Designing Randomized Clinical Trials for Rare Diseases

by

Lusine Abrahamyan

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Health Policy, Management and Evaluation
University of Toronto

© Copyright by Lusine Abrahamyan 2010
Designing Randomized Clinical Trials for Rare Diseases

Doctor of Philosophy

2010

Lusine Abrahanyan

Department of Health Policy, Management and Evaluation

University of Toronto

Abstract

Objectives: 1) To evaluate the quality of randomized clinical trials (RCTs) in rare diseases using Juvenile Idiopathic Arthritis (JIA) as an example, 2) to evaluate the time to treatment response in patients with rheumatic diseases, 3) to evaluate the power of the Randomized Placebo-Phase Design (RPPD) under various response time distributions, and 4) to examine the use of Value of Information (VOI) methodology in the optimal design of clinical trials for rare disease using hemophilia prophylaxis with factor VIII as an example.

Methods. The methods include a systematic review, a secondary analysis of data from an RCT and from a patient registry, a computer simulation study, and an evaluation of hypothetical RCT scenarios with VOI methodology.

Results. The quality of RCTs in JIA based on selected quality indicators was poor with some positive changes over time. In the data sets used for the assessment of hazard distributions, the response times followed mostly generalized gamma or lognormal distributions. The impact of time-to-event distribution on the power of RCTs was assessed in computer simulations. Based on the simulation results, the highest sample sizes were observed for response times following
the exponential distribution. In most scenarios, the parallel groups RCT design had higher power than the RPPD. The conclusion of the VOI analyses indicated that at threshold values lower than $400,000 the current evidence supported the use of on-demand therapy. Threshold values higher than $1,000,000 supported the use of tailored or alternate day prophylaxis. At threshold values between $400,000 - $1,000,000 the optimal decision varied from on-demand to prophylaxis therapies.

**Conclusions.** New, more powerful and acceptable designs should be developed for rare diseases. When time-to-event outcomes are used, investigators should use various sources of information to evaluate response time distributions before the new trial is designed, and consider this information in sample size calculation and analysis. VOI methodology should be used in the planning stage of studies to determine the relevant costs and benefits of future research, and to determine the optimal trial parameters that maximize the cost-benefit trade-off.
Acknowledgments

I would like to dedicate this work to my parents, Edmon and Bavakan Abrahamyan. All my life I have been blessed to have their love, support, and guidance. I want to thank my son Rafael for his understanding and patience. This work could not have been done without the support, trust, and love of my husband Vahan Muradkhanyan.

It is with deep gratitude that I would like to thank my supervisor Dr. Brian Feldman for his superb mentorship during these last five years – starting from the first day I arrived in Canada. His confidence in my abilities, his tremendous support, and his positive approach to all difficulties that one can face in life played a significant role in my life and development as an individual and as a professional. He was the one who taught me that science is not about finding answers but asking right questions.

I am very thankful to Dr. Andy Willan who was always prompt with his advice and help, always kind and supportive. Thanks to him, I learned about the applications of the Value of Information Method – a handy tool for decision makers. I would like to thank Dr. Joseph Beyene for his support in exploring survival distributions and applying various statistical techniques to my work. I am thankful to Dr. Edward Keystone for his interest and support during my research project.

There are many people that influenced my life by setting personal examples and by giving me professional advice; my deepest gratitude to Drs. Jan Barnsley, Ahmed Bayoumi, Marie Diener-West, Anahit Demirchyan, Murray Krahn, Prakesh Shah, and Michael Thompson. I have been blessed to work and study in a very nurturing environment of the CHES program. I am grateful to my colleagues and friends here for their support and understanding - Audrey Abad, Sindhu Johnson, Silvia Li, Marjorie McLimont, and Samantha Stephens.

I want to thank the Canadian Arthritis Network, The Research Institute of The Hospital for Sick Children, and the Department of Health Policy Management and Evaluation for providing me with financial support during my study years.
# Table of Contents

Abstract ......................................................................................................................................... ii
Acknowledgments ........................................................................................................................ iv
List of Tables ................................................................................................................................... vi
List of Figures ................................................................................................................................... viii
List of Appendices ....................................................................................................................... ix
Thesis Overview ........................................................................................................................... x

Chapter 1. Introduction .................................................................................................................. 1
  1.1 Epidemiology of rare diseases .......................................................................................... 1
  1.2 Special policies for rare diseases .................................................................................... 1
  1.3 Difficulties in conducting randomized clinical trials in rare diseases ......................... 2
  1.4 Study framework .............................................................................................................. 3
  1.5 Study objectives .............................................................................................................. 5

Chapter 2. Quality of Randomized Clinical Trials in Juvenile Idiopathic Arthritis .............. 6
  Abstract ..................................................................................................................................... 7
  2.1 Introduction ...................................................................................................................... 8
  2.2 Materials and methods ................................................................................................. 9
    2.2.1 Eligibility criteria and search strategy ........................................................................ 9
    2.2.2 Data extraction .......................................................................................................... 9
    2.2.3 Data analysis ............................................................................................................. 10
  2.3 Results ............................................................................................................................ 11
    2.3.1 Search results .......................................................................................................... 11
    2.3.2 General description of studies .................................................................................. 11
    2.3.3 Quality indicators .................................................................................................... 12
    2.3.4 Quality indicators and study results ........................................................................ 13
    2.3.5 Temporal trends of the quality indicators ................................................................ 13
  2.4 Discussion ....................................................................................................................... 14

Chapter 3. Distribution of Time to Treatment Response in Arthritis Patients ...................... 26
  Abstract ................................................................................................................................... 27
  3.1 Introduction ...................................................................................................................... 28
  3.2 Materials and methods ................................................................................................. 29
    3.2.1 TEMPO trial ............................................................................................................ 29
    3.2.2 German Etanercept Registry .................................................................................. 30
    3.2.3 Statistical analyses ............................................................................................... 30
  3.3 Results ............................................................................................................................ 31
    3.3.1 TEMPO trial ............................................................................................................ 31
    3.3.2 German Etanercept Registry .................................................................................. 32
  3.4 Discussion ....................................................................................................................... 33

Chapter 4. The Impact of Survival Distributions on the Power of Randomized Clinical Trials 43
  Abstract ..................................................................................................................................... 44
  4.1 Introduction ...................................................................................................................... 45
  4.2 Materials and methods ................................................................................................. 47
    4.2.1 Simulation design and parameters ........................................................................... 47
    4.2.2 Simulation outcomes (output variables) ................................................................. 48
    4.2.3 Assumptions ........................................................................................................... 49
    4.2.4 Simulation scenarios .............................................................................................. 49
  4.3 Results ............................................................................................................................ 50
List of Tables

Chapter 2. Quality of Randomized Clinical Trials in Juvenile Idiopathic Arthritis
Table 1. Quality indicators and definitions
Table 2. General characteristics of RCTs
Table 3. Generation of allocation sequence, allocation concealment, and masking
Table 4. Time trends of quality indicators

Chapter 3. Distribution of Time to Treatment Response in Arthritis Patients
Table 1. Parametric estimates of survival distributions: TEMPO trial
Table 2. German Etanercept Registry patient characteristics
Table 3. Parametric estimates of survival distributions: German Etanercept Registry

Chapter 4. The Impact of Survival Distributions on the Power of Randomized Clinical Trials
Table 1. Simulation scenarios and parameters
Table 2. Total sample size required to achieve statistical significance (alpha < 0.05)

Chapter 5. Value of Information in Hemophilia Trials
Table 1. Parameter estimates for the trials
Table 2. Sensitivity analyses for Comparison 1 (threshold value = $1,000,000)
List of Figures

Chapter 2. Quality of Randomized Clinical Trials in Juvenile Idiopathic Arthritis
Figure 1. Flow diagram for selection of articles for the review
Figure 2. Quality ranking of RCTs in JIA

Chapter 3. Distribution of Time to Treatment Response in Arthritis Patients
Figure 1. Parametric and non-parametric estimates for probability of response to Combined treatment (ACR20, TEMPO trial)
Figure 2. Hazard function under the lognormal distribution for ACR20 for the three treatment groups (TEMPO trial)
Figure 3. Parametric and non-parametric estimates for probability of response for ACR30 (German Etanercept Registry, non-interval censoring)

Chapter 4. The Impact of Survival Distributions on the Power of Randomized Clinical Trials
Figure 1. Presentation of the Randomized placebo-phase design and the parallel groups RCT
Figure 2. Power of the RPPD and parallel groups design RCTs for different scenarios

Chapter 5. Value of Information in Hemophilia Trials
Figure 1. Net benefit of all treatment methods calculated as $\lambda \cdot E - C$
Figure 2. Optimal sample size as a function of threshold value ($\lambda$)
Figure 3. Expected value of sample information and expected total cost
List of Appendices

Appendix 1. Articles included in the final review
Appendix 2. Statistical characteristics of survival distributions
Appendix 3. The R code for computer simulations
Appendix 4. Sample size and power calculations by different programs
Appendix 5. Value of information: Components and formulas
Appendix 6. The R code for VOI calculations
Thesis Overview

The overall aim of this project was to develop and evaluate methods for improving the design of randomized clinical trials (RCTs) for therapies for rare diseases. The specific aims were addressed with four projects directed to different methodological aspects of RCTs in rare diseases.

Chapter 1 of this dissertation is an introductory chapter presenting the epidemiology of rare diseases, difficulties in studying rare diseases and special policies existing in different countries in support of research for rare diseases. The study framework along with the list of specific objectives addressed by this dissertation is presented at the end of this chapter.

Chapter 2 is the first of the four projects. This study evaluated the quality of RCTs in rare diseases using six individual quality indicators applied to RCTs in Juvenile Idiopathic Arthritis as an example. The search strategy and the definition of quality indicators used in the study are presented in the text, and the list of RCTs included in the analysis is presented in Appendix 1.

Chapter 3 presents the second project of this dissertation. This study evaluated the treatment response time distributions in arthritis patients, using a data set from a previously conducted RCT in Rheumatoid Arthritis and another data set from a registry of JIA patients. The statistical characteristics of survival distributions considered in the study are presented in Appendix 2.

Chapter 4 is the third project of this dissertation. This study evaluated the impact of the response time distributions on the power of RCTs. Two designs of RCTs, the randomized placebo-phase design (RPPD) and the traditional parallel groups design were compared using Monte Carlo computer simulations. Appendix 3 presents the R program code for the simulations.

Chapter 5 is the fourth project of this dissertation. The study applied value of information (VOI) methods to rare diseases with high treatment costs, using hemophilia as an example. Three currently acceptable hemophilia treatment methods were compared: on-demand therapy, tailored prophylaxis and high dose prophylaxis with Factor VIII. The components and formulas
used in this study are presented in Appendix 4, and the R program code for calculations is presented in Appendix 5.

Chapter 6 is the final, conclusion chapter. It presents the synthesis of each project along with the implications for policy and future research directions.
Chapter 1. Introduction

1.1 Epidemiology of rare diseases
Recently, due to the difficulties in conducting studies and the high costs of drug development, there has been an increasing interest in rare diseases. Currently there are about 5,000 to 6,000 rare diseases registered worldwide. Rare diseases when taken together are not so ‘rare’; their global burden is large, affecting about 25 million North Americans and 30 million Europeans [1]. With the development of medical technologies and a subsequent better understanding of the underlying pathophysiological mechanisms, broad disease categories have been separated into smaller and more well-defined disease entities. In this way, about 250 new rare diseases are described every year [2].

No single definition of rare disorders exists, and different countries accept different prevalence thresholds for purposes of classification. In the United States, any disease with a prevalence of 7 cases per 10,000 population, or fewer, is considered to be rare [3]. By the European Union definition, a disease must affect 5 or fewer cases per 10,000 population to be considered rare [3, 4]. In the UK, there is a working definition for ultra-orphan diseases - a condition with a prevalence of fewer than 1 case per 50,000 population [3]. By the definition of The Canadian Organization for Rare Disorders, a rare or orphan disease affects fewer than 1 in 20,000 people. For most of these rare diseases no effective cure is available. Drugs for rare or orphan diseases are often expensive to develop and to produce [4].

1.2 Special policies for rare diseases
Over the recent decades, different countries have implemented special polices and regulations to support research and development of orphan and ultra-orphan drugs. The first regulation was established in 1983 in the U.S. by the Food and Drug Administration Orphan Drug Act [5, 6]. This legislation provides tax incentives, seven-year market exclusivity, expedited reviews for market approval, and support in developing research protocols to drug developers. Since then, over 280 orphan designated products have been introduced to the U.S. market providing treatment for more than 14 million patients (compared to only 10 drugs in the previous decade) [7].
Similar policies were adopted in 1993 in Japan [8] and six years later in Australia [9]. In 2000, following the recommendations derived from a survey done for the European Parliament, the European Union passed new legislation related to the research, development and marketing of orphan drugs [10]. The EU legislation offered market exclusivity to pharmaceutical companies for ten years in all countries, lower fees for market authorization, assistance with marketing applications, and licensing fee reductions [10]. However, the legislation did not address inequality of access across individual EU states. A recent survey showed that the first five orphan products, authorized by the European Medical Agency, were only available in six countries with variable timing of availability and cost [11].

Currently there is no similar legislation in Canada. In 1996, Health Canada concluded that there was no need for a specific orphan drug policy, and that the existing programs can accommodate the support for research or market access for orphan drugs [12].

Although the existing policies facilitate the market approval of orphan drugs, they generally do not address patient timely access and reimbursement. These decisions are taken at the national, regional, state or provider level [13].

1.3 Difficulties in conducting randomized clinic trials in rare diseases

There are many challenges in conducting randomized clinical trials (RCTs) in rare diseases. The main difficulty is low patient accrual due to the low disease prevalence (and often wide geographical dispersion) [7]. Some RCTs are stopped early because of low enrollment [14]. In many rare diseases, the variable and sometimes unpredictable disease progression is another obstacle in enrolling patients in RCTs, where usually a more homogenous patient population is preferred. Because many rare diseases have a genetic origin with a chronic, disabling and even life-threatening progression, it can take a long time for patients to develop objective outcomes such as disability or death [1, 15, 16]. The validation of subjective outcomes can be time consuming and is not always planned as a part of the study protocol. Because of these difficulties, RCTs in rare diseases may have poor quality and may be less efficient compared to trials in common diseases.
When there is a great uncertainty around treatment efficacy, patients and physicians are more reluctant to enroll in studies. As alternatives to parallel RCTs, other designs have been implemented to increase study power and acceptability. For example, in the cross-over RCT design patients switch allocated treatments after a specific time and a washout period. However, this design is only suitable for patients with chronic, relatively stable conditions, and drugs that do not introduce disease modifying changes. N-of-1 trials enroll single patients with variable disease conditions and the need for an individual treatment [15]. These trials, which require the same conditions as cross-over trials, have low generalizability. Some other new designs have not had a wide application. An example is the randomized withdrawal design, which enrolls only those patients who respond to the treatment in an initial open-label phase [17, 18]. Another innovative design for rare diseases (which alternated placebo and experimental treatment of patients on an annual basis) required ten years to enroll 39 patients – the one time the design was attempted [19].

There would seem to be a need to develop new, more acceptable and more powerful designs for RCTs in rare diseases. The urgent need to improve the efficiency and effectiveness of the clinical trial process, including trial design, endpoints and analysis, was stressed in the Federal Drug Administration (FDA) white paper entitled *Critical Path to New Medical Product Development* released in April 2004 [20]. It emphasized that without refinement of the critical path many opportunities will be missed, creating more frustrations and an escalating financial and moral impact on society. The white paper proposed the utilization of quantitative clinical trial modeling to improve trial designs and to predict outcomes. This thesis aims to improve the designs of randomized clinical trials in rare diseases by validating a new design of the RCT and evaluating strategies that will increase the informativeness of future trials.

### 1.4 Study framework

The RCT design is a dynamic model with multiple components including the disease, the patient population and the drug, trial design specification (parallel-groups or other design), patient enrollment scheme including follow-up visits and planned approaches for patient drop-outs, patient compliance, outcome measurement, data analysis, and techniques to minimize bias. The latter is addressed most often by adequate random allocation, allocation concealment, blinding, selection of objective outcomes, and data analysis techniques to properly account for
drop-outs. We hypothesized that the challenges in conducting RCTs for rare diseases may impact the quality of studies. The first project of this thesis evaluated the quality of RCTs in Juvenile Idiopathic Arthritis (a rare, chronic, childhood disease) and indicated the areas where improvements were needed. Appendix 1 lists all the articles that were included in the final review.

The development of a new design for RCTs is a complex task. Any new design should be tested for its overall feasibility, acceptability and potential threats to the study’s internal and external validity. The randomized placebo-phase design (RPPD) is a relatively new design recommended for conducting randomized clinical trials in rare diseases [21]. This is a type of group sequential design where all patients eventually receive the experimental treatment. Because of the design structure, it is assumed that the RPPD is more suitable for chronic conditions. Because all patients receive the experimental treatment, it is assumed that the design will be more acceptable to patients and physicians, and therefore lead to higher accrual, which may determine the difference between trial success and failure when subjects are rare.

Previously the RPPD was compared – in simulation studies – to the parallel-design RCT using, as an example, arthritis patients where the response time to treatment was assumed to follow an exponential distribution [21]. However, this assumption had never been tested. The assumptions and the overall design of the RPPD were felt to need further validation. The second project evaluated the response time to treatment probability distributions in patients with Rheumatoid Arthritis – using data from a previously conducted RCT – and in patients with Juvenile Idiopathic Arthritis using data from a patient registry. Appendix 2 presents the statistical characteristics of survival distributions that were under consideration.

The third project used the findings from project 2 to evaluate the statistical efficiency of the RPPD and the traditional parallel groups RCT under different survival distributions using Monte Carlo computer simulations. For the purpose of computer simulations, an object-oriented program was developed in the R statistical language. Appendix 3 presents the simulation code. To enhance the use and applicability of the new design by medical and scientific societies, the code was published as an electronic supplement to a published article presenting the simulation results.
As previously discussed, the cost of drug development and the cost of treatment are very high in many rare diseases. One approach to optimize the design of RCTs in rare diseases is the evaluation of the value of information (VOI). This analytical technique compares the potential value of new information against the total trial cost when planning a new study in the context of a Bayesian decision-theoretical framework. Factors taken into account in this technique include the disease annual incidence, current incremental net benefit, time horizon for the decision rule, duration of follow-up, trial costs, and discount rate. The last, fourth project evaluated the value of information for patients with hemophilia, which is a rare disease with high treatment costs. Details about the components and formulas for calculation of the value of information are presented in Appendix 4. The R code for calculations is presented in Appendix 5.

1.5 Study objectives
The specific objectives addressed by this dissertation are the following:
1. Evaluate the quality of RCTs in rare diseases using Juvenile Idiopathic Arthritis as an example (Project One);
2. Estimate the response time to treatment statistical distributions in patients with rheumatic diseases (Project Two);
3. Evaluate the statistical efficiency of the RPPD compared to the parallel RCT design under different survival distributions, using Monte Carlo computer simulations (Project Three);
4. Demonstrate the use of value of information methods in randomized clinical trials of rare diseases, using hemophilia as an example (Project Four);
5. Recommend future policy directions and research directions to improve the designs of RCTs in rare diseases in order to better establish an evidence base for treatment.
Chapter 2. Quality of Randomized Clinical Trials in Juvenile Idiopathic Arthritis

Publication
(Rheumatology Advance Access originally published online on February 2, 2008)
Printed with permission, Oxford University Press

QUALITY OF RANDOMISED CLINICAL TRIALS IN JUVENTILE IDIOPATHIC ARTHRITIS
L Abrahamyan, SR Johnson, J Beyene, PS Shah, BM Feldman
Lusine Abrahamyan MD MPH¹,³,⁴, Sindhu R. Johnson MD FRCPC¹,³,⁴, Joseph Beyene MSc PhD¹,³,⁴,⁵, Prakesh S. Shah MD MSc FRCPC¹,²,³; Brian M. Feldman MD MSc FRCPC¹,²,³,⁴,⁵.
¹The Hospital for Sick Children; ²Departments of Paediatrics, ³Health Policy Management & Evaluation, ⁴Child Health Evaluative Sciences and ⁵Public Health Sciences, University of Toronto, Toronto, Ontario, Canada

Corresponding author: Brian M. Feldman
Tel: +416 813 5828; fax: +416 813 4989; e-mail: brian.feldman@sickkids.ca

Address for reprint requests: The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada

Word Count: 3207

Running Title: Quality of RCTs in JIA

Key words: Juvenile idiopathic arthritis, randomised clinical trials, systematic review, quality
Abstract

Objectives. We evaluated the quality of randomised clinical trials (RCTs) of therapy for Juvenile Idiopathic Arthritis (JIA) using an individual component approach and assessed temporal changes.

Methods. A systematic review of the literature was performed to identify all RCTs involving exclusively JIA patients. Two investigators independently assessed the identified articles for six quality indicators: generation of allocation sequence, allocation concealment, masking, intention-to-treat (ITT) analysis, dropout rates, and clearly stated primary outcome.

Results. Fifty-two RCTs involving JIA patients were assessed. Generation of allocation sequence was unclear in 79% of studies. Reporting of allocation concealment was adequate in only one-third of the studies. Masking was adequate in 73%, inadequate in 19% and unclear in 8% of reports. ITT analysis was employed in 37% of the reports. Per-protocol analysis was used in 40% and in 23% the method was unclear. Most of the reports (67%) had dropout rates ≤ 20%. About half of the reports (n = 25) failed to show a significant effect of the experimental treatment. No significant associations were found between the study results and quality indicators. With the exception of adequate masking and dropout rate, all quality indicators showed a trend of improvement over the decades.

Conclusions. The quality of RCTs in JIA based on the selected indicators was poor. Although there were some positive changes over time, the reporting and methodological quality of trials should be improved. New, more powerful and acceptable RCT designs should be developed in this patient population.
2.1 Introduction
Juvenile idiopathic arthritis (JIA) is the current name used to describe several subsets of childhood arthritis; JIA includes the subsets previously called juvenile rheumatoid arthritis as well as other types of childhood arthritis. JIA is the most common rheumatologic disease in children. The reported prevalence varies from 0.07 to 4.01 per 1000 children in different countries. [22] JIA has six categories, of which the most common are oligoarthritis (40-60% of patients), rheumatoid factor negative polyarthritis (20-25% of patients), and systemic arthritis (10-20% of patients). [23] The disease burden is considerable both in childhood and later in adulthood. Several studies have shown that more than one-third of patients extend the disease into adulthood [24, 25] often with a significant level of disability [26], increased probability of arthroplasty [27], uveitis [28], osteoporosis [29], high treatment costs [30] and decreased quality of life [31].

A recent review of medical treatments of JIA summarized evidence from 34 randomised clinical trials (RCTs) conducted between 1966 and 2005. [23] The authors concluded that despite the significant advances in the treatment of the disease, there is still a lack of evidence in treatment for several subtypes of JIA. The review outlined some of the design features of included RCTs (e.g. type of design, masking, and length of treatment) but the quality of individual RCTs was not evaluated.

There are significant challenges in conducting RCTs in rare diseases such as JIA. In the study of rare diseases a considerable amount of time may be needed to recruit adequate numbers of patients, which, in turn, can become a reason for discontinuing a study. In JIA, the disease progression can be complex and may require an individualized treatment approach. It can also take many years for a patient to develop an objective outcome such as disability. Moreover, in RCTs of paediatric diseases there may be ethical concerns such as an unwillingness to commit children to trials where they may receive a placebo.[32] Therefore, the challenges to adequately conducting RCTs in this population are considerable.

The quality of published RCTs in JIA has not been investigated. Thorough evaluation of the quality of individual studies will improve our understanding of the challenges in conducting
methodologically sound RCTs and hopefully identify avenues for improvement. This will also serve to enhance subsequent systematic reviews and meta-analyses.

Our aim was to evaluate the quality of RCTs that examined all interventions (pharmacological and non-pharmacological) in JIA using an individual component approach. We hypothesized that the overall quality of RCTs in JIA was poor; however, there was an improvement in the quality of reported RCTs in JIA over the last three decades.

2.2 Materials and methods

2.2.1 Eligibility criteria and search strategy
We reviewed all RCTs of interventions (pharmacological and non-pharmacological) where the study population consisted exclusively of JIA patients irrespective of age and JIA subtype. We excluded quasi-randomised studies or studies that had a mixed population of JIA and other patients. If the same trial was published more than once with different follow-up periods or different outcome evaluations, we included only the study with the report of the principal analysis planned \textit{a priori}. When the study had multiple phases (for example, began as an RCT and then was continued as open-label), only data from RCT was extracted. We included articles published in the English language only. We excluded articles published only as an abstract.

We searched MEDLINE (1966- December 2006), EMBASE (1980- December 2006), and the Cochrane Central Register of Controlled Trials (CENTRAL) for all publications describing RCTs in JIA. In addition, the reference lists of identified trials were searched. Potentially relevant articles were first identified on the basis of their title, abstract and keywords by one of the authors (LA). Two authors (LA and SJ) independently assessed the full text of selected articles and determined the set of trials for the final review. MeSH terms for search included: juvenile rheumatoid arthritis; randomised controlled trial; controlled clinical trial; multicentre study; clinical trial; English; and human studies. The terms were adjusted according to database specific terms.

2.2.2 Data extraction
We selected the important quality indicators to reflect areas where bias (selection bias, performance bias, measurement bias and attrition bias) may occur in RCTs. [33] A data
collection form was developed based on the selected quality indicator components as recommended by Khan et al. [33] The form was pilot-tested in three RCTs involving patients who did not have JIA. Two independent reviewers (LA and SJ) used the form to extract data from the selected JIA articles. Disagreements were solved by consensus and consulting the other authors (JB, PS, and BF).

We extracted the following data from the trials: first author’s name, publication year, study design (parallel, cross-over), number of study arms, compared interventions, total sample size, mean age of patients, study duration (in weeks), allocation concealment (adequate, inadequate, unclear), generation of allocation sequence (adequate, inadequate, unclear), masking (adequate, inadequate, unclear, not applicable), intent-to-treat (ITT) analysis (full or modified ITT, per-protocol analysis, treatment-received analysis, unclear), dropout rate (exact rate; no dropout, ≤ 20%, >20%), methods used to handle missing data, clearly stated primary outcome(s) (yes, no), type of primary outcome(s) (single, composite, multiple) and description, participating centres (single, multicentre or multinational), and the final conclusion of the study. We also evaluated the study funding source (not sponsored by industry, industry sponsored, not reported). The study was considered as industry sponsored if the pharmaceutical manufacturer was acknowledged by the authors or if any of the authors was affiliated with a pharmaceutical company. The main quality indicators of interest were allocation concealment, generation of allocation sequence, masking, intention-to-treat analysis (ITT), dropout rate, and clearly stated primary outcome (Table 1). The definitions for the indicators were selected based on the Consolidated Standards of the Reporting of Trials (CONSORT) guideline and previous studies. [34-36]

2.2.3 Data analysis
Descriptive statistical analyses were performed using mean, standard deviation or median for continuous data and frequencies and percentages for categorical data. Continuous data were compared using t test and analysis of variance; categorical data were compared using chi-square test or Fisher’s exact test as appropriate. Two-tailed p-values were reported. Characteristics of the studies and quality indicators were described and the final results of the studies were compared with the quality items and funding source. Finally, we evaluated the
change of quality indicators over time in decades defined by year of publication. SPSS 11.0 statistical software (SPSS, Inc., Chicago, IL, USA) was used for data entry and analysis.

2.3 Results

2.3.1 Search results
After removing the duplicates, the applied search strategy identified 471 abstracts in total. The abstracts presenting non-randomised, cross-sectional comparative or cohort clinical studies were excluded (n = 417). Fifty-four full text articles underwent full review. One study was excluded because it was not a truly randomised trial. [37] Another excluded study described only the pharmacokinetics of the drug.[38] The final review included 52 articles (see Appendix 1). Figure 1 presents the selection process of articles for the final review.

2.3.2 General description of studies
Table 2 presents general characteristics of the reviewed studies. There were eight cross-over studies. The remaining 44 RCTs used a parallel-group study design. Among them two studies randomised each patient’s joints to different treatments. Another three utilized a randomised withdrawal design, where patients entered into the double masked study only after they had responded to the treatment during an initial open-label run-in phase. In one study patients entered into the RCT only if they did not respond to the treatment in the open-label phase. Forty-two percent of the studies included placebo as a comparison group. The remaining studies compared experimental and contemporaneous “standard treatment” methods. Most of the studies compared medical treatments for JIA (n = 36). The other RCTs evaluated treatment of JIA complications such as uveitis (n=1), growth retardation (n=3), a surgical intervention (n=1), orthopaedic splints and shoes (n=2), local analgesia for intra-articular injections (n=1), folic acid supplementation (n=1), behavioural interventions to increase calcium intake (n=2), effect of daily supplementation with calcium (n=1), interventions for medication adherence (n=1), aquatic training (n=1), massage therapy (n=1), and combined hydrotherapy and physiotherapy (n=1). About half (52%) of the studies were single centre studies and 44% (n = 23) of the studies were industry sponsored. In 7 studies pharmaceutical companies provided either the standard or the experimental treatment but it was unclear whether industry funded the study.
Table 2 shows the median sample size of the studies by design. The median sample size was 30 for cross-over RCTs (n=8) and 49 for parallel RCTs (n=44). The median sample size of parallel RCTs in medication treatment of JIA (n=30) was 71. Other general characteristics of RCTs of medical treatment of JIA were quite similar to the results from all studies.

The overall mean or median age of patients in the included RCTs was impossible to evaluate because of the poor and inconsistent reporting in some studies. The range of the reported means was from 5.5 to 15.9 years if excluding one study reporting all patients being 25 years old.

### 2.3.3 Quality indicators

The generation of the allocation sequence and presence/absence of allocation concealment were unclear in the majority of RCTs (Table 3). There were 35 double masked studies. In 7 studies it was reportedly impossible to mask patients from the interventions (such as massage therapy, orthotic devices, and dietary interventions). For such studies we considered that masking was adequate if the outcome assessors were masked (n = 3).

Intention-to-treat analysis was employed in 19 studies (37%); full ITT in 12 studies and modified ITT in 7. Twenty-one (40%) studies conducted per protocol analysis, while in 12 studies (23%) the approach was unclear. Eighteen RCTs (32.7%) had dropout rates of more than 20% of the original sample while 6 studies (12%) had no dropout. Overall, the mean dropout rate was 15% (SD = 13; range from 0 to 52%). In two studies, it was impossible to estimate the exact dropout rate because of the inconsistent number of patients for different measured outcomes. No significant correlation was found between the indicators of study duration and exact dropout rate (Pearson’s correlation coefficient was 0.273; p > 0.05). Among 19 studies that used ITT, only 5 mentioned the method of handling missing data.

Only half of the reports clearly specified the primary outcome (n = 26). Among the 26 studies with clearly stated outcomes, a single primary outcome was evaluated in 16 cases and multiple outcomes in 10 of them. Multiple outcomes mostly included general clinical, articular, and/or laboratory assessments. Overall, 16 studies (31%) measured at least one composite outcome out of which 13 were clearly stated primary outcomes.
About 70% of all RCTs achieved an adequate rating in a maximum of 3 quality indicators out of the evaluated 6; similar results of achieved numbers of quality indicators were observed for RCTs when we considered only medication treatment (Figure 2).

2.3.4 Quality indicators and study results
In 25 (48%) studies, no statistically significant difference was found between the experimental and control treatments. The experimental treatment was superior in 25 RCTs. In one non-inferiority trial the experimental treatment was found to be not inferior to standard treatment. In another study the experimental intervention had less efficacy compared with the standard treatment. In 3 studies patients in the standard treatment (acetylsalicylic acid was used in all cases) had significant adverse events.

We did not find a statistically significant association between the study results and the source of funding (p > 0.05, Chi-square test; raw data not shown). Similarly there was no statistically significant difference in the reported quality indicators between positive and negative studies.

2.3.5 Temporal trends of the quality indicators
In order to show the temporal trends, RCTs were grouped in ten year intervals by their year of publication. The first two RCTs in JIA patients were conducted in 1977. The time periods and respective number of RCTs were: 1976-1985 (n = 11), 1986-1995 (n = 17), and 1996-2006 (n = 24). Compared to the first decade, the number of RCTs had almost doubled during the time period between 1996 and 2006. Table 4 shows the changes of quality indicators over the three decades.

With the exception of adequate masking and dropout rate, all quality indicators showed a trend for improvement over the decades. Compared to the second decade, the number of RCTs reporting adequate allocation concealment increased almost five fold, and the number of RCTs with clearly stated primary outcomes increased almost two times during the third decade. Temporal increases in ITT analysis between the first and third decades, and second and third decades were also statistically significant.
2.4 Discussion

Rare disease trials face a number of methodological challenges in the hope of determining the efficacy of an intervention. Thus, assurance of methodological quality is a critical and necessary endeavour to ensure that limited resources are effectively utilized and trial results are appropriately translated into clinical practice. Our study evaluated 52 RCTs for JIA patients beginning from 1977 to present; a number much lower than might be expected when compared to more common diseases in the adult population. We identified that most of the quality components that we evaluated were seldomly met adequately.

The generation of random number sequences was the most rarely reported quality indicator; this was mostly persistent across the three time periods. The overall reporting of allocation concealment was low; however, in contrast to random number generation, it had a five-fold increase after 1995. In different diseases, the proportion of RCTs with reported adequate randomisation methods and adequate allocation concealment varies from 1 to 52 percent and from 2 to 39 percent respectively.[39] While the overall rates of these quality indicators have been low, significant changes in reporting are likely due to the publication of the CONSORT statement in 1996.[34] A study that compared the quality of RCTs in adult rheumatology between two time periods, 1987-1988 and 1997-1998 suggested that among the studied quality indicators only the reporting of allocation concealment increased significantly.[40] The improvement of quality over time could also be explained by the establishment of international networks such as The Paediatric Rheumatology Collaborative Study Group (PRCSG) in 1973 and The Paediatric Rheumatology International Trials Organization (PRINTO) in 1996, which aim to conduct high quality clinical trials in children with rheumatic diseases.

Although the effect of adequate masking on the study main results is controversial,[39] masking is still one of the current standards for RCTs. Adequate masking should be guaranteed especially when the outcomes are susceptible to measurement bias. In JIA masking is especially important considering the use of subjective outcomes such as patient or physician global assessments using visual analogue scales. In our study double masking was used in 67.3% of studies, and no positive temporal change was observed. This result is inconsistent with some other similar studies in the field. A study that assessed RCTs of rheumatological disorders in adults identified a higher rate of double masking of studies both between 1987 and
In contrast, in systemic lupus erythematosus only 54.3% of RCTs were double masked.[41] The rate of double masked studies might depend on the type of compared interventions. If we remove those studies in which double masking was thought to be unfeasible (n=7), our number of RCTs with double masking rises to 77.8%.

The importance of clearly stated primary outcomes is very high in JIA and other systemic diseases, where outcomes are often subjective, and mortality is not a large concern. Depending on the evaluated treatment, the outcomes of interest can vary from study to study. In JIA, a validated measure of response to treatment was developed in 1997.[42] The American College of Rheumatology (ACR) Paediatric 30 score is a composite outcome consisting of 6 different measures, and is currently the most widely accepted method to estimate the response to treatment for JIA. Most probably two factors, the publication of the CONSORT guideline and the development of a universal validated outcome, played a role in the significant increase in the number of RCTs in JIA with clearly stated outcomes. In the review of RCTs in systemic lupus erythematosus, adequately reported outcome was the only quality indicator that showed a significant improvement over time.[41]

Although in this study the most frequently used approach was per-protocol analysis, a significant temporal increase of RCTs with ITT was observed. Application of ITT methods yields unbiased and valid estimates not compromising the balance between the groups achieved by randomisation.[43] A study that evaluated ITT principles in superiority trials of structural outcomes in adult rheumatic diseases between 1994 and 2003 also found that most of the RCTs employed per-protocol analysis.[35] They reported that about 28.6% of studies had more than 20% missing data. The method of handling missing data was reported in 23.5% of the studies and the most prevalent method was the last observation carried forward approach.[35] In our study most of the RCTs had less than 20% dropout rate and there was no significant increase in the dropout rates over time. However, in our study only five RCTs reported the method of handling missing data, four of which used the last observation carried forward approach and one used sensitivity analysis.
The main objective of our study was to evaluate the quality of RCTs in JIA, which was poor for most of the selected indicators. However, some of the observed characteristics of the studies highlighted the difficulties in conducting RCTs in this population. We found that the sample size was less than 50 in more than half of the included parallel RCTs. Some of the reviewed RCTs discussed explicitly the difficulties related to patient accrual. One of the studies enrolled only 25 patients over 3 year period from 7 centres in the U.S.A. and Canada. Another study could not reach the needed sample size of 15 patients in the planned 18 months and was stopped early. Almost half of the studies failed to show a significant difference between the compared treatments – in some cases likely due to small sample sizes and low power. The results were neither associated with the levels of quality indicators nor with the source of funding. Lack of objective outcomes is another difficulty in this population. Although the ACR Paediatric 30 is a validated response measurement, it is a composite outcome and the clinical interpretation of the results can be challenging. Moreover, depending on the compared treatment methods (e.g. medication, massage therapy, behavioural interventions, surgery) the optimal outcome measurement may not be the ACR Paediatric 30. About half of the studies were either multicentre or multinational as a way to facilitate patient accrual. Alternative designs were also employed to make the studies more acceptable and more powerful. One-fifth of the RCTs in JIA used non-traditional designs such as cross-over and randomised withdrawal designs.

Study quality is relevant to the interpretation of study results. Several studies showed that low quality studies are more likely to yield biased results.[44, 45] In our study we did not see significant associations between the quality indicators and the study results. However, this type of conclusion depends also on the number of included studies (n=52) which is relatively small compared to previous studies.

One of the limitations of our study was that only English language and full text articles were included. We believe that this will not likely bias the results of the study because when the disease is rare and it is not easy to conduct an RCT, there should likely be more desire to report the results in English language journals. Moreover, the inclusion or exclusion of non-English language articles is felt by methodologists to be more of an issue when the systematic review is focused on treatment effectiveness rather than study quality.[46] Another limitation might be
the reporting quality of the studies. It is possible that reporting quality is not always a good measure of the actual methodological quality of a study.[47] For example, we accepted masking as adequate if the authors named or described it as such. We categorized allocation concealment or method of random number generation as unclear if no information was available from the articles. This deficiency can be related to other factors related to reporting (e.g. journal word count restrictions) rather than the methodological quality of RCTs. In contrast, some of the components we have measured were directly related to methodological quality (such as ITT and dropout rate). However, the methodological and reporting quality of the studies are usually thought to be highly related.[48] Due to the constraints of feasibility we did not evaluate the methodological quality of RCTs by JIA subtype. Future investigators should be aware that the response to the treatment and the choice of the appropriate outcome can differ between the different subtypes. This is a further complicating issue for RCTs for this disease.

In summary, the overall quality of RCTs in JIA has been poor and there have been very few RCTs. As a result, there is little high quality evidence for many current practices in JIA. There are positive changes in some of the indicators over time but more efforts should be made for further improvements of the design, conduct and reporting of the studies. In particular, studies should report the method of random number generation and allocation concealment, should make studies double masked whenever possible, conduct ITT analysis and ensure complete follow-up as far as possible. Considering the current difficulties in conducting RCTs in this population – and the likelihood that studies will continue to be relatively few – both patients and physicians will benefit if more powerful and acceptable study designs will be used in the future. Perhaps this can be achieved by the development and testing of new designs for intervention trials for rare diseases.
Key Messages

- Although overall quality of RCTs in JIA is poor, there have been improvements over time.
- The reporting and methodological quality of trials should be improved.

Acknowledgements

Lusine Abrahamyan has been awarded a Canadian Arthritis Network (CAN) Graduate Student Award and Graduate Studentship from the Research Training Centre of the Hospital for Sick Children, Toronto, Canada. Sindhu Johnson has been awarded an Abbott Scholar Award and CAN Fellowship. Brian Feldman is supported by a Canada Research Chair in Childhood Arthritis.

Supplementary data

The list of article included in the review is presented in Appendix 1.
<table>
<thead>
<tr>
<th>Item</th>
<th>Scale and Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Allocation concealment</strong></td>
<td><strong>Adequate</strong> Centralized or pharmacy-controlled randomisation, numbered or coded identical containers administered serially, on-site computer which gave allocation only after participant’s characteristics is entered, serially numbered opaque, sealed envelopes.</td>
</tr>
<tr>
<td></td>
<td><strong>Inadequate</strong> Allocation was done by open, inadequate methods such as alternation of study identification numbers, date of visits or by date of birth, open allocation schedule, unsealed or non-opaque envelopes.</td>
</tr>
<tr>
<td></td>
<td><strong>Unclear</strong> Method was not specified.</td>
</tr>
<tr>
<td><strong>Generation of allocation sequence</strong></td>
<td><strong>Adequate</strong> Sequence was generated by random tables, computer generation, coin-tossing or shuffling.</td>
</tr>
<tr>
<td></td>
<td><strong>Inadequate</strong> Other methods such as case record number, date of birth, date of visit</td>
</tr>
<tr>
<td></td>
<td><strong>Unclear</strong> Method was not specified.</td>
</tr>
<tr>
<td><strong>Masking</strong></td>
<td><strong>Adequate</strong> Defined as double-masked by the authors, or when both patients and assessors are masked, or when assessors were masked in the studies where it was considered impossible to mask patients.</td>
</tr>
<tr>
<td></td>
<td><strong>Inadequate</strong> Either patients or assessors were not masked in studies where it was feasible.</td>
</tr>
<tr>
<td></td>
<td><strong>Unclear</strong> Method was not specified.</td>
</tr>
<tr>
<td><strong>Intent-to-treat analysis (ITT)</strong></td>
<td><strong>Full ITT</strong> All randomised patients were included in the analysis and kept in their original groups decided by randomisation (regardless of any protocol violations, compliance and withdrawals).</td>
</tr>
<tr>
<td></td>
<td><strong>Modified ITT</strong> Analysis included participants with baseline measurement and at least 1 follow-up evaluation, regardless of the time between baseline and the second measurement or analysis excluded patients.</td>
</tr>
</tbody>
</table>
who were defined as ineligible after randomisation or who never received the treatment.

<table>
<thead>
<tr>
<th>Per-protocol</th>
<th>Analysis included only patients who complied with the study protocol including exposure to treatment, availability of the outcome measurement, and absence of major protocol violations (case-complete).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment- received</td>
<td>Analyses were done according to the treatment patients actually received, not according to their randomisation groups or per protocol.</td>
</tr>
<tr>
<td>Unclear</td>
<td>It was impossible to categorize in any group.</td>
</tr>
</tbody>
</table>

**Dropout rate**

- Defined as No dropout, dropout $\leq 20\%$ or $> 20\%$

**Clearly stated primary outcome**

- Yes: If the primary outcome(s) was clearly stated
- No: If there was no clear specification about the primary outcome
Table 2. General characteristics of RCTs

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>All studies n=52</th>
<th>Medication treatment of JIA N=36</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study design</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parallel RCTs</td>
<td>44 (84.6%)</td>
<td>30 (83.3%)</td>
</tr>
<tr>
<td>Cross-over RCTs</td>
<td>8 (15.4%)</td>
<td>6 (16.7%)</td>
</tr>
<tr>
<td><strong>Control treatment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>30 (57.7%)</td>
<td>21 (58.3%)</td>
</tr>
<tr>
<td>Placebo</td>
<td>22 (42.3%)</td>
<td>15 (41.7%)</td>
</tr>
<tr>
<td><strong>Number of arms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two arms</td>
<td>44 (84.6%)</td>
<td>29 (80.6%)</td>
</tr>
<tr>
<td>Three arms</td>
<td>8 (15.4%)</td>
<td>7 (19.4%)</td>
</tr>
<tr>
<td><strong>Involved centres</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>27 (51.9%)</td>
<td>19 (52.8%)</td>
</tr>
<tr>
<td>Multicentre</td>
<td>14 (26.9%)</td>
<td>6 (16.7)</td>
</tr>
<tr>
<td>Multinational</td>
<td>11 (21.2%)</td>
<td>11 (30.5)</td>
</tr>
<tr>
<td><strong>Funding source</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industry sponsored</td>
<td>23 (44.2%)</td>
<td>19 (52.8%)</td>
</tr>
<tr>
<td>Not industry sponsored</td>
<td>13 (25.0%)</td>
<td>4 (11.1%)</td>
</tr>
<tr>
<td>Source not reported</td>
<td>9 (17.3%)</td>
<td>8 (22.2%)</td>
</tr>
<tr>
<td>Unclear</td>
<td>7 (13.5%)</td>
<td>5 (13.9%)</td>
</tr>
<tr>
<td><strong>Median sample size</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parallel RCTs</td>
<td>49 (IQR: 29-80)</td>
<td>71 (IQR: 31-99)</td>
</tr>
<tr>
<td>Cross-over RCTs</td>
<td>30 (IQR: 22-39)</td>
<td>30 (IQR: 29-47)</td>
</tr>
<tr>
<td><strong>Median follow-up duration</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parallel RCTs</td>
<td>24 weeks (IQR: 12-38)</td>
<td>24 weeks (IQR: 12-24)</td>
</tr>
<tr>
<td>Cross-over RCTs</td>
<td>14.5 weeks (IQR: 8-28)</td>
<td>12 weeks (IQR: 8-16)</td>
</tr>
</tbody>
</table>

IQR; Interquartile range
Table 3. Generation of allocation sequence, allocation concealment, and masking

<table>
<thead>
<tr>
<th>Generation of allocation sequence</th>
<th>Allocation concealment</th>
<th>Masking</th>
</tr>
</thead>
<tbody>
<tr>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>Adequate</td>
<td>11 (21.2)</td>
<td>19 (36.5)</td>
</tr>
<tr>
<td>Inadequate</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Unclear</td>
<td>41 (78.8)</td>
<td>33 (63.5)</td>
</tr>
</tbody>
</table>
Table 4. Time trends of quality indicators

<table>
<thead>
<tr>
<th>Quality indicators</th>
<th>Overall</th>
<th>Year of publication</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Adequate generation of allocation sequence (%)</td>
<td>21</td>
<td>0</td>
<td>18</td>
<td>33(^a)</td>
<td></td>
</tr>
<tr>
<td>Adequate allocation concealment (%)</td>
<td>37</td>
<td>9</td>
<td>12</td>
<td>67(^{a,b})</td>
<td></td>
</tr>
<tr>
<td>Adequate masking* (%)</td>
<td>73</td>
<td>82</td>
<td>71</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>ITT** (%)</td>
<td>37</td>
<td>18</td>
<td>18</td>
<td>58(^{a,b})</td>
<td></td>
</tr>
<tr>
<td>Dropout rate ≤ 20% (%)</td>
<td>67</td>
<td>55</td>
<td>71</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>Clearly stated primary outcome (%)</td>
<td>50</td>
<td>36</td>
<td>23</td>
<td>75(^{a,b})</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Adequate masking was compared to inadequate and unclear masking. \(^b\) Full or modified ITT was compared to per-protocol and unclear analysis. \(^a\) P<0.05 between the first and the third decades. \(^b\) P<0.05 between the second and the third decades.
Figure 1. Flow diagram for selection of articles for the review
Figure 2. Quality ranking of RCTs in JIA
Chapter 3. Distribution of Time to Treatment Response in Arthritis Patients

In Press in the Journal of Clinical Epidemiology.

RESPONSE TIMES FOLLOW LOGNORMAL OR GAMMA DISTRIBUTION IN ARTHRITIS PATIENTS

L Abrahanyan, J Beyene, JY Feng, Y Chon, AR Willan, H Schmeling, G Horneff, EC Keystone, BM Feldman

Lusine Abrahanyan MD MPH\textsuperscript{1,3}, Joseph Beyene MSc PhD\textsuperscript{1,3,6}, JingYuan Feng MS\textsuperscript{8}, Yun Chon PhD\textsuperscript{8}, Andrew R. Willan PhD\textsuperscript{1,6}, Heinrike Schmeling MD\textsuperscript{7}, Gerd Horneff MD\textsuperscript{9}, Edward C Keystone MD FRCP\textsuperscript{4,5}, Brian M. Feldman MD MSc FRCP\textsuperscript{1,2,3,6}

\textsuperscript{1}Child Health Evaluative Sciences program, The Hospital for Sick Children Research Institute,\textsuperscript{2}Departments of Paediatrics, \textsuperscript{3}Health Policy Management & Evaluation, \textsuperscript{4}Medicine and Immunology, \textsuperscript{5}Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; \textsuperscript{7}Division of Rheumatology, Department of Pediatrics, Alberta Children's Hospital, University of Calgary, Calgary, Canada, \textsuperscript{8}Amgen Inc., Thousand Oaks, CA; \textsuperscript{9}Department of Pediatrics, Asklepios Clinic, Sankt Augustin, Germany

Corresponding author: Brian M. Feldman

Tel: +416 813 6117; fax: +416 813 4989; e-mail: brian.feldman@sickkids.ca

Address for reprint requests: The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada

Word Count: 3061

Running Title: Distribution of time to treatment response in arthritis

Key words: Survival distribution, rheumatoid arthritis, juvenile idiopathic arthritis, Methotrexate, Etanercept
Abstract

Objective: This study evaluated the statistical distribution of time to treatment response in patients with rheumatic diseases.

Study Design and Setting: The study utilized a secondary data analysis design. Data from the TEMPO trial was used to model the response times for etanercept (ETN), methotrexate (MTX), and combined ETN+MTX in patients with rheumatoid arthritis. The German Etanercept Registry was used to evaluate the response time distributions in patients with juvenile idiopathic arthritis.

Results: For MTX, the lognormal distribution was considered to be the best model for the outcome of ACR20, lognormal, generalized gamma, and log-logistic distributions for ACR50, and lognormal and generalized gamma for ACR70. For ETN, the lognormal model was best for ACR20, the generalized gamma for ACR 50, and both lognormal and generalized gamma distributions for ACR70. For Combined treatment, the best model was the log-logistic distribution for ACR20, generalized gamma for ACR50, and both lognormal and generalized gamma distributions for ACR 70. For the German Etanercept Registry, the lognormal distribution was the best model for all three outcomes of Pediatric ACR30, ACR50, and ACR70 without interval censoring.

Conclusion: Study designs might be more efficient if the response distributions are taken into consideration during planning.

Abstract word count: 196
3.1 Introduction
Survival analysis is used to model the time to an event and to answer important questions about the probability of the event over time. When planning randomized clinical trials (RCTs) for which time-to-event is the primary outcome, it is important to properly describe the probability distribution of the response time to treatment. Although there are several non-parametric tests for comparing survival distributions (e.g. Cox-Mantel test, logrank test, Gehan’s generalized Wilcoxon test), it has been shown that the power of these tests varies significantly depending on the underlying probability distribution [49]. The response time distribution can also have an impact on sample size calculation; accurately considering response distribution would lead to more efficient trial designs. For example, in most widely accepted sample size calculators for survival analysis it is assumed that the response time follows an exponential distribution [50, 51]. In other words, it is assumed that the hazard is constant over time for all patients in the study. However, this assumption is rarely tested. It may be, for example, that the hazard rate has an increasing or decreasing function or has a bathtub shape or other changing pattern over time. Ignoring this information can reduce the study power and, ultimately, affect the study conclusions, treatment guidelines and the planning of future studies.

This study evaluated the statistical distribution of time to response to drugs in patients with rheumatic diseases. The burden of rheumatic diseases is substantial; the incidence of rheumatoid arthritis (RA) alone is about 1-2% of the world’s population [52]. There have been significant improvements in the management of patients with rheumatic diseases after the introduction of biologic therapies such as tumor necrosis factor (TNF-α) inhibitors (e.g. etanercept, infliximab, adalimumab), and there are ongoing studies testing new medications and biologic agents [52]. The understanding of treatment response time distributions can improve patient management and improve planning of RCTs with new agents in the future.

The aim of this study was to evaluate the time to drug response distributions of patients with rheumatic diseases. We evaluated the data of the TEMPO (Trial of Etanercept and Methotrexate with Radiographic Patient Outcomes) trial which was a randomized, double-blind, parallel-group study of RA patients with three treatment arms: etanercept only (ETN), methotrexate only (MTX) or etanercept combined with methotrexate (Combined ETN+MTX) [53]. We also analyzed data from the German Etanercept Registry which is a prospective
registry of patients with Juvenile Idiopathic Arthritis (JIA) [54, 55]. This registry started enrolling patients in 2001 and currently has more than 1000 patients.

3.2 Materials and methods
A secondary data analysis design was used to estimate the probability distribution of time to response to drugs in patients with rheumatic diseases. Information about the patient population, treatment methods and outcome assessments of the TEMPO trial and the German Etanercept Registry is briefly summarized here and has been published in more detail elsewhere [54, 55].

3.2.1 TEMPO trial
The TEMPO trial enrolled patients who were 18 years of age or older, with a disease duration of six months to 20 years and active, adult-onset RA who had a less than satisfactory response to at least one disease modifying drug other than MTX. Patients were randomized into one of three treatment groups: ETN 25 mg twice a week subcutaneously and oral placebo once a week (n=223), MTX 7.5 mg escalated to 20 mg oral capsules once a week within 8 weeks if patients had any painful or swollen joints, and placebo subcutaneous injection twice a week (n=228), or a combination of 25 mg subcutaneous etanercept injections twice a week and oral MTX capsules once a week (n = 231). Patients were followed up to 196 weeks; however, frequent visits were planned only for the first year in the study when patients were seen for the assessment of response at weeks two, four and then once monthly. For the purpose of this study, we used data from the first 52 weeks. The primary outcome of the TEMPO trial was the numeric index of the American College of Rheumatology (ACR) response (ACR-N) area under the curve over the first 24 weeks. ACR Core Data Set response criteria ACR20, ACR50, and ACR70 were secondary endpoints, reported as the proportion of patients who met the criteria for each follow-up visit during 52 weeks, and were used in our study. The ACR response criteria is a composite outcome of swollen joint count and tender joint count, physician and patient global assessment of disease activity, patient assessment of physical function, patient assessment of pain, and laboratory evaluation of an acute-phase reactant such as erythrocyte sedimentation rate or C-reactive protein [56]. The ACR20 criteria requires a minimum 20% improvement of both tender and swollen joint counts, in addition to a 20% reduction in three of the other five measures of the ACR core data set. For our analyses, we evaluated the time to the first outcome events – first ACR20, ACR50 or ACR70. These outcomes were not mutually
exclusive – if the patient met ACR70 criteria at the first visit it was considered that ACR20 and ACR50 were also met. However, each outcome was analyzed in a separate data set.

3.2.2 German Etanercept Registry

The German Etanercept Registry is a prospective registry enrolling patients with JIA since 2001 [54, 55]. The registry contains detailed information about patients’ history, previous treatments, and concomitant medications. Patients are enrolled as they start treatment with ETN after they failed to respond to at least one disease modifying anti-rheumatic drug (DMARD), mainly MTX. Generally, regular follow-up visits for clinical assessments are required at one, three, and six months, and then every six months. For the purpose of this study we selected only patients with JIA, with the onset age of disease < 16 years (n = 521). We evaluated the probability distributions of time to response (meaning Pediatric ACR30, ACR50 and ACR70 criteria – validated measures of response to treatment in JIA [42]). These outcomes were evaluated in separate data sets.

3.2.3 Statistical analyses

Since patients are followed at discrete time intervals in RCTs, the event of interest is only known to take place between two follow-up visits, and the exact time-to-event is unknown. In this case, the response time is interval censored. For this study, we had both interval censored data and right censored data (patients lost to follow-up). The censoring status was determined by whether the two values of the upper and lower time interval were present or whether the upper time interval was not available. The following variables were collected for each patient: treatment group, response criteria, response status, and response time (interval or right censored).

We analyzed data using both non-interval and interval censoring methods for the German Etanercept Registry and only non-interval censoring method for the TEMPO trial (because of frequent and regular visits). For non-interval censoring, we treated the first time on the record for each response criteria as the exact time to event; subjects were right censored if the trial ended before a response or the patient dropped out. We estimated the survival function of time to treatment response by the Kaplan-Meier method where the survival probability indicated the probability of not responding and the hazard rate indicated the hazard of responding. For
interval censoring we used two estimating techniques; we estimated the time to response using an interval censoring technique known as the Turnbull algorithm and also by simply using interval midpoints [57].

After obtaining the non-parametric maximum likelihood estimator, a series of parametric distributions were explored to identify an appropriate distribution form which could adequately describe the data. We evaluated the parametric distributions of survival times that we thought would be most appropriate for this patient population, i.e. the exponential, Weibull, lognormal, log-logistic and generalized gamma distributions. Goodness-of-fit of parametric survival models were estimated using a graphical check – by plotting the observed response times and model predicted times by the Kaplan-Meier survival probability estimates – as well as by calculating the Akaike information criterion (AIC) for each model [49, 58]. For each model $i$ we calculated $\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}}$, where $\text{AIC}_{\text{min}}$ is the AIC for the “best” model, i.e. the model with the lowest AIC value. The models having $\Delta_i \leq 2$ were considered as suggesting substantial evidence for fit, $4 \leq \Delta_i \leq 7$ as suggesting considerably less support, and $\Delta_i > 10$ as basically no support [58, 59]. Analyses were conducted using SAS 9.1 (SAS Institute Inc., Cary, NC, USA), Stata 8.2 (Stata Corporation, TX, USA) and R 2.6.1 (The R Foundation for Statistical Computing) statistical software.

The study protocol was approved by the Research Ethics Board at the Hospital for Sick Children, Toronto, Canada and by corresponding stakeholders of the primary data. The comparative effectiveness of the studied drugs was not assessed in any way as it was not the focus of this study.

3.3 Results

3.3.1 TEMPO trial

Patient characteristics of the TEMPO trial have been presented elsewhere [53]. Parametric estimates for the various survival distributions by treatment group and outcome are given in Table 1. Time in the analysis was measured in weeks. For MTX, the lognormal distribution was considered to be the best model for the outcome of ACR20, lognormal, generalized gamma, and log-logistic distributions for ACR50, and lognormal and generalized gamma for ACR70 (all $\Delta_i \leq 2$). Our results suggested that for treatment with ETN the lognormal model was best
for ACR20, the generalized gamma for ACR 50, and both lognormal and generalized gamma
distributions for ACR70. For Combined treatment, the best model fit was under the log-logistic
distribution for ACR20, generalized gamma for ACR50, and both lognormal and generalized
gamma distributions for ACR 70. The difference between the observed and model predicted
survival probability estimates for Combined treatment and ACR20 is given in Figure 1. The
Weibull distribution was not significantly different from the exponential distribution in this
outcome and was not depicted. The change in hazard over time under the lognormal
distribution for all three treatment groups and ACR20 is given in Figure 2. The hazard of
response to treatment increased at the start of the treatment towards a peak (four weeks) and
monotonically decreased to the initial level by the end of the first year of follow-up (52 weeks).

3.3.2 German Etanercept Registry
The characteristics of patients included in the analysis of the German Etanercept Registry data
are given in Table 2. In this data set, most of the patients were females (67.6%) predominantly
with polyarticular JIA. Mean age of therapy start was about 12 years, and mean disease
duration before therapy start was 4.8 years. Most of the patients received concomitant
medications at baseline and / or after enrolment including non-steroidal antiinflammatories
(NSAIDs) (n=456, 87.5%), corticosteroids (n=282, 54.1%), and MTX (n=431, 82.9%).

We calculated non-parametric survival probabilities using three different censoring approaches
as described previously (non-interval censoring, interval censoring by Turnbull algorithm and
midpoints). We found that the non-parametric survival probability estimates were quite similar
with both interval censoring methods. However, these estimates were different than when
interval censoring was ignored. For example, the non-parametric, median survival time for the
ACR30 outcome was 8.7 weeks with no interval censoring and 4.3 weeks when using
midpoints. We fitted parametric distribution models using both approaches (Table 3). For
simplicity, we report the results for interval censoring by midpoints. Convergence was not
achieved when we tried to fit the data using the generalized gamma distribution. We applied
different likelihood maximization techniques available from the Stata streg function, and we
also changed the initial values for parameters. However, these techniques were not successful
and after the investigation of the smoothed, non-parametric hazard function we concluded that
the generalized gamma distribution was not reasonable for this data set.
When ignoring interval censoring the lognormal distribution is the best fit for all three outcomes, as shown in Table 3. When we used interval censoring by midpoints, the log-logistic distribution was the best for ACR30 and the lognormal distribution for ACR50 and ACR70. The parameters of the fitted distributions were different when estimated according to the two censoring approaches indicating that ignoring interval censoring was more conservative in estimating the response times compared to the interval censoring approach with midpoints. For example, the median response time under the lognormal distribution was about nine weeks for ACR30 without interval censoring and five weeks with interval censoring by midpoints.

The parametric (ignoring interval censoring) and non-parametric probability estimates for achieving an ACR30 response are given in Figure 3. Almost all models underestimate the probability of response before 24 weeks and overestimate it after 48 weeks.

### 3.4 Discussion

RCTs are still considered the best methodology for providing evidence to inform decisions regarding the adoption of new treatments. However, the increasing costs of conducting RCTs and difficulties in recruiting patients with rare diseases require that we revisit the accepted methods and look for methodologies that can maximize the efficiency and the information of trials [60]. The investigation of probability distributions of drug response time provides additional information about drug efficacy/effectiveness and improves the planning of future RCTs. This study aimed to describe the parametric survival distributions of widely used medications in RA and JIA. Parametric distributions are able to show the change of hazard over time while the patient is on treatment. Our analyses were not directed towards between-drug comparisons. Instead, we aimed to understand the hazard function (shape) that describes both the frequency and timing of drug response events.

The results of the original TEMPO trial showed that the Combined treatment was superior to ETN or MTX alone in both clinical and radiographic primary endpoints. The study was not designed to evaluate the outcomes using survival analysis, and the sample size calculation was based on finding pairwise differences between groups in comparing the area under the curve of the numeric index of the ACR response over the first 24-weeks [53]. The trial also compared
the proportions of patients with ACR20, ACR50, and ACR70 response criteria at 52-weeks. The results of our study suggested that, at least in our sample, the hazard of responding could be best explained by a lognormal or generalized gamma distribution and not by an exponential distribution. Taking the fitted lognormal distribution as an example, the hazard for response increases from the time of treatment start until it reaches a plateau; the hazard then falls monotonically. Median response times can be easily calculated for all treatments and outcomes using the reported distribution parameters (under the lognormal distribution, for example, it is equal to the exponent of μ).

In the German Etanercept Registry, the treatment response time followed a lognormal distribution for most outcomes. It was interesting to observe the effect of interval censoring on the evaluation of distribution fit. For the ACR 30, the best distribution that fit the observed data was log-logistic when interval censoring was considered and lognormal when interval censoring was ignored. Another important observation was the difference in estimated model parameters between these two approaches. While interval censoring did not have a major role in the TEMPO trial with frequent follow-up visits, this was not true for the Registry data with unequal and relatively infrequent visits. In a rare disease such as JIA, different methodological approaches should be used to maximize the information that can be drawn from the data. It is for that purpose that we present the results from both censoring approaches without any direct comparison between the results.

The identification and description of the appropriate parametric model is important for the understanding of disease-drug interactions, patient management, study planning, and accurate data analysis. A detailed review of the parametric survival distributions and their applications in different disease conditions is presented by Lee and Wang [49]. The exponential distribution, for example, has a constant hazard rate over time. The Weibull distribution has wider application and can model increasing, decreasing, and constant risk. The lognormal distribution is suitable for modelling hazard rates that are initially increasing to a maximum and then decreasing. This distribution was best to fit our observed data in the majority of evaluations in our analyses. Survival analyses are widely used in medical research and, particularly, in RCTs. We conducted a systematic review of abstracts in Ovid MEDLINE using the MeSH term “randomized controlled trial” and limiting our search to human studies; we found that 21%
(128/611) of all RCTs published from 2007 to 2008 in 4 major medical journals (New England Journal of Medicine, Lancet, Journal of the American Medical Association, and British Medical Journal) used time-to-event analysis to report the primary outcomes. It is unclear, though, how often different time-to-event distributions were considered when planning these studies. Our study demonstrated that in chronic conditions such as arthritis treated with highly effective medications, the assumption of response time following the exponential distribution may not be valid. The impact of the underlying survival distribution on sample size requirement and study power can be shown by computer simulations. However, this was not the main focus of this study, and should be the focus of subsequent investigation.

Our results should be interpreted in the context of potential study limitations. We were able to obtain and investigate only two data sources, and the results are limited to the patient population and treatment methods applied in these studies. It is possible, however, that under different disease and treatment conditions the best fitting distribution would be different. We also limited our choice to probability distributions that we thought as appropriate for these patient populations and did not try to fit data into other distributions such as piecewise-exponential or Gompertz models. On the other hand, these last two distributions are extensions of the exponential distribution, and, as the analysis showed, the exponential distribution was not suitable to fit the observed data. Another potential limitation is the fact that we used the data from the German Etanercept Registry which did not have frequent patient follow-up visits. We tried to overcome this limitation by presenting the findings with and without interval censoring.

Information about the parametric time to response distribution for specific treatments and specific populations is a valuable tool for patient treatment decisions and should be considered in the study planning phase to inform the sample size calculation and proposed statistical analyses. We evaluated the distributions of response time to treatments in patients with RA (ETN, MTX, Combined ETN+MTX) and JIA (ETN) and found that in our sample of patients the response times follow mostly lognormal or gamma distribution. Future work should explore the impact of the distribution selection on sample size calculations in time-to-event analyses.
Acknowledgements: Lusine Abrahamyan has been awarded a Canadian Arthritis Network (CAN) Graduate Student Award and Graduate Studentship from the Research Training Centre of the Hospital for Sick Children, Toronto, Canada. Brian Feldman is supported by a Canada Research Chair in Childhood Arthritis. Andrew R. Willan is funded by the Discovery Grant Program of the Natural Sciences and Engineering Research Council of Canada (grant number 44868-08).

The authors wish to thank Michele Hooper (Amgen Inc., Thousand Oaks, CA) for her assistance in planning the study.

Conflict of Interest statement:
Dr. Horneff received fees for invited presentations and for participation at advisory committees of the pharmaceutical companies Wyeth, Abbott, Bristol-Myers-Sqibb, Roche, Novartis, and Genzyme. Dr. Keystone has conducted research and has served as a Consultant or a member of the Advisory Board for Amgen and other companies that make products for use in patients with rheumatoid arthritis.

What is new?

- In most widely accepted sample size calculators for survival analysis it is assumed that the response time follows an exponential distribution. This assumption is made for simplicity and is rarely tested. This study showed that the exponential distribution may not always be the best distribution for patients with rheumatic diseases.
- In the study sample of patients with rheumatic diseases treated with MTX and biologic therapies the response time to treatments follows mostly a lognormal or gamma distribution.
- Careful consideration of response distribution might make study designs more efficient. Future investigations will determine the impact of distribution selection on sample size calculations for clinical trials.
Table 1. Parametric estimates of survival distributions: TEMPO trial

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Parameters</th>
<th>AIC, Δ</th>
<th>Parameters</th>
<th>AIC, Δ</th>
<th>Parameters</th>
<th>AIC, Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACR20</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential</td>
<td>λ=0.089</td>
<td>644.1, 38.5</td>
<td>λ=0.114</td>
<td>671.1, 70.7</td>
<td>λ=0.133</td>
<td>707.0, 94.7</td>
</tr>
<tr>
<td>Weibull</td>
<td>λ=0.085, γ=1.124</td>
<td>641.0, 36.3</td>
<td>λ=0.116, γ=0.962</td>
<td>672.6, 72.2</td>
<td>λ=0.137, γ=0.935</td>
<td>707.0, 94.7</td>
</tr>
<tr>
<td>Lognormal</td>
<td>μ=2.000, σ=0.930</td>
<td>605.6, 0.0</td>
<td>μ=1.650, σ=0.960</td>
<td>600.5, 0.0</td>
<td>μ=1.490, σ=0.930</td>
<td>615.02, 2.7</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>α=0.029, γ=1.786</td>
<td>617.9, 11.2</td>
<td>α=0.062, γ=1.786</td>
<td>605.2, 4.8</td>
<td>α=0.075, γ=1.887</td>
<td>612.4, 0.0</td>
</tr>
<tr>
<td><strong>ACR50</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential</td>
<td>λ=0.026</td>
<td>575.5, 23.3</td>
<td>λ=0.035</td>
<td>641.6, 23.6</td>
<td>λ=0.042</td>
<td>702.4, 31.1</td>
</tr>
<tr>
<td>Weibull</td>
<td>λ=0.027, γ=1.190</td>
<td>571.9, 19.4</td>
<td>λ=0.035, γ=0.990</td>
<td>643.6, 25.5</td>
<td>λ=0.042, γ=0.943</td>
<td>703.3, 32.0</td>
</tr>
<tr>
<td>Lognormal</td>
<td>μ=3.230, σ=1.060</td>
<td>552.5, 0.0</td>
<td>μ=2.860, σ=1.220</td>
<td>620.6, 2.5</td>
<td>μ=2.640, σ=1.23</td>
<td>675.2, 4.0</td>
</tr>
<tr>
<td>GG</td>
<td>λ=0.044, α=-0.206, γ=-4.000</td>
<td>553.6, 1.1</td>
<td>λ=0.083, α=-0.442, γ=-1.471</td>
<td>618.0, 0.0</td>
<td>λ=0.108, α=-0.501, γ=-1.340</td>
<td>671.3, 0.0</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>α=0.005, γ=1.639</td>
<td>554.2, 1.7</td>
<td>α=0.020, γ=1.370</td>
<td>627.1, 9.1</td>
<td>α=0.029, γ=1.351</td>
<td>682.2, 10.9</td>
</tr>
<tr>
<td><strong>ACR70</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential</td>
<td>λ=0.009</td>
<td>395.9, 14.7</td>
<td>λ=0.012</td>
<td>459.3, 9.1</td>
<td>λ=0.019</td>
<td>607.1, 14.7</td>
</tr>
<tr>
<td>Weibull</td>
<td>λ=0.011, γ=1.282</td>
<td>392.8, 11.6</td>
<td>λ=0.013, γ=1.190</td>
<td>458.1, 7.9</td>
<td>λ=0.019, γ=0.980</td>
<td>609.1, 16.7</td>
</tr>
<tr>
<td>Lognormal</td>
<td>μ=4.260, σ=1.200</td>
<td>383.0, 1.8</td>
<td>μ=4.05, σ=1.260</td>
<td>450.2, 0.0</td>
<td>μ=3.53, σ=1.370</td>
<td>593.12, 0.7</td>
</tr>
<tr>
<td>GG</td>
<td>λ=0.021, α=-0.504, γ=-0.943</td>
<td>381.2, 0.0</td>
<td>λ=0.023, α=-0.328, γ=-1.471</td>
<td>450.6, 0.4</td>
<td>λ=0.041, α=-0.288, γ=-1.587</td>
<td>592.4, 0.0</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>α=0.002, γ=1.449</td>
<td>389.0, 7.8</td>
<td>α=0.004, γ=1.370</td>
<td>454.9, 4.7</td>
<td>α=0.012, γ=1.250</td>
<td>598.1, 5.7</td>
</tr>
</tbody>
</table>

In the exponential distribution λ is the hazard rate; in the Weibull distribution λ is the scale and γ is the shape parameter; in the lognormal distribution μ is the mean and σ² is the variance of the log survival time; in the GG (generalized gamma distribution) γ is the shape, λ is the scale, and α is the locator parameter; in the log-logistic distribution α is the scale and γ is the shape parameter.

GG: generalized gamma distribution

AIC: Akaike information criteria was calculated by -2(log-likelihood) + 2K, where K is the number of estimated parameters in the model (K=1 for the exponential, K=2 for Weibull, lognormal, and log-logistic distributions, K=3 for generalized gamma distribution). Δ_i = AIC_i-minAIC

* Convergence was not achieved under the generalized gamma distribution for ACR20.
Table 2. German Etanercept Registry patient characteristics*

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>N = 521</th>
<th>Characteristics</th>
<th>N = 521</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>JIA subtypes (n, %)</strong></td>
<td></td>
<td><strong>Weight at the start, kg</strong></td>
<td>42.0 (18.0)</td>
</tr>
<tr>
<td>Systemic Arthritis</td>
<td>57 (10.9)</td>
<td><strong>Height at the start, cm</strong></td>
<td>145 (22.8)</td>
</tr>
<tr>
<td>Seronegative polyarticular JIA</td>
<td>158 (30.3)</td>
<td><strong>CHAQ at the 1st visit</strong></td>
<td>0.93 (1.79)</td>
</tr>
<tr>
<td>Seropositive polyarticular JIA</td>
<td>59 (11.3)</td>
<td><strong>Age at diagnosis, years</strong></td>
<td>7.9 (4.5)</td>
</tr>
<tr>
<td>Persistant oligoarticular JIA</td>
<td>30 (5.8)</td>
<td><strong>Age at disease onset, years</strong></td>
<td>7.3 (4.5)</td>
</tr>
<tr>
<td>Extended polyarticular JIA</td>
<td>80 (15.4)</td>
<td><strong>Age at therapy start, years</strong></td>
<td>12.1 (4.3)</td>
</tr>
<tr>
<td>Enthesis related arthritis</td>
<td>67 (12.9)</td>
<td><strong>Response achieved (n, %)</strong></td>
<td></td>
</tr>
<tr>
<td>Psoriasis and arthritis</td>
<td>43 (8.3)</td>
<td>ACR30</td>
<td>486 (93.3)</td>
</tr>
<tr>
<td>Unclassified JIA</td>
<td>27 (5.2)</td>
<td>ACR50</td>
<td>461 (88.5)</td>
</tr>
<tr>
<td><strong>Female gender (n, %)</strong></td>
<td>352 (67.6)</td>
<td>ACR70</td>
<td>399 (71.6)</td>
</tr>
<tr>
<td><strong>Enbrel start dose, mg/kg</strong></td>
<td>0.5 (0.23)</td>
<td><strong>Follow-up length, weeks</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median (IQR)</td>
<td>61 (39-78)</td>
</tr>
</tbody>
</table>

*All variables are presented as mean (standard deviation) unless specified.

N/A, not applicable; CHAQ, Childhood Health Assessment Questionnaire; IQR, Interquartile range, NSAID, non-steroidal anti-inflammatory drugs.
Table 3. Parametric estimates of survival distributions: German Etanercept Registry

<table>
<thead>
<tr>
<th>Distribution*</th>
<th>Non-interval censoring</th>
<th>Interval censoring (midpoints)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AIC, $\Delta_i$</td>
<td>Parameters</td>
</tr>
<tr>
<td>Exponential</td>
<td>$\lambda = 0.070$</td>
<td>1522.6, 169</td>
</tr>
<tr>
<td>Weibull</td>
<td>$\lambda = 0.070, \gamma = 1.002$</td>
<td>1524.6, 171</td>
</tr>
<tr>
<td>Lognormal</td>
<td>$\mu = 2.181, \sigma = 0.924$</td>
<td>1353.6, 0.0</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>$\alpha = 0.022, \gamma = 1.849$</td>
<td>1365.0, 11.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distribution*</th>
<th>Non-interval censoring</th>
<th>Interval censoring (midpoints)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AIC, $\Delta_i$</td>
<td>Parameters</td>
</tr>
<tr>
<td>Exponential</td>
<td>$\lambda = 0.049$</td>
<td>1601.0, 124.9</td>
</tr>
<tr>
<td>Weibull</td>
<td>$\lambda = 0.049, \gamma = 1.061$</td>
<td>1600.2, 124.1</td>
</tr>
<tr>
<td>Lognormal</td>
<td>$\mu = 2.481, \sigma = 1.071$</td>
<td>1476.1, 0.0</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>$\alpha = 0.024, \gamma = 1.548$</td>
<td>1499.4, 23.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distribution*</th>
<th>Non-interval censoring</th>
<th>Interval censoring (midpoints)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AIC, $\Delta_i$</td>
<td>Parameters</td>
</tr>
<tr>
<td>Exponential</td>
<td>$\lambda = 0.029$</td>
<td>1549.7, 62.5</td>
</tr>
<tr>
<td>Weibull</td>
<td>$\lambda = 0.030, \gamma = 1.030$</td>
<td>1551.2, 64</td>
</tr>
<tr>
<td>Lognormal</td>
<td>$\mu = 3.018, \sigma = 1.194$</td>
<td>1487.2, 0.0</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>$\alpha = 0.016, \gamma = 1.377$</td>
<td>1511.3, 24.1</td>
</tr>
</tbody>
</table>

In the exponential distribution $\lambda$ is the hazard rate; in the Weibull distribution $\lambda$ is the scale and $\gamma$ is the shape parameter; in the lognormal distribution $\mu$ is the mean and $\sigma^2$ is the variance of the log survival time; in the log-logistic distribution $\alpha$ is the scale and $\gamma$ is the shape parameter.

AIC: Akaike Information Criteria was calculated by $-2(\text{log-likelihood}) + 2K$, where $K$ is the number of estimated parameters in the model ($K=1$ for the exponential, $K=2$ for Weibull, lognormal, and log-logistic distributions). $\Delta_i = \text{AIC}_i - \text{minAIC}$

* Convergence was not achieved under the generalized gamma distribution.
Figure 1. Parametric and non-parametric estimates for probability of response to Combined treatment (ACR20, TEMPO trial)
Figure 2. Hazard function under the lognormal distribution for ACR20 for the three treatment groups (TEMPO trial)
Figure 3. Parametric and non-parametric estimates for probability of response for ACR30 (German Etanercept Registry, ignoring interval censoring)
Chapter 4. The Impact of Survival Distributions on the Power of Randomized Clinical Trials

The article has been accepted for publication in the Journal of Clinical Epidemiology.

THE IMPACT OF SURVIVAL DISTRIBUTIONS ON THE POWER OF RANDOMIZED CLINICAL TRIALS

L Abrahamyan, CS Li, J Beyene, AR Willan, BM Feldman
Lusine Abrahamyan MD MPH\textsuperscript{1,3}, Chuan Silvia Li\textsuperscript{1,4}, Joseph Beyene MSc PhD\textsuperscript{1,3,5},
Andrew R. Willan PhD\textsuperscript{1,5}, Brian M. Feldman MD MSc FRCPC\textsuperscript{1,2,3,5}

\textsuperscript{1}Child Health Evaluative Sciences, Research Institute, The Hospital for Sick Children,
\textsuperscript{2}Departments of Paediatrics, \textsuperscript{3}Health Policy Management & Evaluation, \textsuperscript{4}Pharmacology, \textsuperscript{5}Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

Corresponding author: Brian M. Feldman
Tel: +1 416 813 6117; fax: +1 416 813 4989; e-mail: brian.feldman@sickkids.ca

Address for reprint requests: The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada

Word Count: 2950

Running Title: The power of RCTs under different survival distributions

Key words: randomized clinical trial design, randomized placebo-phase design, power, survival distributions, clinical trial simulations
Abstract

Objective: This study evaluated the power of the Randomized Placebo-Phase Design (RPPD) – a new design of randomized clinical trials (RCTs), compared to the traditional parallel groups design, assuming various response time distributions. In the RPPD, at some point, all subjects receive the experimental therapy and the exposure to placebo is for only a short, fixed period of time.

Study design and Setting: For the study, an object-oriented simulation program was written in R. The power of the simulated trials was evaluated using six scenarios where the treatment response times followed the exponential, Weibull or lognormal distributions. The median response time was assumed to be 355 days for the placebo and 42 days for the experimental drug.

Results: Based on the simulation results, the sample size requirements to achieve the same level of power were different under different response time to treatment distributions. The scenario where the response times followed the exponential distribution had the highest sample size requirement. In most scenarios, the parallel groups RCT had higher power compared to the RPPD.

Conclusion: The sample size requirement varies depending on the underlying hazard distribution. The RPPD requires more subjects to achieve a similar power to the parallel groups design.

Abstract word count: 199

Running Title: The power of RCTs under different response time distributions

Key words: randomized clinical trial design, randomized placebo-phase design, power, sample size, survival distributions, clinical trial simulation
4.1 Introduction
There are several difficulties in conducting RCTs, especially when studying rare diseases. The main difficulty in the case of rare disease is low patient accrual due to the low disease prevalence [7]. A recent review of the quality of RCTs in Juvenile Idiopathic Arthritis (JIA – an uncommon illness affecting children) found that more than half of all RCTs conducted from 1976 to 2006 (n=52) had sample sizes of fewer than 50 patients, and about 50% of the trials failed to show a difference between the experimental and standard treatments [61]. Some RCTs are stopped early because of low enrolment, subjecting participants to potentially risky interventions with no guarantee that their participation will lead to results of scientific value [14, 62]. In rare diseases, the variable and sometimes unpredictable disease progression along with lack of established objective outcome measures is another obstacle in enrolling patients in RCTs – where usually a more homogenous patient population and objective outcomes are preferred [1, 15, 16]. High costs of drug development and treatment are other characteristics of rare diseases creating debates around the reimbursement issues [3, 4, 13, 63]. Because of these difficulties, RCTs in rare diseases may suffer from poor quality and may be less efficient when compared to trials in more common diseases.

When there is a strong preference for one of the study arms, the parallel groups RCT design may not be acceptable, nor the preferred design for clinical research. Subjects’ strong preferences for one of the treatment choices in an RCT is one of the important factors playing a role in enrolment decisions, and can strongly affect accrual success [64-66]. In order to increase study power and acceptability, other designs have been implemented as alternatives to parallel groups RCTs. Most of these designs, however, are only applicable when specific conditions of the patient population, disease progress model or medications are met and therefore they are not widely applicable. For example, cross-over trials are more suitable for patients with chronic conditions, and drugs that do not introduce disease modifying changes. N-of-1 trials enroll single patients and assign individual treatment plans [15]. The randomized withdrawal design enrolls only those patients who respond to the experimental treatment in the initial open-label phase [18].

The randomized placebo-phase design (RPPD) is a relatively new trial design that was developed with the primary objective of increasing accrual in difficult to study areas, such as
the study of therapy for rare disease. The validity of the design and investigation of the power of the design was first published in 2001 [21]. In an RPPD trial, patients are randomly assigned to different starting times for experimental therapy at study enrollment. Some patients receive the experimental treatment at the outset and some receive a placebo treatment for a short period of time after which they receive the experimental treatment (Figure 1). Thus, the main difference between the RPPD and the parallel groups RCT is that in the RPPD at some point all patients receive the experimental therapy, and the exposure to placebo is for only a short, fixed period of time. It is assumed that this characteristic will make RPPD more acceptable and preferable to patients and physicians, especially in cases where there is a preference for the experimental therapy [21].

The main assumption of the RPPD is that if the treatment is effective, those who receive the treatment sooner will respond to it sooner, on average. The outcome is a time-to-event variable assumed to follow one of the so-called ‘survival distributions’. A previous study assessed the power of the RPPD under the assumption of an exponential distribution describing treatment response times [21]. The RPPD was compared with the parallel groups RCT in terms of its sample size and power when a high, moderate, intermediate or low potency drug was assigned to patients with childhood arthritis. The results showed that for the same level of power the RPPD generally requires more subjects than the parallel groups design. It also showed that a fixed placebo phase (in which subjects were assigned to either no placebo phase, or the maximally acceptable placebo phase length – in order to maximize variance) was more powerful than designs where the placebo phase length was allowed to be more variable. The RPPD was relatively more efficient for high potency drugs [21].

The assumption of treatment response times following an exponential distribution has rarely been tested. However, the most widely accepted sample size calculators for survival time analysis assume that the response times follow an exponential distribution [50, 51]. We have recently studied the response time distributions for patients with rheumatoid arthritis (RA) and juvenile idiopathic arthritis (JIA) treated with methotrexate and/or etanercept (In Press in the Journal of Clinical Epidemiology). Our study found that the response time to treatment followed a lognormal distribution in the majority of comparisons, less often a generalized
gamma or log-logistic distribution; the exponential distribution described treatment responses relatively less well.

The objective of this study was to evaluate the statistical efficiency of the RPPD compared to the parallel groups RCT design assuming various response time distributions, using Monte Carlo computer simulations. This study will help to define the scope of applicability of the RPPD.

4.2 Materials and methods

We used Monte Carlo computer simulations to describe the performance characteristics of the RPPD – as compared to the parallel groups RCT – under a variety of scenarios.

The important attributes of any computer simulation include model parameters (constants such as total trial or follow-up duration), assumptions (taken from real-world systems), input variables (such as sample size), algorithms that mimic the process of the trial, and outputs (for example, proportion of false positive and false negative trials for a set of any given input structures) [67, 68]. For our study purposes, an object-oriented, discrete-time simulation program was developed using R statistical software (Version 2.6.1, The R Foundation for Statistical Computing). The full code, along with further explanations is presented in the Appendix 3. A short description of the design parameters specified for the simulations is presented in the Appendix of this chapter. All simulations were carried out using the high performance computer at the Centre for Computational Biology at the Hospital for Sick Children, Toronto, Canada. Each scenario was simulated 10 000 times. The simulation program was validated against two available sample size calculators: PS – power and sample size calculator developed by Dupont and Plummer [69] and SSP – an internet-based power and sample size calculator which uses a method described by Lachin and Folkes [51, 70]. Results are presented in the Appendix 4.

4.2.1 Simulation design and parameters

In the RPPD, subjects were randomized to start the experimental therapy or placebo at study entry. Placebo subjects were switched to the experimental therapy after a defined placebo phase duration (45 days or 60 days in simulations). For the parallel groups RCT design, subjects were
randomized into the experimental or control groups and received the assigned treatment for the entire trial duration. All other parameters were identical between the two designs.

It was assumed that even under the placebo treatment there might be a spontaneous response; the placebo group (and in the RPPD, subjects under the placebo phase) were assigned a median response time of 355 days. This corresponds to a 30% 'placebo response' at 6 months (under the exponential distribution) – which is commonly seen in clinical trials in chronic diseases [71]. The experimental drug was considered to be of intermediate potency with a median response time of 42 days (similar to some disease modifying drugs or biologic therapies in diseases like rheumatoid arthritis). These median times were used to calculate the parameters of the different survival distributions for the simulation scenarios.

Subjects entered into the study during 90 days of accrual time. Entry followed a uniform distribution ranging from 0 to the maximum length of the accrual phase. After enrollment, for the RPPD, subjects were randomly allocated with a 50% probability to starting experimental therapy at enrolment for a duration of 275 days or to a placebo phase (which under different simulations was set to either 45 days or 60 days) followed by the experimental therapy (275 days in total). The total maximum trial duration of each trial was 365 days – 90 days for accrual and 275 days for patient follow-up. To investigate the effect of the placebo phase duration on the study power, we repeated all scenarios for the RPPD with 45 days and 60 days of placebo phase. To better simulate a real life clinical trial, subjects’ evaluations for response status were done at discrete follow-up times - at every two weeks after the enrollment until the patient responded or the trial ended.

Each simulation run had 10 000 iterations (or 10 000 trials for each evaluated sample size). The randomly generated randomization seed was saved after each run for possibility of future replications.

**4.2.2 Simulation outcomes (output variables)**

In this simulation project, the unit of replication was each clinical trial and the outcome of the simulation experiment was based on the analysis of the outcome measures obtained from all clinical trial replications. For each clinical trial in our simulations, the outcome for each subject was the time from accrual to the time a response occurred or the study ended (subject
censored), whichever happened first. The trial outcome was evaluated by the Cox proportional hazards model. At the end of each simulation scenario we evaluated whether, on average, there was a significant difference between the survival times of the two treatment groups. For the parallel groups RCT this was achieved by taking the proportion of trials where the experimental treatment was effective and the likelihood ratio statistic from the Cox proportional hazards model was greater or equal to the critical value at alpha = 0.05 in the two-sided and one-sided chi-square distribution with one degree of freedom. For the RPPD a one-sided test was applied - the null hypothesis stated that there is no difference between the treatment groups, and the alternative hypothesis stated that on average, subjects who received the treatment sooner (experimental subjects) responded sooner. The power of the RPPD was defined by the proportion of trials where the experimental treatment had shorter observed response time and the likelihood ratio statistic was significant at one-sided alpha = 0.05 of the chi-square distribution with one degree of freedom.

4.2.3 Assumptions
The RPPD is prone to the same threats to validity as other randomized clinical trials. For the purpose of this study, we assumed that subjects’ randomization, allocation assignment / concealment, and masking (i.e. double masking) were adequate in all trials. We assigned 100% accrual probability and 10% random patient drop-out probability to all scenarios. Subjects who dropped out, and those not responding to the treatment by the end of the study were censored. We also assumed that there were no patient crossovers, no contamination and full compliance with treatment protocols. The analysis for each trial used the intent-to-treat principle.

4.2.4 Simulation scenarios
To investigate the effect of baseline response time distributions on the power of the studied designs, we planned and conducted six simulation scenarios with different survival distributions and varying sample sizes (Table 1). The first four scenarios (Scenario 1 - 4) assumed that the baseline hazard (i.e. the spontaneous response seen on placebo) followed the exponential distribution – which may be true for chronic diseases such as arthritis – and the response time for the experimental group followed an exponential, Weibull or lognormal distribution. Scenario 5 assigned a lognormal distribution and Scenario 6 assigned a Weibull
distribution (with different increasing hazard shapes) to both placebo and experimental treatment groups / phases.

4.3 Results
The results of the simulations for Scenarios 1-6, in terms of sample size requirements to reach 80% or 90% power to detect statistically significant differences between the treatments, are presented in Table 2. The results for the RPPD and one-sided parallel groups RCT are presented graphically in Figure 2.

The parallel group RCT had lower sample size requirements in most of studied scenarios when compared with the RPPD. In two scenarios in which we simulated a Weibull distribution for time to response (Scenario 3 and 6), and 60 days of placebo phase duration, the RPPD and the two-sided parallel groups RCT had similar power. In all scenarios the RPPD design with 60 days placebo phase duration had higher power compared with the RPPD that had a placebo phase duration of 45 days.

4.4 Discussion
Investigation of the underlying survival distribution for a particular patient population and medication is not widely applied when planning RCTs where the outcome is evaluated as time-to-event data. The impact of the underlying survival distribution on study power is unclear. The results of our study show that under different survival distributions the sample size requirements can be quite different for the same level of power. The highest sample size requirement was observed in the scenario where the response times followed an exponential distribution. In contrast, the smallest sample size requirements were observed for a Weibull distribution. The majority of available sample size calculators assume an underlying exponential distribution, and ignorance of the underlying hazard distribution can provide inadequate power or can over-power planned studies. However, if the underlying survival distribution is not exponential, the consideration of the exact distribution can improve the trial efficiency and decrease the societal opportunity cost of the trial (e.g. total trial duration, financial costs and number of patents enrolled in the study). These factors are especially important in rare diseases where adequate patient enrollment is one of the challenges, and drugs for these diseases are often expensive to develop [4].
The study results showed also that in the majority of simulation scenarios the parallel groups RCT required fewer subjects than the RPPD. Considering the structure of the designs - all subjects in the control group received the placebo treatment for the total duration of trial in the parallel groups RCT and for only 45 or 60 days in the RPPD, this was an expected finding [21]. Surprisingly, in two scenarios with a Weibull distribution and 60 days of placebo phase duration, we observed similar levels of power between the RPPD with 60 days of placebo phase duration and the two-sided, parallel groups RCT. This indicates that for diseases with increasing hazard (i.e. in progressive conditions), the RPPD can be used with similar efficiency as the parallel groups design RCT. We believe that the parallel groups RCT with a one-sided test should have a wider application in clinical effectiveness studies; however, the majority of currently conducted parallel RCTs use a two-sided hypothesis. The RPPD is one-sided by definition, and for all scenarios it is therefore relevant to compare the RPPD with a two-sided parallel groups RCT as illustrated in Figure 2.

The power of the RPPD, as shown by the simulations, can be altered by changing the placebo phase duration; in the simulations the RPPD with 60 days of placebo phase duration was more efficient than the RPPD with 45 days of placebo phase duration. The advantage of the RPPD design over the parallel-design is that in the RPPD all patients receive the experimental drug during the trial – half of them after being in the placebo phase for a short time. However, although increasing the placebo phase duration can be appealing, there are ethical issues in enrolling patients into the placebo phase for a longer time than is necessary (in our scenarios, the 60 days placebo phase duration was longer than the median response time). The parallel groups RCT will require a smaller sample size in most cases but half of the patients will not get the experimental treatment during the trial. Thus, in selecting the trial design we should consider trade-offs between patient preferences for treatment allocation, design acceptability, study feasibility and overall study sample size.

The optimal placebo phase duration for the RPPD design should be decided individually for each clinical trial scenario. The factors that should be considered for defining the optimal placebo phase duration include the disease progression model (acute versus chronic conditions), drug potency (high or low potency, disease modifying or not), response time to
drug distributions, and the effect size. Using the simulation program we have developed, investigators can evaluate the power of the RPPD and the parallel design under various clinical scenarios. The list of the clinical trial design parameters that can be specified in the program is provided in the Appendix.

Although we considered several components of the RCT design, it is possible that we did not include all factors that may impact the study results. For example, study design acceptability is an important factor for adequate patient enrollment. We did not consider this parameter in our simulations because no formal study has compared these two study designs in the past. In our simulations we made several assumptions about trial designs such as randomization, allocation assignment, masking, accrual probability, drop-out probability, patient crossovers, contamination and compliance which may not be true in all real life situations. As with all simulation studies, we chose a limited number of parameters to vary in order to make the number of simulations manageable. The impact of additional factors (separately or in combination) can be investigated in the future using the simulation program that we have provided.

In conclusion, we have demonstrated that the sample size requirements for randomized clinical trials with time-to-event outcomes can vary for different treatment response hazard distributions. This finding was observed for both the RPPD and the parallel groups RCT design. In most of the evaluated scenarios the RPPD required more subjects to achieve similar power to the parallel design. The difference in the sample size requirement between the designs can be decreased by prolonging the placebo phase duration of the RPPD design.

Appendix: The simulation program

For this study we developed a discrete time, object-oriented Monte Carlo simulation program using R statistical software (Version 2.6.1, The R Foundation for Statistical Computing). The program allows to define several important clinical trial parameters to calculate the power of the parallel design RCT and the RPPD for different sample sizes when the outcome is a time-to-event variable. The following parameters of the clinical trials design can be specified in the program:
• **Trial design:** defined as ‘RPPD’ or ‘PARALLEL’;

• **Survival distributions:** the distribution of the time to response for the experimental or the placebo treatment specified as ‘exponential’, ‘Weibull’, or ‘Lognormal’;

• **Survival distribution parameters:** specified for the placebo and experimental treatment separately; ‘lambda’ for the exponential distribution, and ‘shape’ and ‘scale’ for the ‘Weibull’ or lognormal distribution;

• **Sample size:** defined as a total sample of subjects enrolled in the trials;

• **Study duration parameters:** specified as accrual phase duration; follow-up visit interval duration or ‘cycle time’ to evaluate the response to treatment; placebo phase duration and experimental treatment duration;

• **Number of trials:** corresponds to the number of simulations/iterations;

• **Accrual probability:** specified separately for the RPPD or the parallel groups RCT design;

• **Drop-out probability:** specified separately for each treatment arm.

The full code of the “**SIMULATION-BASED SAMPLE SIZE CALCULATOR FOR THE RPPD**” with accompanying explanations is available on the website of the Child Health Evaluative Sciences program at the Research Institute of The Hospital for Sick Children: [http://www.sickkids.ca/Research/Child-Health-Evaluative-Sciences/Lab-and-research-project-sites/index.html](http://www.sickkids.ca/Research/Child-Health-Evaluative-Sciences/Lab-and-research-project-sites/index.html)
Acknowledgements: Lusine Abrahamyan has been awarded a Canadian Arthritis Network (CAN) Graduate Student Award and Graduate Studentship from the Research Training Centre of the Hospital for Sick Children, Toronto, Canada. Brian Feldman is supported by a Canada Research Chair in Childhood Arthritis. Andrew R. Willan is funded by the Discovery Grant Program of the Natural Sciences and Engineering Research Council of Canada (grant number 44868-08).

What is new?

- The sample size requirements for randomized clinical trials (RCT) with time-to-event outcomes can vary significantly for different treatment response hazard distributions. This finding was demonstrated by Monte Carlo computer simulations for two RCT designs – the traditional parallel groups and the Randomized Placebo-Phase Design (RPPD).
- The RPPD is assumed to be more acceptable to patients and physicians because in this design at some point all subjects receive the experimental therapy. Simulations showed that although in the majority of scenarios the RPPD required more subjects to achieve a power similar to the parallel groups design, the difference was small for some scenarios.
- The consideration of the exact distribution can improve the trial efficiency and decrease the societal opportunity cost of the trial (e.g. total trial duration, financial costs and number of patents enrolled in the study).
- The discrete time, object-oriented simulation program written for this study allows us to define several important clinical trial parameters to calculate the power of the parallel design RCT and the RPPD for different sample sizes when the outcome is a time-to-event variable following exponential, Weibull or lognormal distribution.
<table>
<thead>
<tr>
<th>Simulation scenario</th>
<th>Placebo phase/group</th>
<th>Experimental phase/group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Distribution</td>
<td>Parameter(s)</td>
</tr>
<tr>
<td>1</td>
<td>Exponential</td>
<td>$\lambda = 0.00195$</td>
</tr>
<tr>
<td>2</td>
<td>Exponential</td>
<td>$\lambda = 0.00195$</td>
</tr>
<tr>
<td>3</td>
<td>Exponential</td>
<td>$\lambda = 0.00195$</td>
</tr>
<tr>
<td>4</td>
<td>Exponential</td>
<td>$\lambda = 0.00195$</td>
</tr>
<tr>
<td>5</td>
<td>Lognormal</td>
<td>$\mu = 5.87$, $\sigma = 0.95$</td>
</tr>
<tr>
<td>6</td>
<td>Weibull 3</td>
<td>$\lambda = 0.00243$, $\gamma = 2.5$</td>
</tr>
</tbody>
</table>

*Weibull 1, 2, 3: different shapes of increasing hazard*
Table 2. Total sample size required to achieve statistical significance (alpha < 0.05)

| Simulation | Power 80% | | Power 90% | |
|------------|-----------|----------------|-----------|
|            | RPPD*     | Parallel RCT   | RPPD*     | Parallel RCT |
|            | 45 days   | 60 days 1-sided | 2-sided | 45 days   | 60 days 1-sided | 2-sided |
| 1          | 70        | 47        | 15        | 20        | 100       | 70        | 20        | 25        |
| 2          | 50        | 35        | 14        | 18        | 70        | 50        | 20        | 25        |
| 3          | 23        | 11        | 8         | 11        | 32        | 15        | 12        | 14        |
| 4          | 35        | 19        | 10        | 12        | 50        | 27        | 13        | 15        |
| 5          | 40        | 27        | 11        | 14        | 60        | 40        | 15        | 19        |
| 6          | 19        | 7         | 6         | 7         | 27        | 10        | 7         | 9         |

*The RPPD was simulated for two different placebo phase durations: 45 days and 60 days.*
Both designs, the randomized placebo-phase design and the parallel groups RCT, originally randomize four subjects to each treatment arm. The major difference between two designs is between the placebo arm in the RCT, and the placebo phase of the RPPD, in which subjects in the RPPD are assigned a fixed placebo phase duration after which they switch to the experimental treatment. Patients that do not respond by the end of trial, are censored.

Figure 1. Presentation of the Randomized placebo-phase design and the parallel groups RCT
Figure 2. Power of the RPPD and parallel groups design RCTs for different scenarios*

Scenario 1

Scenario 2

Scenario 3

Scenario 4
… RPPD (Placebo phase 45 days) - - - RPPD (Placebo phase 60 days) \* Parallel groups design (2-sided)

*Please, refer to Table 1 for distribution parameters and to the text for simulation parameters.*
Chapter 5. Value of Information in Hemophilia Trials

USING VALUE OF INFORMATION METHODS WHEN THE DISEASE IS RARE AND THE TREATMENT IS EXPENSIVE – THE EXAMPLE OF HEMOPHILIA TRIALS

L Abrahamyan, BM Feldman, J Beyene, AR Willan

Lusine Abrahamyan MD MPH\(^1,4\), Joseph Beyene MSc PhD\(^1,4,5\), Victor Blanchette MD FRCP FRCP\(^1,3,6\), Brian M. Feldman MD MSc FRCP\(^1,2,4,5\), Andrew R. Willan PhD\(^1,5\)

\(^1\)Child Health Evaluative Sciences, Research Institute, The Hospital for Sick Children, Departments of Paediatrics, \(^2\)Division of Rheumatology, \(^3\)Division of Hematology/Oncology, \(^4\)Health Policy Management & Evaluation, \(^5\)Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada, \(^6\)on behalf of the Canadian Dose-Escalation Primary Prophylaxis (CHPS) Study

Corresponding author: Andrew R. Willan
Tel: +1 416 813 2166; fax: +1 416 813 5979; e-mail: andy@andywillan.com
Address for reprint requests: \(^1\)Child Health Evaluative Sciences, Research Institute, The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada

Word Count: 4119

Running Title: Expected value of sample information in hemophilia trials

Key words: expected value of sample information; hemophilia; prophylaxis, factor VIII

Acknowledgements: Lusine Abrahamyan has been awarded a Canadian Arthritis Network (CAN) Graduate Student Award and Graduate Studentship from the Research Training Centre of the Hospital for Sick Children, Toronto, Canada. Brian Feldman is supported by a Canada Research Chair in Childhood Arthritis. Andrew R. Willan is funded by the Discovery Grant Program of the Natural Sciences and Engineering Research Council of Canada (grant number 44868-08).
Abstract

Objective. The research objective is to assess, by applying value of information (VOI) methods, the evidence in support of using (i) alternate day prophylaxis (AP), (ii) tailored prophylaxis (TP) or (iii) on-demand treatment (OD) with factor VIII in the initial management of hemophilia. VOI methodology – using cost-benefit in a decision analytic framework – identifies those situations where the evidence is sufficient to adopt one treatment strategy over another or, where the evidence is insufficient it provides the optimal sample size for additional research.

Methods. The net monetary benefit of each treatment modality was plotted against the threshold value for the health outcome. Two threshold values – $400,000 and $1,000,000 – were selected for comparing the treatment modalities. To apply VOI methods, the following parameters were identified: current estimate of mean and variance of incremental net benefit, incidence, patient enrollment fraction, duration of follow-up, time horizon for the decision, trial costs and discount rate.

Results. Overall, for lower threshold values (<$400,000) OD had a higher net benefit than either prophylaxis modalities, while for high threshold values (> $800,000) prophylaxis modalities had a higher net benefit. At a threshold value of $400,000, TP and OD were superior to AP, since there was sufficient evidence to adopt either one over AP. However, a new clinical trial with 42 patients per arm is needed to make an optimal choice between TP and OD. A trial of this size would yield an expected net gain of over $28 million. For a threshold value of $1,000,000, both AP and TP were superior to OD, however the evidence was insufficient to choose between AP and TP, and the optimal trial to compare the two would require 48 patients per arm, yielding an expected net gain of over $125 million.

Conclusion. VOI methods can be used to determine the relevant costs and benefits of future research. The results can help in decision making for treatment strategies in hemophilia and in planning of future studies for this condition and for other rare diseases.

Word count: 326
5.1 Introduction
There are many challenges in conducting studies in rare diseases, including low patient accrual, variable disease progression, lack of objective outcomes, and relatively high overall trial costs. Currently there are about 5,000 to 6,000 rare diseases registered worldwide together affecting about 25 million North Americans and 30 million Europeans[1]. For most rare diseases no effective cure is available.

Over recent decades, different countries have implemented special polices and regulations to support research and development of orphan and ultra-orphan drugs [6, 9, 10]. Although the existing policies facilitate the market approval of orphan drugs, the decisions for timely access and reimbursement are not universal and are taken at national, regional, state or provider level [13]. This is because most orphan drugs are expensive and do not meet currently acceptable ‘rational’ cost-effectiveness thresholds used in resource allocation decisions. Furthermore, in some cases even when cost-effectiveness is near the acceptable threshold values, there is a great deal of uncertainty around the estimates of cost-effectiveness, and a drug may receive conditional funding with the decision to revise as new evidence becomes available [3, 4].

Value of information (VOI) methodology compares the expected value of information from additional research against its expected cost and can be applied when undertaking therapeutic research, i.e. a new randomized clinical trial. VOI methods are also used to identify the sample size that maximizes expected net gain (defined as “value” minus “cost”).

In the presence of uncertainty, there is an opportunity loss associated with making a particular decision – defined as the difference between the net benefit of the best decision and the net benefit of the decision taken [72]. The prior expected opportunity loss (EOL₀) is determined by taking the expectation with respect to available current information and multiplying it by the number of patients to whom the decision applies. EOL₀ is the maximum amount a decision maker should be willing to pay to decrease the uncertainty associated with a decision [73]. The expected value of sample information (EVSI) of a trial is the reduction of expected opportunity loss it provides, and is a function of the sample size. The expected net gain (ENG) is the difference between the EVSI and the expected total costs (ETC) of the trial. The aim of the VOI exercise is to identify the sample size that maximizes ENG. If the maximum ENG is
positive then the optimum decision is to conduct a new trial. On the other hand, if the maximum ENG is negative the optimum decision is to adopt the treatment with the largest mean net benefit based on current information. This method has been previously described in the medical decision-making literature for various clinical settings [73-80].

The aim of the present study is to evaluate the use of VOI methodology in a rare disease with high treatment costs, by determining the EVSI and the ETC. For that purpose, we selected hemophilia A which is a sex-linked, rare genetic disorder with an annual incidence of 1 in 5000 live male births [81]. The worldwide prevalence of hemophilia is about 400,000, and globally only 25% of affected patients receive adequate treatment [82]. The disease is clinically manifested by repeated bleeding due to deficiency of clotting factor VIII (FVIII) (hemophilia A) or factor IX (FIX) (hemophilia B). About 80-85% of all cases are hemophilia A. About 50% of all hemophilia patients have severe hemophilia defined by lower than 1 or 2% FVIII levels and recurrent bleeds, mostly in joints, either spontaneously or after minor injuries. Over time, repeated joint bleeding leads to arthropathy and disability [83]. The main treatment in many developed countries is prophylactic FVIII replacement for bleeding prevention; however, the approaches for bleeding prophylaxis vary across countries. High dose primary prophylaxis is FVIII infusions given on alternate days. It is considered to be one of the best methods to prevent bleeding and improve patient outcomes, and is implemented in many Scandinavian and Northern European countries [84, 85]. In many other countries, patients are treated with an on-demand regimen, that is, FVIII infusions only when bleeding occurs.

Hemophilia is a very expensive disease to treat. The high treatment costs are explained by the high cost of FVIII which accounts for 90% of total treatment costs in hemophilia [86, 87]. In a study in France, the mean annual per patient treatment costs with on-demand therapy was estimated to be USD $73,029 [88].

Until now, there has been only one completed randomized clinical trial (RCT) in hemophilia; it compared alternate day prophylaxis (AP) with on-demand (enhanced episodic) treatment (OD) [89]. This multi-centre study was conducted in the U.S. among 65 boys with severe hemophilia. The study found that by the age of six, the OD group had a relative risk of magnetic resonance imaging (MRI) detected joint damage of 6.1 (95% Confidence Interval from 1.5 to 24.4)
compared to the AP group. The trial estimated that the average annual cost of treatment with alternate day prophylaxis for a 50kg person would reach USD $300,000.

Concurrent with the U.S. study, a single-arm, multi-centre trial was launched in Canada where patients, who met the same eligibility criteria as in the U.S. trial, were treated with tailored prophylaxis (TP). All patients in this cohort (n=25) started with once weekly prophylaxis and escalated to twice weekly or alternate day prophylaxis if the bleeding pattern met specified escalation criteria [90]. At 5 years of follow-up 40% of subjects were on once-weekly prophylaxis, 32% on twice weekly prophylaxis, and 28% on alternate-day prophylaxis. Joint examination scores at the end of 5 years, obtained from radiographs and physical examination, were close to normal [90].

In this report we use VOI methods to determine whether or not current evidence is sufficient to adopt one treatment modality over the others. If not, VOI methods were used to determine the optimal sample size for additional RCTs. All three paired comparisons were considered. Comparison 1 is AP vs TP; Comparison 2, TP vs OD; and, Comparison 3, AP vs OD. The current information on costs and effectiveness of the treatment modalities were taken from the U.S. [89] and the Canadian trials [90].

5.2 Materials and methods

5.2.1 VOI methodology

VOI methodology considers current information and information from future trials, and is supported by Bayesian statistical decision theory. Incremental net benefit \((b)\) of treatment \(A\) in comparison to \(B\) is the net benefit of \(A\) minus the net benefit of \(B\). It can defined as \(b = \Delta E \lambda - \Delta C\), where \(\lambda\) is the willingness-to-pay or threshold value for an additional unit of effectiveness, \(\Delta E\) is the difference \((A - B)\) in effectiveness, and \(\Delta C\) is the difference \((A - B)\) between the treatment costs. The incremental net benefit based on current information is assumed to follow a normal distribution with mean \(b_0\) and variance \(v_0\). To reduce uncertainty, a new trial can be conducted which provides updated values for the mean and variance \((b_1,v_1)\), and post-trial per-patient expected opportunity loss \((EOL_1)\). The expected value of sample information (EVSI) from a proposed new trial is the reduction of the per-patient expected opportunity loss it provides \((EOL_0-EOL_1)\), multiplied by the number of patients that can benefit from the decision.
based on the trial results. The expected total cost (ETC) is calculated considering the fixed and per-patient variable financial costs of the trial, as well as the opportunity costs for patients who receive the less favorable treatment (based on current knowledge) because of the trial. The ENG is calculated as the difference between EVSI and ETC. The optimal sample size for the new trial is the sample size that maximizes the ENG. If the maximum ENG is negative, i.e. EVSI is less than ETC for all sample sizes, the existing information is sufficient for decision making, selecting the treatment with the largest net benefit, and the optimal sample size is zero. However, if the maximum ENG is positive, doing a new trial is the optimal decision. More detailed description of methods and formulas applied in this study are described in articles by Willan and Pinto[78], Eckermann and Willan [75-77, 79], Willan and Eckermann[79], and Willan[80].

The hypothetical scenario assumes that a new, multi-centre RCT is planned in the U.S. and Canada to enroll patients with the same eligibility criteria as the previous trials, namely children with severe hemophilia, less than 30 months old, with no joint damage and no clotting factor inhibitors [89]. The hypothetical sample size per arm \( n \) was varied from 0 to 100 for each scenario. Considering the number of annual live male births in Canada [91] and the U.S. [92] in 2006, the annual incidence of severe hemophilia \( k \) was assumed to be 240 in the countries combined. The time horizon \( h \), which is the time horizon for the decision of which potential treatment to adopt before a new set of treatments might be considered, was set to 20 years for the base-case, and to 15 and 25 years in the sensitivity analyses. The patient enrollment fraction \( f \) or proportion of incident cases willing to enroll into the study was set to 20% in the base-case, and to 15% and 25% for sensitivity analyses. It was assumed that the decision would affect only new incident cases. The number of eligible patients who would potentially benefit from the decision based on the trial results \( N \) is given by \( (h-(t_a + \tau))k \) where \( t_a = (2n)/(fk) \) is the duration of trial accrual and \( \tau \) is the duration of follow-up (i.e. from the end of patient accrual to when the data is available to update the evidence). Following the example of the U.S trial, the proposed study would have five years of follow-up with an additional one year for data analysis and dissemination of the results (i.e. \( \tau = 6 \)). The fixed cost \( (C_f) \) to set up the trial was estimated to be $1,000,000 in the base-case, and $500,000 and $2,000,000 in the sensitivity analyses. All costs estimates were in US dollars. Per-patient variable cost \( (C_v) \) was estimated to be $5,000 for the total trial duration in the base case and
$2,500 and $10,000 respectively with low and high fixed cost scenarios in the sensitivity analyses. The number of patients receiving the treatment with the smallest net benefit ($\bar{n}$) while the trial is underway was given by $k(t_e + \tau) - n$.

The mean incremental net benefit ($b_0$) based on current knowledge was calculated using the cost and effectiveness estimates from all three potential treatments as determined by the U.S. trial and the Canadian tailored prophylaxis cohort studies [89, 90]. Effectiveness was measured as the proportion of children without any joint damage at the age of six estimated by MRI findings. MRI joint damage was considered positive if the image demonstrated a subchondral cyst, surface erosion, or joint-space narrowing [89, 93]. For the cost estimates, which are based on FVIII use only, we used the average FVIII units infused during the total trial duration accounting in this way for about 90% of total treatment costs [87]. The cost of FVIII was estimated as $1 per unit.

### 5.2.2 Prophylaxis (AP) versus Episodic Treatment (OD) in Hemophilia

The U.S. trial reported that by the age of six 93% (25/27) of patients in the AP arm and 55% (16/29) in the OD arm did not have MRI detected joint damage [89]. Using our estimate of FVIII cost of $1 per unit, the average per patient treatment cost was $352,973 ± 150,454 for patients in the AP arm (n = 32) and $113,237 ± 65,494 for patients in the OD arm (n=33).

### 5.2.3 Canadian Hemophilia Primary Tailored Prophylaxis (TP) Cohort

At the time of the MRI joint assessment, children in the Canadian cohort were older (average 8.7 ± 1.5 years, range 6-12 years) than children in the U.S. trial (6 years old), and 12 out of 24 subjects on the trial had MRI detected joint damage. To estimate the MRI score at the age of six, we examined clinical variables that were able to predict MRI joint damage at the actual time of MRI assessment. The following variables were available at the time of MRI and at the age of six: number of total lifetime joint hemorrhages (average 5.92 ± 4.87 and 12.88 ± 14.33 at the age of six and at the time of MRI), number of hemorrhages per year (average 0.99 ± 0.812 and 1.48 ± 1.52), target joint defined as at least three consecutive bleeds into the same joint within six months (48% and 60%), and joint physiotherapy score (modification of the World Federation of Hemophilia physical examination score [94], average 3.12 ± 2.15 and 4.35 ± 3.54). The largest correlations were observed between the MRI scores and the number of target
joints (point-biserial correlation coefficient = 0.38) and the MRI scores and the number of hemorrhages per year (Pearson correlation coefficient = 0.30). Using the MRI joint score linear prediction rule at the time of MRI assessment (‘predicted MRI score’ = 4.681 + 1.926*’target joint at the time of MRI’ + 0.218*number of hemorrhages per year at the time of MRI), the predicted number of children with MRI detected joint damage at the age of six would have been 5/24 corresponding to 79% with no damage.

The average per patient treatment cost with tailored prophylaxis was $275,303 ± 103,936. In order to account for potential weight differences between the Canadian and the U.S. patients, we estimated the average weight of the subjects in alternate day prophylaxis and in tailored prophylaxis by dividing the reported total FVIII use of each cohort (11,289,372 for AP and 6,882,579 for TP) by their FVIII use per average kilogram (sum of FVIII use for bleeding events – average annual bleeds*average study duration*n*FVIII use for bleedings – and FVIII use as assigned per protocol). We found that, on average, children in the Canadian cohort had a 1.17 times higher weight than children in the U.S. study. This number was used to standardize the average per patient cost of the tailored prophylaxis as $235,302 ± 88,834.

5.2.4 Other parameters for calculations

Table 1 presents for all three comparisons the estimated differences in mean effectiveness (ΔE) and in mean costs (ΔC), variances of differences (V(ΔE), V(ΔC)) and the covariance (C(ΔE, ΔC)), and the sum across treatment modality of the between-patient variance of net benefit (σ2). The differences taken for the three comparisons are: Comparison 1, AP – TP; Comparison 2, TP – OD; and, Comparison 3, AP – OD. For study purposes, the threshold value (λ) to prevent one child from developing joint damage was varied from 0 to $1,200,000. The correlation (ρ) between cost and joint damage was evaluated in the Canadian cohort and was assumed to be the same in the U.S. trial. The point-biserial correlation coefficient was -0.04 and was used to calculate the covariance (C(ΔE, ΔC)). Base-case analyses were performed ignoring discounting for future costs and benefits, and 5% and 3% discounting rates were investigated in the sensitivity analyses. Prior mean incremental net benefit (INB) was calculated by
\[(\hat{\lambda}_e \lambda - \hat{\lambda}_c),\] where \(\lambda\) was the threshold value for avoiding a damaged joint. Variance of the prior INB \(v_0\) was calculated as \((\lambda^2 V(\hat{\lambda}_e) + V(\hat{\lambda}_c) - 2\lambda \hat{C}(\hat{\lambda}_e, \hat{\lambda}_c))\).

### 5.2.5 Assumptions

It was assumed that for any particular comparison if a trial was not done, patients would receive the treatment modality with the highest net benefit, and if a trial was done, that those patients not in the trial would receive the treatment modality with the lowest net benefit. It was also assumed that no new treatment was developed and implemented during the time horizon, and that no planned interim analysis or early stopping rule would be applied. For simplicity, we ignored the risk of FVIII inhibitor development and its effect on ETC. (There were only two cases of FVIII inhibitor development in the U.S. RCT and only one case of a transient inhibitor development in the Canadian cohort [89, 90].) The disutility of having more frequent injections or infection complications in high dose or tailored prophylaxis compared with on-demand prophylaxis was also ignored.

### 5.3 Results

We estimated the mean net benefit of all treatment modalities for threshold values from $0 to $1,200,000. Figure 1 shows that for threshold values lower than $400,000 OD has a higher net benefit than either AP or TP. The difference between the treatment modalities is less pronounced for threshold values between $400,000 and $800,000. For threshold values above $800,000 AP and TP have a higher net benefit than OD.

The relationship between the threshold value and the optimal decision was evaluated by calculating the optimal sample size for different threshold values. The optimal sample size, as a function of the threshold value, is plotted in Figure 2 for all comparisons. In Comparison 1, for threshold values between 0 and $480,000 (area A) there is sufficient evidence to adopt TP, and the optimal sample size is 0. For threshold values greater than $1,350,000 (area D) there is sufficient evidence to adopt AP. Between the values of $480,000 and $1,350,000, the EVSI exceeds ETC and the optimal sample size is positive. In Comparison 2, for threshold values between 0 and $363,000 (Area A) there is sufficient evidence to adopt OD, and the optimal sample size is zero. For threshold values over $680,000 (area D) the evidence supports TP (optimal ample size is zero). For values between $363,000 and $680,000 the optimal sample
size is positive and new trials will have positive ENG. In Comparison 3, for threshold values between 0 and $510,000 (area A) there is enough evidence to adopt OD without a trial. For threshold values over $790,000 (area D) the evidence is enough to adopt AP. For the values in between, the sample size is positive and a new trial is warranted.

We conducted two base-case analyses, one using a threshold value of $400,000 and one using a threshold value of $1,000,000. For a threshold value of $400,000, there was enough evidence to choose TP over AP (Comparison 1) and OD over AP (Comparison 3). However, there was not sufficient evidence to choose OD over TP (Comparison 2) even though it had a larger net benefit, and a new trial to reduce the uncertainty was worthwhile. The optimal sample size for the new trial, using base-case parameters, was 42 patients per arm with the ETC of $48,847,548 (financial cost of $1,420,000 and opportunity cost of $47,427,548) and an ENG of $28,070,005. The return on investment, calculated by dividing ENG by ETC, was 57%. The EVSI and ETC, as functions of sample size for this comparison are presented graphically in Figure 3. As demonstrated in Figure 3, there is a positive ENG starting from a sample size of 9 patients per arm which is maximized at 42 patients per arm.

For a threshold value of $1,000,000, there was sufficient evidence to choose either prophylaxis method over OD (Comparisons 2 and 3). However, a new trial comparing the two prophylaxis methods was the optimal decision (Comparison 1). Using the base-case parameters for Comparison 1, the optimal sample size was 48 patients per arm with the ETC of $32,869,806 (financial cost of $1,480,000 and opportunity cost of $31,389,806) and an ENG of $125,542,013. The return on investment was 380% corresponding to 19% annualized rate of return for the 20 years time horizon.

Table 2 and Figure 4 present the results of sensitivity analyses for Comparison 1 with a threshold value of $1,000,000. In the sensitivity analyses of Comparison 1, the change from base-case parameters to both high and low trial costs offered little difference in the ENG and the optimal sample size. Applying a discounting rate of 3% and 5% decreased the EVSI, ETC, and ENG. Overall, the largest variation of the ENG was observed when varying the time horizon \((h)\) (Figure 4). ENG was positive in all sensitivity analyses of Comparison 1 for all sample sizes indicating the need to conduct a new trial before adopting AP.
5.4 Discussion

In decision analysis, the recommendation for the adoption of new health interventions is made on the basis of a positive expected incremental net benefit. However, in rare diseases often the evidence is scarce and insufficient for decision making. In rare diseases, although the number of people affected by a decision can be small, the overall opportunity loss can be significant. In considering the funding of new research and patient reimbursement in rare diseases, VOI methodology provides more relevant determinations of the value and costs of additional research.

Using VOI methodology and the cost and effectiveness estimates of hemophilia treatment from the U.S. and Canadian trials, we were able to compare current treatment modalities of hemophilia. We found that the sign of the incremental net benefit for the paired comparison depended on the threshold value. Low threshold values favored on-demand therapy, while higher threshold values favored prophylactic treatment modalities.

Using a specific threshold level and applying VOI methodology, a decision maker can identify treatments for which there is enough evidence and treatment for which there are not enough evidence and new trials are warranted. Moreover, VOI methodology can identify the optimal sample size for a new trial that maximizes the expected net gain. For example, in our study, for a threshold value of $400,000 there is enough evidence to choose either tailored prophylaxis or on demand therapy over alternate day prophylaxis. However, a new trial with a sample size of 42 patients per arm would be needed to make an optimal choice between on-demand therapy and tailored prophylaxis. With a threshold level of $1,000,000 to avoid joint damage, there is sufficient evidence to choose either prophylactic modality over on-demand therapy. However, a new trial would be needed to make an optimal choice between the prophylaxis modalities.

The robustness of findings can be evaluated by estimating the impact of various assumptions and parameters using sensitivity analyses. We illustrate this using the example of Comparison 1 with a threshold value of $1,000,000; the results were robust for parameters of discount rate, trial fixed costs and variable costs, enrollment fraction, and the time horizon. The small impact of changes in the trial costs on the ENG was explained by the fact that the EVSI was several
times larger than the ETC. A lower enrollment fraction decreased the optimal sample size and
the ENG because the increased trial duration decreased both the time in which an optimal
decision might be applied and reduced the number of patients who would benefit from the trial.
The time horizon had the largest impact on the optimal sample size because it directly affected
the number of people who would benefit from the treatment after an optimal decision would be
made.

To our knowledge, no study has attempted to estimate the threshold value for avoiding joint
damage in hemophilia. A study in Sweden used a contingent valuation method to assess the
willingness-to-pay (WTP) for on-demand therapy and once to three-time weekly prophylaxis in
a representative sample of the Swedish population [95]. The study found that the mean WTP
from an individual’s annual income was EUR 39 (USD$57 using 2008 yearly average
exchange rate) for on-demand therapy and EUR 65 (USD$95) for prophylaxis. Using these
estimates, it can be inferred that if one million taxpayers agree to pay $95 from their annual
income for hemophilia prophylaxis, treatment for at least 100 new incident cases using tailored
or alternate day prophylaxis could be provided by a publicly-funded health care system.

Our study results must be interpreted in the light of potential limitations. We made assumptions
about several parameters including patient accrual fraction, time horizon of the decision, and
trial costs that may have been incorrect. However, these assumptions were explored in
sensitivity analyses, and our results were found to be robust. There are recent articles about EVI
methodology that have tested the impact of other parameters and assumptions not considered in
this study [72, 75, 76, 78].

The result of our study can help to plan RCTs in rare diseases with high treatment costs. There
are other rare diseases that require expensive, continuous enzyme replacement, for example,
and may need more trials to establish the optimal schemes for disease management. For
example, the per patient/per year cost of enzyme replacement in Gaucher’s disease by
Imiglucerase is £70,100, by Agalsidase in Fabry’s disease £109,600, and by Laronidase in
Mucopolysaccharidosis 1 approximately £311,000 [4]. A careful analysis of future expected
costs and benefits in planning RCTs in these diseases using VOI methodology can increase the
informativeness of future trials.
In conclusion, using the VOI methodology we estimated the expected value of sample information, the expected trial costs and the expected net gain for therapeutic trials of hemophilia – a rare disease with high treatment costs. We hope that the results of this study will help in decision making for treatment strategies in hemophilia and in planning of future studies in this condition and in other rare diseases.
Table 1. Parameter estimates for the trials

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Description</th>
<th>Comparison 1(^A)</th>
<th>Comparison 2(^B)</th>
<th>Comparison 3(^C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{\Delta}_E)</td>
<td>Difference in effectiveness: (\hat{\Delta}_E = E_T - E_S)</td>
<td>0.134</td>
<td>0.240</td>
<td>0.374</td>
</tr>
<tr>
<td>(\hat{\Delta}_C)</td>
<td>Difference in costs: (\hat{\Delta}_C = C_T - C_S)</td>
<td>$117,491</td>
<td>$122,065</td>
<td>$239,556</td>
</tr>
<tr>
<td>(\hat{V}(\hat{\Delta}_E))</td>
<td>Variance of difference in effectiveness: (\hat{V}(\hat{\Delta}_E) = V(E_T) + V(E_S))</td>
<td>0.00868</td>
<td>0.0143</td>
<td>0.00981</td>
</tr>
<tr>
<td>(\hat{V}(\hat{\Delta}_C))</td>
<td>Variance of difference in costs: (\hat{V}(\hat{\Delta}_C) = V(C_T) + V(C_S))</td>
<td>30,527,881,724</td>
<td>12,180,939,644</td>
<td>26,925,870,152</td>
</tr>
<tr>
<td>(\hat{C}(\hat{\Delta}_E, \hat{\Delta}_C))</td>
<td>Covariance of differences in effectiveness and costs: (\hat{C}(\hat{\Delta}_E, \hat{\Delta}_C) = \rho \sqrt{\hat{V}(\hat{\Delta}_E)\hat{V}(\hat{\Delta}_C)})</td>
<td>-651</td>
<td>-528</td>
<td>-651</td>
</tr>
<tr>
<td>(\sigma_+^2)</td>
<td>Between-patient standard deviation: (\sigma_+^2 = \frac{1}{2} (\sigma_T^2 + \sigma_S^2))</td>
<td>((777,928)^2)</td>
<td>((478,887)^2)</td>
<td>((685,058)^2)</td>
</tr>
</tbody>
</table>

\[\begin{align*}
\sigma_T^2 &= \lambda \sigma_{E_T}^2 + \sigma_{C_T}^2 - 2 \lambda \hat{C}(\hat{\Delta}_E, \hat{\Delta}_C) \\
\sigma_S^2 &= \lambda \sigma_{E_S}^2 + \sigma_{C_S}^2 - 2 \lambda \hat{C}(\hat{\Delta}_E, \hat{\Delta}_C)
\end{align*}\]

\(A: \text{Compares alternate day prophylaxis with tailored prophylaxis.}\)

\(B: \text{Compares tailored prophylaxis with on-demand therapy.}\)

\(C: \text{Compares alternate day prophylaxis with on-demand therapy.}\)
Table 2. Sensitivity analyses for Comparison 1 (threshold value = $1,000,000)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Optimal n</th>
<th>EVSI</th>
<th>ETC</th>
<th>ENG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base-case</td>
<td>48</td>
<td>158,411,861</td>
<td>32,869,745</td>
<td>125,542,116</td>
</tr>
<tr>
<td>High trial costs*</td>
<td>48</td>
<td>158,411,819</td>
<td>34,349,806</td>
<td>124,062,013</td>
</tr>
<tr>
<td>Low trial costs*</td>
<td>49</td>
<td>158,570,237</td>
<td>32,285,719</td>
<td>126,284,518</td>
</tr>
<tr>
<td>Enrolment fraction 15%</td>
<td>40</td>
<td>149,049,666</td>
<td>33,818,247</td>
<td>115,231,418</td>
</tr>
<tr>
<td>Enrolment fraction 25%</td>
<td>56</td>
<td>165,390,369</td>
<td>32,279,084</td>
<td>133,111,285</td>
</tr>
<tr>
<td>Time horizon 15yrs</td>
<td>36</td>
<td>92,420,772</td>
<td>30,938,856</td>
<td>61,481,916</td>
</tr>
<tr>
<td>Time horizon 25yrs</td>
<td>59</td>
<td>227,776,092</td>
<td>34,639,844</td>
<td>193,136,247</td>
</tr>
<tr>
<td>Discount rate 3%</td>
<td>48</td>
<td>127,997,006</td>
<td>29,318,064</td>
<td>98,678,942</td>
</tr>
<tr>
<td>Discount rate 5%</td>
<td>48</td>
<td>101,468,977</td>
<td>27,548,307</td>
<td>73,920,670</td>
</tr>
</tbody>
</table>

*High-cost scenario assumes a fixed cost of $2,000,000 and a variable cost of $10,000.

**Low-cost scenario assumes a fixed cost of $500,000 and a variable cost of $2,500.
Figure 1. Net benefit of all treatment methods calculated as $\lambda * E - C$. 

<table>
<thead>
<tr>
<th>Net Benefit ($$, mln)</th>
<th>Threshold value, mln</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>1.2</td>
</tr>
</tbody>
</table>
Comparison 1: Alternate day vs Tailored prophylaxis

Comparison 2: Tailored prophylaxis vs On-demand Therapy
Figure 2. Optimal sample size as a function of threshold value ($\lambda$). For all sample size and threshold value combinations, in area A the optimal decision is to retain the standard treatment; in area B the current evidence supports the standard (lower cost, lower effectiveness) treatment but a new trial offers a reduction of uncertainty and an expected net gain if a trial is done; in area C the current evidence favors the experimental treatment (higher cost, higher effectiveness) but a new trial is warranted to reduce the uncertainty; and in area D there is enough evidence to adopt the superior experimental treatment.
Figure 3. Expected value of sample information and expected total cost as a function of sample size for Comparison 2 (TP vs. OD therapy) and threshold value of $400,000. EVSI is the expected value of sample information, ETC is the expected total cost, $n^*$ is the optimal sample size, and ENG is the expected net gain.
Figure 4. Change of the expected net gain with changing of the time horizon \((h)\) and enrollment fraction \((f)\) in Comparison 1 (threshold value = $1,000,000).
Chapter 6. Concluding Remarks

Conducting research studies for rare diseases can be an arduous process. Over the years, investigators and patients have created international networks that support research in rare diseases. For example, the Rare Diseases Clinical Research Network (http://rarediseasesnetwork.epi.usf.edu/), National Organization of Rare Disorders (http://www.rarediseases.org/), and Orphanet (http://www.orpha.net/) provide detailed information about rare diseases as well as about ongoing studies (patient population, eligibility criteria, intervention, and stage of the study), newly available treatment methods, and existing patient registries. Sharing available information and encouraging the development of new research methods are key factors towards the improvement of management and treatment of rare diseases. This thesis project demonstrates that there are methods that can be applied individually or in combination to improve the planning and execution of RCTs for rare diseases.

6.1 Synthesis

Quality of RCTs in Rare Diseases – the case of JIA

It is well established that the quality of evidence obtained from a study depends highly on its methodological and reporting quality. The quality of a study is one of the main criteria used in systematic reviews or meta-analyses when the decision is made to include or exclude a study in the data synthesis. The objective of the first project of this dissertation was to evaluate the quality of RCTs in rare diseases using Juvenile Idiopathic Arthritis (JIA) as an example. A systematic review identified all RCTs that involved exclusively JIA patients. Six main individual quality indicators were used to assess the quality of each RCT. The study concluded that the quality of RCTs in JIA based on the selected indicators was poor. There were some positive changes over time but it was recommended that more efforts are needed to ensure adequate quality.

The study indicated that the introduction of the CONSORT statement resulted in the improvement of reporting of RCTs – a finding confirmed in other similar studies. For rare diseases, where the evidence is often limited, good methodological and reporting quality is an
additional effort to increase the informativeness of a study. In 2007 a similar guideline, the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE), was developed for reporting observational studies [96]. This guideline should be used for reporting of observational research that involves a rare disease. Recently, a study evaluating the quality of reporting in sports injury abstracts using 17 CONSORT and 22 STROBE criteria found significant improvement of several indicators over the last three years both in RCTs and in observational studies [97].

Our study of the quality of RCTs in JIA indicated that the use of individual component approach is a valid and feasible method for evaluating study quality. This approach avoids any potential bias associated with the calculation of total/summary quality scores. It allows addressing specific indicators as well as other important characteristics. For example, we were able to highlight the difficulties in conducting RCTs in rare diseases by observing small sample sizes, and high proportion (50%) of studies that failed to reject the null hypothesis. These findings indicate that new, more powerful and more acceptable designs are needed for rare diseases.

**Response time distributions**

This project demonstrated that information from past trials and patient registries can be used to estimate the exact treatment response time distributions for a given patient population and intervention. It is generally taught that the exponential distribution can provide a reasonable estimation of response times over a short period of time. However, with highly effective treatments this assumption may not always be true, and other distributions need to be assessed (e.g., lognormal, log-logistic, generalized gamma, Weibull or piecewise distribution that are more flexible to model a changing hazard function). Similarly, the exponential distribution may not be the best fit for time-to-event outcomes if one expects a change of patient condition over a short period of time or when aging over a long follow-up period is expected.

In this study we used data from an RCT (TEMPO trial) and a patient registry (German Etanercept Registry) to evaluate the response time distributions for highly effective treatments such as Etanercept and Methotrexate in patients with rheumatic arthritis. Our study found that in the majority of evaluated outcomes (ACR20, ACR50, and ACR70 in patients with rheumatic
arthritis and ACR30, ACR50, and ACR70 in patients with juvenile idiopathic arthritis) the response times followed mostly lognormal or generalized gamma distribution.

The identification and description of the appropriate parametric model is important for the understanding of disease-drug interactions, patient management, study planning, and accurate data analysis. Maximizing the use of available information from previous studies can be a cost-effective way to increase the efficiency of future trials.

**Power of RCTs under different response time distributions**

We used two RCT designs - the Randomized Placebo-Phase Design (RPPD) and the traditional parallel groups RCT design to evaluate the effect of the underlying hazard distribution on the power of a clinical trial. For study purposes a discrete time, object-oriented simulation program was developed. The results of simulations showed that the power of RCTs, both for the RPPD design and the parallel groups RCT, is dependent on the underlying hazard distribution. The smallest sample size requirements were observed for a Weibull distribution and the largest for an exponential distribution, supporting findings from previous studies. For example, a study found that when the response times follow Weibull distribution, the sample sizes heavily depend on the shape parameter, and the difference between the power when assuming an exponential or Weibull distribution can be very large [98].

Our study not only compared the differences in power but also developed a sample size calculator for various response time distributions that can be used by the wider scientific community. The program developed for this project allows investigators to do sample size calculations for time-to-event outcomes when the response times follow exponential, Weibull or lognormal distribution, taking into account other important trial design parameters (e.g., accrual and follow-up duration, patient visit times, distribution parameters, accrual and drop-out probability). Consideration of the exact distribution can improve the trial efficiency and decrease societal opportunity costs by decreasing the number of patients enrolled in a study, the total trial duration, and trial costs.
**VOI in rare diseases**

Very often there is a wide uncertainty regarding the evidence for a new health care intervention in rare diseases, and it is not easy to make a decision whether the new treatment method should be adopted without further study. This dilemma can be addressed by taking into consideration the opportunity loss associated with any decision. Value of information (VOI) methodology compares the expected value of information from an additional trial against the expected cost of the trial, and identifies the decision that maximizes the expected net gain. In this study, VOI methodology was applied to a rare disease with high treatment costs, using hemophilia as an example.

To apply VOI methodology, an investigator must identify several parameters (more than needed for traditional sample size calculations) and make assumptions. Parameters for calculations can be obtained from various sources including published studies and patient registries, and assumptions can be checked in sensitivity analyses. With VOI methodology, the study sample size calculation is not restricted to traditional, pre-defined Type I and Type II error levels; the main purpose, rather, is to identify the opportunity loss and maximize the expected net gain of the obtained information. Moreover, this methodology explicitly considers both the prevalence of the disease (the rarity in this case) and the feasibility of enrollment in calculations while this is not the case with regular sample size calculators. We hope that after publication of our results more investigators will become aware of the method and more decision makers will apply it in the process of making reimbursement decisions.

**6.2 Implications and policy recommendations**

This thesis project demonstrates that the improvement of existing methodological approaches and development of new approaches can increase the efficiency of future trials for rare diseases. First, the deficiencies in reporting and methodological quality of RCTs in rare diseases using RCTs in JIA as an example were highlighted. To minimize bias and improve the quality of evidence, investigators should pay more attention to the conduct and reporting of random number generation and allocation concealment, make studies double masked whenever possible, conduct ITT analysis and ensure complete follow-up as far as possible.
The second project showed that the response times for time-to-event outcomes may follow different distributions, and the exponential distribution may not be accurate to represent the response hazard in all conditions. In order to do more accurate sample size calculations and statistical analysis, investigators should evaluate the response time distributions using information from available sources such as previous RCTs, observational studies, and patient registries. Patient registries are widely used in rare diseases for the assessment of the natural progress of the disease, the burden and cost of the illness, drug effectiveness, practice patterns in various locations, and long-term outcomes. This project showed that patient registries can also be used for the assessment of the treatment response time distributions for specific drugs and patient populations. This information can be used for planning of future studies.

Further, we found that the sample size requirement can be different for various combinations of treatment response time distributions. By considering the exact survival distribution, investigators may be able to decrease the required sample size or make sure the study has enough power to detect the proposed difference in the effectiveness (i.e. maximize the efficiency of proposed trials). We hope that after the publication of the code the calculation of the sample size for time-to-event outcomes that follow an exponential, Weibull, or lognormal distribution will be more feasible for other investigators.

Our computer simulations showed that for some scenarios the power of the RPPD and the traditional parallel groups RCT were not very different. Assuming that the RPPD will have a higher acceptability and higher recruitment rates, clinical investigators may consider this design for conditions when patients have a preference for the experimental medication (e.g. rare diseases with no established standard treatment, cancer trials, AIDS trials). This design can be used for both Phase II and Phase III RCTs.

Computer simulations have been used in medical research for more than twenty years. Simulation models can effectively combine information from various sources (e.g., previous trials, patient registries, literature search) and various methods of statistical computing to inform investigators about the potential limitations/outcomes of future trials. There should be a wider application of modeling approaches in rare diseases where the evidence is scarce, the resources are limited, and the planned studies must therefore be highly efficient.
We also believe that the VOI methodology is a valuable tool for planning studies for rare diseases where the uncertainty around the outcomes can be quite large. This methodology can indicate whether there is sufficient evidence to adopt the proposed new treatment or whether a new study is needed to decrease the uncertainty around the findings. Moreover, VOI methodology will indicate the optimal sample size that will maximize the difference between the cost of the trial and the value of additional information.

All these methodological changes can be applied to future RCTs of rare diseases – either separately or in combination. For example, an investigator who plans an RCT for a rare disease with time-to-event outcomes should use VOI methods to determine if a new study is worthwhile, or if the evidence supports the benefit of known treatments. If a new study is called for, the sample size that maximizes the expected net gain received (using VOI methodology) should be determined. The investigator should use previous studies to estimate the response time distributions for the particular population and treatment method. Then, the investigator should select a study design that is more feasible and more acceptable for patients and physicians. The sample size of the study can be further refined using our “Simulation-based sample size calculator for the RPPD” program after defining trial parameters. As is the case for all clinical trials, at the planning stage the investigator should make sure that the randomization, the treatment allocation, and masking will be adequate. The final analysis should be planned following the ITT principle with a priori defined approaches for missing data. The trial details should be finalized after considering the trade-offs between patient preferences for treatment allocation, design acceptability, and study feasibility.

6.3 Future research directions
Recently, there have been several articles debating whether rare diseases deserve a special status in health technology assessment (HTA) and, subsequently, in patient reimbursement decisions [3, 13, 63, 99, 100]. Many rare diseases are expensive to treat; the resources used to treat one patient with a rare disease can save the lives of many others with more common and less expensive diseases. However, it is unclear whether society values rare diseases differently considering the fact that many of these diseases have a genetic origin (i.e. not the “fault” of the sufferer as opposed to so-called lifestyle diseases) and affect children (who appear to be more
highly valued in many societies). Research studies are needed to obtain comparative values or weights for the societal threshold / willingness-to-pay (WTP) for disease characteristics such as rarity, severity, and societal opportunity costs. Surveys evaluating WTP should be conducted among general public, patients, and decision makers. The results may vary between countries but they may improve HTA and patient care for rare diseases in general.

Our results suggest a future program of research. The randomized clinical trial is an established method for obtaining high quality evidence about treatment methods. There are several designs of RCTs; however, patients’ and physicians’ preferences for these designs are not well established. A systematic review should be conducted to define all RCT designs that have been reported. A subsequent study can use a preference elicitation method to evaluate the acceptability of different designs or design components among patients and physicians. Patient and disease characteristics that can potentially affect the preferences should also be considered (for example, patient gender, age, disease duration, severity, cost of treatment, previous treatment failures, preference to placebo/experimental treatment, chronic versus acute condition, etc). This type of evaluation will facilitate the development of new, more acceptable designs of RCTs. The new designs can then be evaluated with computer simulations before applying to real studies.

In conclusion, the efficiency of future RCTs for rare diseases can be improved by refining the existing methodologies, by maximizing the use of available information, and by encouraging the development of new approaches. The present work evaluated methods that can be used to increase the efficiency of studies in rare disease – by using data available from previous RCTs and from patient registries, by computer simulations, and by evaluating the costs and benefits of future research studies. The study findings will be presented in scientific conferences and will be submitted for publication in peer-reviewed journals.
Bibliography


75. Eckermann S, Willan AR. Expected value of information and decision making in HTA. Health Econ 2007;16(2):195-209.


Appendices

Appendix 1. Articles included in the final review


Appendix 2. Statistical characteristics of survival distributions

<table>
<thead>
<tr>
<th>Statistical Distributions*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameters</strong></td>
</tr>
<tr>
<td>Parameters</td>
</tr>
<tr>
<td><strong>Probability density function</strong></td>
</tr>
<tr>
<td><strong>Surviviorship function</strong></td>
</tr>
<tr>
<td>$\int_{-\infty}^{\gamma} e^{-\frac{1}{2\sigma^2} \left(\log x - \mu\right)^2} dx$</td>
</tr>
<tr>
<td><strong>Hazard function</strong></td>
</tr>
<tr>
<td><strong>Median</strong></td>
</tr>
</tbody>
</table>

The exponential, Weibull, lognormal, and gamma distributions are special cases of a **generalized gamma (GG)** distribution with three parameters, $\lambda$ (scale), $\gamma$ (shape), and $\alpha$ (locator parameter). The density function of the GG is defined as:

$$f(t) = \frac{\alpha \gamma t^{\gamma-1}}{\Gamma(\gamma)} \exp\left[-(\lambda t)\gamma\right]$$

where $t$, $\lambda$, $\alpha > 0$ and $\Gamma(\gamma)$ is the gamma function defined as $\Gamma(\gamma) = \int_0^\infty x^{\gamma-1} e^{-x} dx$. The GG equals to exponential distribution when $\gamma = \alpha = 1$, the Weibull when $\alpha > 0$ and $\gamma = 1$, the lognormal when $\gamma \to \infty$, and gamma when $\alpha = 1$.

Appendix 3. The R code for computer simulations

SIMULATION-BASED SAMPLE SIZE CALCULATOR FOR THE RPPD

This discrete time, object-oriented simulation program is developed for the investigation of the statistical efficiency (power) of the Randomized Placebo-Phase Design (RPPD) and for parallel group RCTs with time-to-event outcomes, using R statistical software (Version 2.6.1, The R Foundation for Statistical Computing). The program calculates the power of trials for different sample sizes when the response times to experimental or standard treatment follow exponential, Weibull or lognormal distributions. Appendix A presents the program flow chart. Appendix B is the program code with clinical trial design input parameters that can be modified by the user. Appendix C is the source code (‘RPPD.R’) for the simulation function that should be saved in the same directory as the program code (Appendix B).

Directions: In order to run the simulation and calculate study power, input the corresponding parameters included in the “CLINICAL TRIAL DESIGN MODIFIABLE PARAMETERS” section of Appendix B and submit the code. For large sample sizes (total n ≥ 50) and for a large number of simulations (≥ 1000 iterations) you may consider using a high performance computer.

Notes:
More information about ‘R.oo’ package for object-oriented simulations is available at:
http://www1.maths.lth.se/help/R/R.oo/

Appendix 3A. Simulation Program Flow Chart

Start

Initialize patients

Did patient drop-out during current visit?

Yes

Assign current time and censor

No

Calculate conditional probability of survival

Did patient respond during current visit?

Yes

Assign current time and response status

No

Is this the last visit?

Yes

Assign current time and censor

No

Go to next cycle/visit ➔ Go to 1

Drop current trial if all patients are in one group or all have the same survival time

Calculate the hazard ratio and likelihood ratio test statistic of the trial

Calculate the overall Power

Stop
Appendix 3B. Simulation program code

#==========PROGRAM START==========#
library(survival) # load R package for survival analysis
library(R.oo) # load R package for object-oriented simulations
source("RPPD.r") # source 'RPPD.R' function used in simulation
patients <- list() # define patient object as a list (Global Variable)
EXPONENTIAL = 1
WEIBULL = 2
LOGNORMAL = 3
EXP_PHASE = 1
PLACEBO_PHASE = 0
FILE = 1
SCREEN = 2
RPPD = 1
PARALLEL = 2

#==========CLINICAL TRIAL DESIGN MODIFIABLE
PARAMETERS==========#
simulation = PARALLEL # the trial design is 'PARALLEL' OR 'RPPD'
placebo_distribution = EXPONENTIAL # placebo treatment distribution is 'EXPONENTIAL' or 'WEIBULL' or 'LOGNORMAL'
exp_distribution = EXPONENTIAL # experimental treatment distribution is 'EXPONENTIAL' or 'WEIBULL' or 'LOGNORMAL'
MAX_PLACEBO_LENGTH = 60 # maximum placebo phase duration for the RPPD design in days
ACCRUAL_PHASE = 90 # accrual time in days
MAX_EXP_LENGTH = 275 # maximum duration for the experimental treatment in days
CYCLE_TIME = 14 # interval duration for patient follow-up visits in days
NUM_PATIENT = 8 # total number of patients in the study
num_trials = 100 # number of simulated trials/iterations
alpha=0.05 # Type I error rate
exp_drop_out = 0.10 # patient drop-out probability (random) for the experimental group
control_drop_out = 0.10 # patient drop-out probability (random) for the control group
rppd_accrual_prob = 1.0 # patient accrual probability in the RPPD design
parallel_accrual_prob = 1.0 # patient accrual probability in the parallel group design

baseline_lambda = 0.00195 # baseline (daily) lambda for the exponential distribution
treatment_lambda = 0.0165 # experimental treatment lambda (daily) for the exponential distribution

baseline_d_shape = 3 # baseline shape for the Weibull for mean logT for the lognormal distribution
baseline_d_scale = 0.021 # baseline scale for the Weibull or SD of logT for the lognormal distribution
treatment_d_shape = 3.74 # experimental treatment shape for the Weibull or mean logT for the lognormal distribution

treatment_d_scale = 0.95 # experimental scale for the Weibull or SD of logT for the lognormal distribution

output_option = SCREEN # display the results on the 'SCREEN' or save in the 'FILE'
output_file = "output.txt" # specify the output file name if 'output_option = FILE'

#==========SEED OPTIONS==========#

seed = Sys.time() # set the seed equal to the system time
n_seed = as.numeric(seed) # convert the seed to the numeric value
output("Seed: ",n_seed, ",\n") # print the seed
set.seed(as.numeric(seed)) # set the randomization seed equal to system time

#==========PLACEBO PHASE DURATION DEPENDING ON THE RCT DESIGN==========#

max_exp_cycle= MAX_EXP_LENGTH%/%CYCLE_TIME # calculate the number of cycles in the experimental phase

if (simulation == RPPD) {
  max_placebo_cycle= MAX_PLACEBO_LENGTH%/%CYCLE_TIME # calculate the number of cycles/patient visits in the placebo-phase of the RPPD
  accrual_prob = rppd_accrual_prob # assign accrual probability for the RPPD
} else if (simulation == PARALLEL) {
  MAX_PLACEBO_LENGTH = MAX_EXP_LENGTH + 1
  max_placebo_cycle= MAX_PLACEBO_LENGTH%/%CYCLE_TIME # calculate the number of cycles/patient visits in the placebo group of the parallel group RCT
  accrual_prob = parallel_accrual_prob # assign accrual probability for the parallel group RCT
}

start_simulation () # run the simulation

#==========PROGRAM END==========#
Appendix C. ‘RPPD.R’ function

#==========PROGRAM START==========#

#==========FUNCTION TO DISPLAY RESULTS/PARAMETERS==========#
output <- function(msg1='', msg2='', msg3='', msg4='', msg5='', msg6='', msg7 = '', msg8 = '',
msg9 = '', msg10 = '', msg11 = '', msg12 = '') {
  if (output_option == FILE)
    cat(paste(msg1, msg2, msg3, msg4, msg5, msg6, msg7, msg8, msg9, msg10, msg11, msg12,
sep =""), file=output_file, append = TRUE)
  else if (output_option == SCREEN)
    cat(paste(msg1, msg2, msg3, msg4, msg5, msg6, msg7, msg8, msg9, msg10, msg11, msg12,
sep =""))
}

#==========DEFINE CHARACTERISTICS OF THE "Patient" - OBJECT==========#
setConstructorS3("Patient", function(patientID, lambda, d_scale, d_shape,
placebo_phase_length, status, survival_time, probSurv, accrual_time, drop_out, is_enrolled,
phase) {
  if (missing(patientID)) patientID <- NA;
  if (missing(lambda)) lambda <- NA;
  if (missing(d_scale)) d_scale <-NA;
  if (missing(d_shape)) d_shape <-NA;
  if (missing(probSurv)) probSurv <- NA;
  if (missing(placebo_phase_length)) placebo_phase_length <- NA;
  if (missing(status)) status <- NA;
  if (missing(survival_time)) survival_time <- NA;
  if (missing(accrual_time)) accrual_time <- NA;
  if (missing(drop_out)) drop_out <- NA;
  if (missing(is_enrolled)) is_enrolled <- NA;
  if (missing(phase)) phase <- NA;
  extend(Object(), "Patient",
    .patientID=patientID,
    .lambda=lambda,
    .probSurv=probSurv,
    .placebo_phase_length=placebo_phase_length,
    .d_shape = d_shape,
    .d_scale= d_scale,
    .status=status,
    .survival_time=survival_time,
    .accrual_time=accrual_time,
    .drop_out=drop_out,
    .is_enrolled=is_enrolled,
    .phase=phase
  )
})
# ================ SIMULATE ALL PATIENTS FOR ALL TRIALS =============== #

runTrial <- function() {
  lrt_list <- list() # store the likelihood Ratio Test (LRT) statistic from each trial in a list
  realTrialNum = 1 # set the real trial number as one

  for (j in 1:num_trials) { # initialize trials
    if ((realTrialNum == 1) || (realTrialNum %% 10 == 0) || (!is.null(j))) # print the trial number for every i-th iteration to observe simulation progress
      output("\nTrial ", realTrialNum, "-----------------------------\n")

    allInOneGroup = TRUE # set 'all patients are randomized to one group' as true

    for (i in 1:NUM_PATIENT) { # initialize patients
      patients[[i]] <<- Patient()
      patients[[i]].patientID <<- i
      patients[[i]].is_enrolled <<- ifelse(runif(1) < accrual_prob, 1, 0) # assign accrual probability
      patients[[i]].placebo_phase_length <<- ifelse(runif(1) < 0.5, 0, MAX_PLACEBO_LENGTH) # assign placebo phase duration
      patients[[i]].phase <<- ifelse(patients[[i]].placebo_phase_length == 0, EXP_PHASE, PLACEBO_PHASE) # assign treatment arms

      if (patients[[i]].is_enrolled == 1) { # if the patient is enrolled
        patients[[i]].status <<- 0 # assign initial values for status, survival time and survival probability
        patients[[i]].survival_time <<- 0
        patients[[i]].probSurv <<- 1

        # Check if all patients are in the same group
        if (allInOneGroup == TRUE && i!=1 && patients[[i]].placebo_phase_length!=patients[[i-1]].placebo_phase_length)
          allInOneGroup = FALSE # Assign distribution parameters based on treatment allocation at enrollment
          patients[[i]].lambda <<- ifelse(patients[[i]].placebo_phase_length==0, treatment_lambda, baseline_lambda)
          patients[[i]].d_scale <<- ifelse(patients[[i]].placebo_phase_length==0, treatment_d_scale, baseline_d_scale)
          patients[[i]].d_shape <<- ifelse(patients[[i]].placebo_phase_length==0, treatment_d_shape, baseline_d_shape)
          patients[[i]].accrual_time <<- runif(1, 0, ACCRUAL_PHASE) # assign accrual time
        patients[[i]].drop_out <<- 0 # assign initial drop-out status
      }
    }

    if (allInOneGroup == FALSE) { # if patients are randomized to different arms
      runPhase() # simulate current trial
      allsameSurvivaltime = TRUE # set 'all patients have the same survival time' as true
    }
  }
}

# ================ SIMULATE ALL PATIENTS FOR ALL TRIALS =============== #
if (patients[[i]]$.survival_time != patients[[i-1]]$.survival_time) {  # compare survival times
    allsamesurvivaltime = FALSE
    break
}
}

if (allsamesurvivaltime == FALSE) {  # if all patients do not have the same survival time
    lrt_list[[realTrialNum]] = fitCox(realTrialNum)  # add the LRT from the current trial to the list
    realTrialNum = realTrialNum + 1  # add the trial to real trial number after removing trials that are 'allInOneGROUP' or have 'allsamesurvival'
}
}
}

output("\n\n~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
\n")

# Calculate the Power of the simulation scenario after all iterations
if (simulation == RPPD) {
    lrt_list_count = 0
    for (k in 1:(realTrialNum-1)) {
        # Check if the current trial has a hazard ratio < 1 (observed difference) and LRT equal or greater than the critical value of the Chi square test for 1 df and one-sided alpha
        if ((lrt_list[[k]][[2]] < 1) && (lrt_list[[k]][[1]] >= qchisq(1-2*alpha, 1))){
            lrt_list_count = lrt_list_count + 1  # add to LRT list if condition satisfied
        }
    }
    coxPowerRPPD = lrt_list_count /(realTrialNum-1)#calculate the power of the simulation scenario for the RPPD
    output("RPPD Cox Power: ", coxPowerRPPD, "\n")
} else if (simulation == PARALLEL) {
    lrt_list_count = 0
    for (k in 1:(realTrialNum-1)) {
        # Check if the current trial has a hazard ratio < 1 (observed difference) and LRT greater than the critical value of the Chi square test for 1 df and one or two-sided alpha
        if ((lrt_list[[k]][[2]] < 1) && (lrt_list[[k]][[1]] >= qchisq(1-alpha, 1))){
            lrt_list_count = lrt_list_count + 1  # add to LRT list if condition satisfied
        }
    }
    coxPowerParallel = lrt_list_count/(realTrialNum-1) # calculate the power of the simulation scenario for the parallel group RCT
    output("Parallel Cox Power: ", coxPowerParallel, "\n")
}
runPhase <- function() {
  for (j in 1:NUM_PATIENT) {
    if (patients[[j]]$.is_enrolled == 1) { # for all enrolled patients
      for (i in 1:max_exp_cycle) { # initialize cycles/visits
        if (i == 1) {
          patients[[j]]$.drop_out <<- 0 # assign a drop-out probability of 0 at the entry
        } else {
          if (patients[[j]]$.placebo_phase_length == 0) { # assign a drop-out probability
            drop_out = exp_drop_out
          } else {
            drop_out = control_drop_out
          }
          patients[[j]]$.drop_out <<- ifelse(runif(1) < drop_out, 1, 0) # calculate the drop-out probability
        }
      }
      if (patients[[j]]$.drop_out == 0) { # if the patient did not drop-out at the current cycle
        findprobSurv(i, j) # calculate the probability of survival
        patients[[j]]$.status <<- ifelse(runif(1) < patients[[j]]$.probSurv, 0, 1) # define the response status
        if (patients[[j]]$.status == 1) { # if the patient status is 1 (patient responded)
          patients[[j]]$.survival_time <<- CYCLE_TIME*i # calculate the survival time
          break # go to next patient
        }
      }
    }
  }
}

if (i==max_exp_cycle && patients[[j]]$.status==0) { # if the trial ends and the status is 0
  patients[[j]]$.survival_time <<- i * CYCLE_TIME # calculate the survival time
  patients[[j]]$.drop_out <<- 1 # censor the patient
  break # go to next patient
}

else if (patients[[j]]$.drop_out == 1) { # if the patient is censored/drops out from the current cycle
  patients[[j]]$.survival_time <<- CYCLE_TIME*i # calculate the survival time
  break # go to next patient
}
#==========CALCULATE THE PROBABILITY OF SURVIVAL FOR EACH PATIENT==========#

findprobSurv <- function(current_cycle_num, patientIndex) {
  # Check if the patient was assigned to placebo and finished the placebo-phase
  if (current_cycle_num == (patients[[patientIndex]]$.placebo_phase_length%/%CYCLE_TIME + 1) & & patients[[patientIndex]]$.placebo_phase_length != 0) {
    patients[[patientIndex]]$.phase <<- EXP_PHASE  # change to experimental phase (happens only in the RPPD)
    # Change the distribution parameters from baseline to treatment specific
    if (exp_distribution == EXPONENTIAL)
      patients[[patientIndex]]$.lambda <<- treatment_lambda
    else if (exp_distribution == WEIBULL || exp_distribution == LOGNORMAL) {
      patients[[patientIndex]]$.d_scale <<- treatment_d_scale
      patients[[patientIndex]]$.d_shape <<- treatment_d_shape
    }
  }

  # Assign the survival distribution to each arm
  if (patients[[patientIndex]]$.phase == EXP_PHASE)
    distribution_model = exp_distribution
  else if (patients[[patientIndex]]$.phase == PLACEBO_PHASE)
    distribution_model = placebo_distribution

  # Calculate the conditional probability of survival using S(t+1)/S(t) formula according to distribution
  # For exponential distribution, S(T) = exp(-lambda*t) where lambda is the daily hazard and t is the time
  if (distribution_model == EXPONENTIAL)
    patients[[patientIndex]]$.probSurv <<- exp((-1)*(patients[[patientIndex]]$.lambda*CYCLE_TIME))
  # For Weibull distribution, S(t) = exp[-(scale*t)^shape]
  else if (distribution_model == WEIBULL)
    patients[[patientIndex]]$.probSurv <<- (exp(-
                                             (patients[[patientIndex]]$.d_scale*((current_cycle_num+1)*CYCLE_TIME))^patients[[patientIndex]]$.d_shape))/(exp(-
                                             (patients[[patientIndex]]$.d_scale*(current_cycle_num*CYCLE_TIME))^patients[[patientIndex]]$.d_shape))
  # For lognormal distribution, S(t) = 1-((log(exp(-mean)*t))/SD) where G is the cumulative distribution function of a standard normal variable
  else if (distribution_model == LOGNORMAL) {
    a<-exp(-patients[[patientIndex]]$.d_shape)
    patients[[patientIndex]]$.probSurv <<- (1-
                                             pnorm(log(a*(current_cycle_num+1)*CYCLE_TIME)/patients[[patientIndex]]$.d_scale))/(1-
                                             pnorm(log(a*(current_cycle_num*CYCLE_TIME)/patients[[patientIndex]]$.d_scale))
  }
}

111
# STORE PATIENT INFORMATION AND CALCULATE THE HR & LRT FOR EACH TRIAL #

`fitCox <- function(realTrialNum) {
  # Create an empty data frame
  survival.data <- data.frame(time=numeric(length(patients)), status=numeric(length(patients)),
                              group=numeric(length(patients)))
  counter = 1
  # Store the information of each enrolled patient into the data frame
  for(i in 1:length(patients)) {
    if (patients[[i]]$.is_enrolled == 1) {
      survival.data[counter,] <- c(patients[[i]]$.survival_time,
                                   patients[[i]]$.status,patients[[i]]$.placebo_phase_length)
      counter = counter + 1
    }
  }
  # Fit the Cox proportional hazard model using 'coxph' function from R
  coxfit <- coxph(Surv(time, status)~group, data=survival.data)
  HazardRatio <- exp(coxfit$coef) # calculate the Hazard ratio
  loglik <- coxfit$loglik # obtain the log-likelihood statistic from the model
  warning("Convergence problem in 'loglik'", call. = FALSE)
  lrt <- -2*(loglik[1]-loglik[2]) # calculate the LRT statistic
  retval<-data.frame(lrt,HazardRatio) # store the values to perform power calculation
  return(retval)
}

start_simulation <- function () {
  # start the simulation of the current scenario
  # DISPLAY PARAMETERS USED IN CURRENT SIMULATION #

  output("n~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
"
"
")
  if (simulation == RPPD) {
    output("RPPD Design"
")
  } else if (simulation == PARALLEL){
    output("Parallel Design"
")
  if (exp_distribution == EXPONENTIAL) {
    output("Distribution for experimental phase/group: ")
    output("Exponential Distribution"
")
    output("Daily Treatment Lambda: ", treatment_lambda, "\n")
  } else if (exp_distribution == WEIBULL) {
    output("Distribution for experimental phase/group: ")
    output("Weibull Distribution"
")
    output("Shape Parameter: ", treatment_d_shape, "\n")
    output("Scale Parameter: ", treatment_d_scale, "\n")
  } else if (exp_distribution == LOGNORMAL) {
    output("Distribution for experimental phase/group: ")
    output("Lognormal Distribution"
")
    output("Mean: ", treatment_d_shape, "\n")
  }
output("SD: ", treatment_d_scale, ",\n")
}
if (placebo_distribution == EXPONENTIAL) {
    output("Distribution for placebo phase/group: ")
    output("Exponential Distribution\n")
    output("Daily Baseline Lambda: ", baseline_lambda, ",\n")
} else if (placebo_distribution == WEIBULL) {
    output("Distribution for placebo phase/group: ")
    output("Weibull Distribution\n")
    output("Shape Parameter: ", baseline_d_shape, ",\n")
    output("Scale Parameter: ", baseline_d_scale, ",\n")
} else if (placebo_distribution == LOGNORMAL) {
    output("Distribution for placebo phase/group: ")
    output("Lognormal Distribution\n")
    output("Mean: ", baseline_d_shape, ",\n")
    output("SD: ", baseline_d_scale, ",\n")
}
output("Total Sample Size: ", NUM_PATIENT, ",\n")
output("Number of Trials: ", num_trials, ",\n")

output("~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~\n")
runTrial()

#==========PROGRAM END==========#
Appendix 4. Sample size and power calculations by different programs

Calculation parameters

Time-to-event distribution: exponential in both groups;

Study design: parallel groups RCT;

Baseline daily hazard (placebo): $\lambda = 0.00195$ (median time to response $= 365$ days);

Experimental treatment daily hazard (low potency drug): $\lambda = 0.00825$ (median time to response $= 84$ days);

Accrual period: 90 days; Treatment duration: 275 days;

Number of simulations for the parallel RCT: 1000 trials;

Alpha: 0.05 (two-sided test);

Calculators: R program (Appendix 3); PS – power and sample size calculator developed by Dupont and Plummer [69] and SSP – an internet-based power and sample size calculator which uses a method described by Lachin and Folkes [51, 70].
Appendix 5. Value of information: Components and formulas

The following information has been developed from the published articles by Willan, Pinto and Eckermann [75, 76, 78].

**Incremental Net Benefit**

Incremental net benefit denoted $b$, is the net benefit created by one treatment (for example, experimental treatment or $T$) over another (standard treatment or $S$) comparing both costs and effectiveness. It is defined as $b = \Delta e \lambda - \Delta c$, where $\lambda$ is the willingness-to-pay for an additional unit of effectiveness, $\Delta e$ is the difference in effectiveness, and $\Delta c$ is the difference between the treatment costs. Effectiveness can be expressed in events averted, additional QALYs, life years gained, etc. In this way, $\Delta e \lambda$ will be the expected increase in the value of effectiveness expressed in monetary terms and the subtraction of $\Delta c$ (the expected difference in costs) will yield the incremental net benefit of adopting $T$ over $S$.

INB is positive if either $\Delta e$ is sufficiently positive or $\Delta c$ sufficiently negative. When INB is positive, the optimal decision is to adopt $T$. In contrary, when $b$ is negative, continuing the standard treatment will maximize the net benefit for future patients. The opportunity loss function is depicted in Figure 1.

**Figure 1. Opportunity loss function**
**Expected Value of Perfect Information**

Assuming that the prior distribution of INB is normal with mean $b_0$ and variance $v_0$, and ignoring discounting for future benefits, the per patient expected opportunity loss of the decision rule (adopting T if $b>0$, retaining S, otherwise) is measured by:

$$EPVI_0 = D(b_0, v_0) = \int_{-\infty}^{\infty} \left[ I(b_0 > 0)(g(b) - b) + I(b_0 \leq 0)g(b) \right] f_{\theta}(b) \, db$$

$$= \left[ \frac{v_0}{(2\pi)} \right]^{\frac{1}{2}} \exp \left[ -\frac{b_0^2}{2v_0} \right] - b_0 \left[ \Phi \left( -\frac{b_0}{\sqrt{v_0}} \right) - I(b_0 \leq 0) \right]$$

where the function $g(.)$ is such that $g(x) = x$ if $x>0$ and 0 otherwise, I(.) is the indicator function and $\Phi(.)$ is the cumulative distribution function for a standard normal random variable. The variance of INB is calculated by $v_0 = \lambda^2 v(\hat{\Lambda}_v) + v(\hat{\Lambda}_v) - 2\lambda\text{cov}(\hat{\Lambda}_v, \hat{\Lambda}_v)$

The expected value of perfect information (EVPI) or the expected opportunity loss (EOL) is the expected opportunity loss of the best decision that can be made based on current information. It is called EVPI because the loss can be avoided in the presence of the perfect information about $b$ (i.e. $v_0 = 0$). A future trial will obtain a new INB in order to make a better decision about the treatment under study.

**Figure 2. Positive but uncertain INB prior to a trial**

Let the new trial have a sample size of $n$ per arm. Let $\hat{b}$ be the estimator of $b$ applying the data from the new trial. Assuming normality, $\hat{b} \sim n(b_0, v_0 + 2\sigma^2 / n)$ with density function $f(\cdot)$ and between-patient variance of $\sigma^2$. The posterior distribution of $b$ with the density function $f_1(\cdot)$,
mean $b_1 = v_1((b_0 / v_0) + n\hat{b} / 2\sigma^2)$ and variance $v_1 = ((1 / v_0) + (n / 2\sigma^2))^{-1}$ is presented in Figure 3.

**Figure 3. Posterior density after new trial**

The per patient post trial EOL can be expressed as:

$$EOL_{pp1} = D(b_1, v_1) = \int_{-\infty}^{\infty} \{I(b_1 > 0)(g(b) - b) + I(b_1 \leq 0)g(b)\}f_1(b)db.$$  The reduction in the expected value of information after conducting the new trial is measured as $EOL_{pp0} - EOL_{pp1}$ and is referred as the Expected Value of Sample Information (EVSI). EVSI is the reduction of the per-patient opportunity loss the new trial provides. It is a function of $n$, $\hat{b}$ and is given by:

$$EVSI(n, \hat{b}) = N[D(b_0, v_0) - D(b_1, v_1)]$$ where $N$ is the number of patients who would benefit from the decision calculated depending on the disease progression, its prevalence and incidence (Table 1).

$$E_\hat{b} EVSI(n, \hat{b}) = \int_{-\infty}^{\infty} EVSI(n, \hat{b}) \tilde{f}(\hat{b}) d\hat{b} = I_1 + I_2 + I_3$$

$$I_1 = \sqrt{\frac{2v_0}{\pi}} \frac{\sigma^2 e^{\left(-\frac{\hat{b}^2}{2v_0}\right)}}{n\sigma_b^2}, \text{ where } \sigma_b^2 = V(\hat{b}) = v_0 + 2\sigma^2 / n$$

$$I_2 = -b_0\Phi\left(\frac{b_0}{\sqrt{v_0}}\right) + v_0^{2/3} e^{\left(-\frac{\hat{b}^2}{2v_0}\right)} \frac{\sigma^2}{\sigma_b^2 \sqrt{2\pi}}$$
\[ I_3 = b_0 \Phi \left( -b_0 \sqrt{\frac{\sigma_b^2}{V_0}} \right) = \frac{v_0 e^{-\frac{b_0^2 \sigma_b^2}{2V_0}}}{\sqrt{2\pi \sigma_b^2}} \]

The optimal sample size is that which maximizes the difference between the information provided by the trial (EVSI) and the trial total cost.

**Total Trial Cost**

**Trial total cost** $TC = C_f + 2nC_v + \tilde{n} \cdot |b_0|

- trial fixed cost ($C_f$) or the cost to set up, run and report the trial;
- variable cost ($C_v$) measured by the total number of patients ($2n$ in two arm RCTs) times per patients cost for accrual, treatment and follow-up;
- the opportunity cost for $\tilde{n}$ patients receiving the less favorable treatment based on current knowledge, which is calculated depending on the disease progression, its incidence and prevalence (see Table 1).

**Table 1. Benefits and opportunity loss associated with the new trial**

<table>
<thead>
<tr>
<th>Acute disease treatment ($p = 0$)</th>
<th>Chronic disease (person-years)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$fp &gt; 2n$</td>
</tr>
<tr>
<td><strong>Number of patients who would benefit (N):</strong> all new cases that occur after the trial</td>
<td>$(h - (t_a + \tau))k$</td>
</tr>
<tr>
<td><strong>Opportunity costs ($\tilde{n} \cdot b_0$)</strong></td>
<td>$\left[(p - n)\tau + k\tau \cdot \frac{\tau}{2}\right]b_0$</td>
</tr>
</tbody>
</table>

where $(p - n)$ are...
prevalence cases receiving S, $k\tau$
are incident patients receiving S during the follow-up

where $p(1-f) + \frac{fp}{2}$ are prevalent patients receiving S, $kt_a - (2n - fp)/2$
incident patients receiving S, $\kappa\tau$ are incident patients receiving S during follow-up, and $f = a/k$ is the accrual fraction.

## Assumption
considering staggered entry, all new cases that happen during the time horizon ($h$), are assigned the average half-duration times.

*Variables in the table: $p$ = number of prevalent cases at the time of the decision, $k$ = annual incidence, $h$ = time horizon, $t_a$ = accrual time calculated as $2n/a$, where $a$ is the annual rate of accrual depending on the incidence and patients’ willingness to enroll, $\tau$ = duration of follow-up in years plus the time of data analysis.

## Expected Net Gain
The Expected net gain (ENG) is the difference between the EVSI and the trial TC measured as $ENG(n) = EVSI(n) - TC(n)$. The optimal sample size is that positive value of $n$ that maximizes this difference. TC and EVSI can be calculated and plotted as a function of the sample size (Figure 4).

If for all the values of $n$, the trial cost is greater than the EVSI, no trial would provide information whose value is greater than what it will cost, and the optimal decision should be based on current information. If, however, for some values of $n$ the EVSI is greater than the trial TC, there is not enough existing information to maximize net-benefit, and a new trial should be done. In this case, the optimal sample size is the one that maximizes the difference between EVSI and TC. The optimal sample size is the point in $E_b.EVSIs(n,\hat{b})$ that has a tangent parallel to the slope of TC (shown as $n^*$ on the Figure 4).
Figure 4. Expected total cost and expected value of information as a function of sample size

Discounting future costs and opportunity losses

The following formulas reflect discounting ($r$) of future benefits, costs and opportunity losses:
Assuming a discount rate of $r > 0$, the expression for expected net gain becomes:

$$\text{EVSI} = \left\{ (t^L - t)(1 + r)^{-t^H} + \sum_{t=1}^{\hat{t} - 1} (1 + r)^{-t} \right\} k (\text{EOLpp}_0 - \text{EOLpp}_1)$$

Financial Costs = \left\{ C_f + \left[ (t_\alpha - t^L)(1 + r)^{-t_\alpha} + \sum_{t=0}^{\hat{t}_\alpha - 1} (1 + r)^{-t} \right] a C_v \right\}

Opportunity Costs = \left\{ (t - t^L)(1 + r)^{-t} + \sum_{t=0}^{\hat{t} - 1} (1 + r)^{-t} \right\} k b_0

\left[ - \left[ (t_\alpha - t^L)(1 + r)^{-t_\alpha} + \sum_{t=0}^{\hat{t}_\alpha - 1} (1 + r)^{-t} \right] \right] \left( a / 2 \right) b_0$

where $t = 2n / a + \tau$ is the trial duration; $t^L$ is the integer part of $t$; $t^H = t^L + 1$; $t_\alpha = 2n / a$ is the duration of accrual; and, $t^L_\alpha$ is the integer part of $t_\alpha$.

$k$ – annual incidence

$a$ – annual accrual

$h$ – time horizon

$\tau$ – time (in years) from the end of accrual to when the results are available

EOLpp$_0$ – expected opportunity cost per patient, pre-trial

EOLpp$_1$ – expected opportunity cost per patient, post-trial
Appendix 6. The R code for VOI calculations

cat(paste("Comparison 1: Alternate day prophylaxis (AP) vs. tailored prophylaxis (TP)\n"))
#-------------------Variables and Constants----------------------------------------#
fCost = 1000000 # 1000000 # 2000000 # 5000000-- for sensitivity analyses
vCost = 5000 # 5000 # 10000 # 2500 -- for sensitivity analyses
cat(paste("Fixed Cost: ",fCost, ",\n"))
cat(paste("Variable Cost",vCost, ",\n"))
n = 100 # vary sample size per group from 1 to 100
k = 240 # annual incidence of hemophilia, US & Canada
TimeHor = 20 # 20 year # 25 # 15 -- for sensitivity analyses
cat(paste("Time horizon: ",TimeHor, ",\n"))
EnrFrac = 0.20 # 0.15 # 0.25 -- for sensitivity analyses
cat(paste("Enrollment fraction: ",EnrFrac, ",\n"))
rho = -0.04# calculated in the Canadian cohort, point-biserial
FollTime = 6 # 5 years of follow-up time plus 1 year for analysis & report
AccrRate = EnrFrac * k # a=f*k
FactorUnitCost = 1 # USD$1 per unit of FVIII
discRate=0 # no discounting # 0.05 # 0.03 -- for sensitivity analyses
EffT = 0.925925926 # % with no MRI detected joint damage 5 years for AP group
cat(paste("Proportion with no joint damage in AP group: ",EffT, ",\n"))
EffS = 0.7916667 # % with no MRI detected joint damage 5 years for TP group
cat(paste("Proportion with no joint damage in TP group: ",EffS, ",\n"))
VarEffT = (EffT*(1-EffT))/33 # AP variance of effectiveness
VarEffS = (EffS*(1-EffS))/25 # TP variance of effectiveness
VarDeltaE = VarEffT + VarEffS # Variance of difference in effectiveness
cat(paste("VarDeltaE: ",VarDeltaE, ",\n"))
CostT = 352793*FactorUnitCost # AP group costs
CostS = 235301.8*FactorUnitCost # TP group costs
VarCostT = (150454*FactorUnitCost)^2 # AP variance of costs
VarCostS = (88833.97778*FactorUnitCost)^2 # TP variance of costs
VarDeltaC = VarCostT + VarCostS # Variance of difference in costs
cat(paste("VarDeltaC: ",VarDeltaC, ",\n"))
# for OD, Eff = 0.5517241; Cost = 113237*FactorUnitCost; VarCostS = (65494*FactorUnitCost)^2; sample size = 32

Cov_DeltaE_DeltaC = rho*sqrt(VarDeltaC*VarDeltaE) # Covariance of differences in effectiveness and costs
cat(paste("Cov_DeltaE_DeltaC: ",Cov_DeltaE_DeltaC, ",\n"))

lambda =1000000# USD$ # threshold value/WTP to avoid MRI detected joint damage
cat(paste("lambda: ",lambda, ",\n"))
b0=(EffT-EffS)*lambda-(-(CostT-CostS)) #Prior INB (based on current information)
b0=-(EffT-EffS)*lambda-(-(CostT-CostS)) #Prior INB (change the sign when EffS>EffT)
\[ v_0 = \lambda^2 (\text{VarDeltaE}) + \text{VarDeltaC} - 2\lambda \text{Cov}_{\Delta \text{E} \Delta \text{C}} \]  

\[ \text{SigmaSqT} = \lambda^2 (\text{VarEffT})^3 + \text{VarCostT}^3 - 2\lambda \rho \sqrt{\text{VarEffT}(\text{VarCostT})^3} \]  

\[ \text{SigmaSqS} = \lambda^2 (\text{VarEffS})^2 + \text{VarCostS}^2 - 2\lambda \rho \sqrt{\text{VarEffS}(\text{VarCostS})^2} \]  

\[ \text{sigmaSqP} = 1/2 (\text{SigmaSqT} + \text{SigmaSqS}) \]  

\[ \text{EVPI0} = (\sqrt{v_0/(2\pi)}*\exp(-b_0^2/(2v_0))) - b_0 (\text{pnorm}(-b_0/sqrt(v_0)) - (\text{as.numeric}(I(b_0<=0)))) \]  

\[ \text{EVSI} = N*(\text{EVPI0} - \text{EPVI1}) \]  

\[ \text{EPVI1} = I_1 + I_2 + I_3 \]  

\[ \text{EVSI} = N*(\text{EVPI0} - \text{EPVI1}) \]
tCost = fCost+2*i*vCost+nhat*b0 # expected total cost of trial
OppCost = nhat*b0 # opportunity cost portion of ETC
#cat(paste(i, "\t"))
#cat(paste("Opportunity cost: ",OppCost, "\n"))

ENG = EVSI - tCost # find the expected net gain

#==== Calculations considering discounting======================
trialT = as.integer(trialTime)
accD = as.integer(AccrDur)
discEVSI = (((trialT+1-trialTime)*((1+discRate)^(-trialT)))+ (((1+discRate)^(1-trialT+1)))-(1+discRate)^(1-TimeHor))/discRate)*k*(EVPI0-EPVI1)
discOpCost = ((((trialTime-trialT)*((1+discRate)^(-trialT))) + (1+discRate - (1+discRate)^(1-trialT)))/discRate)*k*b0 - (((AccrDur - accD)*((1+discRate)^(1-accD)))+((1+discRate - (1+discRate)^(1-accD)))/discRate)*(AccrRate/2)*b0
#cat(paste("Discounted Opportunity cost: ",discOpCost, "\n"))
discCost = fCost + (((AccrDur-accD)*((1+discRate)^(1-accD)))+ (1+discRate-((1+discRate)^(1-accD))))*vCost*AccrRate + discOpCost

#====Store data for each sample size===========================
EVI.data[count,1]= i
EVI.data[count,2]= EVSI
EVI.data[count,3]= discEVSI
EVI.data[count,4]= tCost
EVI.data[count,5]= discCost
EVI.data[count,6]= ENG

# plot EVSI, ETC, & ENG
attach(EVI.data)
par(mfrow=c(1,1))
plot(sampleSize,EVSIf/1000000, type='l', xlab="n", ylab="Cost(mln)", xlim=c(0,n),ylim = c(-200,500), main = "EVSI and ETC", col = "red")
lines(sampleSize,totalCost/1000000, col = "blue")
plot(sampleSize,ExpNG/1000000,axes = T, type='l', xlab="n", ylab="Cost in mln(CAD)", xlim=c(0,n), ylim =c(-100, 300), main = "Expected Net Gain", col = "red")
abline(0,0, lty=5)

# display the optimal n, or n that maximized the ENG
EVI.data1 <- EVI.data[order(EVI.data,"ExpNG"), decreasing=TRUE,]
cat(paste("----------------------------------------------------------\n"))
print(EVI.data1 [1,])
cat(paste("----------------------------------------------------------\n"))
detach(EVI.data)

runEVI() # run the function for specified parameters