>0.011</td>
<td>0.009</td>
<td>0.005</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.117</td>
<td></td>
<td>0.000</td>
<td>0.001</td>
<td>0.005</td>
<td>0.014</td>
<td>0.024</td>
<td>0.029</td>
<td>0.024</td>
<td>0.014</td>
<td>0.005</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>0.205</td>
<td></td>
<td>0.000</td>
<td>0.002</td>
<td>0.009</td>
<td>0.024</td>
<td>0.042</td>
<td>0.050</td>
<td>0.042</td>
<td>0.024</td>
<td>0.009</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>0.246</td>
<td></td>
<td>0.000</td>
<td>0.002</td>
<td>0.011</td>
<td>0.029</td>
<td>0.050</td>
<td>0.061</td>
<td>0.050</td>
<td>0.029</td>
<td>0.011</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>0.205</td>
<td></td>
<td>0.000</td>
<td>0.002</td>
<td>0.009</td>
<td>0.024</td>
<td>0.042</td>
<td>0.050</td>
<td>0.042</td>
<td>0.024</td>
<td>0.009</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>0.117</td>
<td></td>
<td>0.000</td>
<td>0.001</td>
<td>0.005</td>
<td>0.014</td>
<td>0.024</td>
<td>0.029</td>
<td>0.024</td>
<td>0.014</td>
<td>0.005</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>0.044</td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.002</td>
<td>0.005</td>
<td>0.009</td>
<td>0.011</td>
<td>0.009</td>
<td>0.005</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.010</td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.001</td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.000</td>
<td>0.0098</td>
<td>0.0439</td>
<td>0.1172</td>
<td>0.2051</td>
<td>0.2461</td>
<td>0.2051</td>
<td>0.1172</td>
<td>0.0439</td>
<td>0.0098</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

**Table B.1** Overall type I error 0.058 (in red) computed according to values for \( p((y_2 - y_1) | H_0) \) at the end of the trial, given \( n = 10 \).

**Figure B.1** Example of a frequentist stopping rule and its error probabilities.
Figure B.2 Example of three different families of stopping rules implemented and plotted.

Figure B.3 Example of stopping rules and two ways of computing conditional power.
Figure B.4 Example of an early trend of efficacy and an implemented stopping rule.

Figure B.5 Example of an early trend of harm and an implemented stopping rule.