**PhD Dissertation**

**Statistical Issues in the Design and Analysis of Early Phase Proof of Concept Clinical Trials in Rheumatoid Arthritis**

by

**Feng Liu**

**A Thesis submitted in fulfilment of the requirement for the**

**Degree of Doctor of Philosophy in the Medical Statistics, School of Health and Related Research, University of Sheffield**

**August 2017**

**Table of Contents**

[1. Introduction 9](#_Toc489029476)

[1.1. Introduction to the Drug Development Process 10](#_Toc489029477)

[1.2. Introduction to Rheumatoid Arthritis and the Treatment Being Evaluated 12](#_Toc489029478)

[1.3. Research Questions 13](#_Toc489029479)

[1.4. Rationale for the Research 14](#_Toc489029480)

[1.5. The Objective of the Dissertation 16](#_Toc489029481)

[1.6. Outline of Dissertation 18](#_Toc489029482)

[2. Common study designS in TRIALS of PATIENTS WITH Rheumatoid Arthritis 22](#_Toc489029483)

[2.1. The Aims of the Chapter 22](#_Toc489029484)

[2.2. Reporting Disease Activity in Clinical Trials for Rheumatoid Arthritis 22](#_Toc489029485)

[2.2.1. Composite Assessment 23](#_Toc489029486)

[2.2.1.1. Disease Activity Score based on 28 Joint Count (DAS28) 23](#_Toc489029487)

[2.3. Common Designs in Dose Selection of RA trials 25](#_Toc489029488)

[2.4. Summary 26](#_Toc489029489)

[3. Literature review OF Adaptive designS 27](#_Toc489029490)

[3.1. The Aims of the Chapter 27](#_Toc489029491)

[3.2. Motivation for using Adaptive Design in Drug development 28](#_Toc489029492)

[3.2.1. Efficient Design 29](#_Toc489029493)

[3.3. Types of Adaptation in Clinical Trials 33](#_Toc489029494)

[3.4. Drop-the-loser or Pick-the-Winner design 34](#_Toc489029495)

[3.5. Group Sequential Design 35](#_Toc489029496)

[3.6. Design Modification Based on Conditional and Predictive Power 39](#_Toc489029497)

[3.6.1. Conditional Power and Predictive Power 39](#_Toc489029498)

[3.7. Drug Response Model in Drug Development 41](#_Toc489029499)

[3.7.1. Multiple Comparison Methods in Dose Response Analysis 42](#_Toc489029500)

[3.7.1.1. Multiplicity Adjustments 42](#_Toc489029501)

[3.7.2. Model-based Dose Finding 43](#_Toc489029502)

[3.7.3. Linear and polynomial regression model 43](#_Toc489029503)

[3.7.4. Parametric Emax Model 44](#_Toc489029504)

[3.7.5. Contrast Test 46](#_Toc489029505)

[3.7.5.1. Emax model using Linear Contrast Method 46](#_Toc489029506)

[3.7.5.2. Trend Test in Dose Response Analysis 47](#_Toc489029507)

[3.7.6. MCP-MOD Method 48](#_Toc489029508)

[3.7.7. Semi-Parametric Normal Dynamic Linear Model 49](#_Toc489029509)

[3.8. Existing Literature in Comparison of Emax and NDLM model 51](#_Toc489029510)

[3.9. Design based Dose Response Model 53](#_Toc489029511)

[3.10. Statistical Issues in Implementing the Adaptive Design 53](#_Toc489029512)

[3.11. Operational Issues in Implementing the Adaptive Design 54](#_Toc489029513)

[3.12. Summary 54](#_Toc489029514)

[4. ADaptive Design in Proof of concept study of GSK123456– a case study 56](#_Toc489029515)

[4.1. The Aim of the Chapter 57](#_Toc489029516)

[4.2. Study Design 57](#_Toc489029517)

[4.2.1. Design Considerations 57](#_Toc489029518)

[4.2.2. Study Design 58](#_Toc489029519)

[4.2.3. Endpoints and Hypothesis 62](#_Toc489029520)

[4.2.4. Adaptive Dose Escalation Strategies 63](#_Toc489029521)

[4.2.5. Bayesian *Emax* model 64](#_Toc489029522)

[4.3. Results of Proof of Concept Study of GSK123456 68](#_Toc489029523)

[4.4. Summary of Study Results 69](#_Toc489029524)

[4.5. Outcome of Compound GSK123456 70](#_Toc489029525)

[4.6. Adaptive Design Discussion 71](#_Toc489029526)

[5. Simulation to support Adaptive Design in Proof of concept Study in follow-on compound GSK654321 73](#_Toc489029527)

[5.1. The Aims of the Chapter 73](#_Toc489029528)

[5.1.1. Recommendation on PoC design of GSK654321 Based on Learnings from GSK123456 74](#_Toc489029529)

[5.1.2. Part A of Phase 2a PoC study of GSK654321 74](#_Toc489029530)

[5.1.3. Part B Design of GSK654321 75](#_Toc489029531)

[5.1.3.1. Simulation in Group Sequential design (Part B) 77](#_Toc489029532)

[5.1.4. Planned Statistical method at final analysis 81](#_Toc489029533)

[5.2. Summary 82](#_Toc489029534)

[6. Improving RA Proof of concept trial with A better statistical model 83](#_Toc489029535)

[6.1. The Aims of the Chapter 83](#_Toc489029536)

[6.2. Frequency of U-shaped Curve and possible explanation 83](#_Toc489029537)

[6.2.1. Literature review on the frequency of the U-shaped dose response curve 86](#_Toc489029538)

[6.2.2. The frequency of U-shaped curve in biological to treat RA patients 86](#_Toc489029539)

[6.2.3. Possible Explanation of U-shaped curve to GSK123456 88](#_Toc489029540)

[6.2.4. Retrospectively Analysis of DAS28 data using Permutation Test on Polynomial Regression model 89](#_Toc489029541)

[6.2.5. Implication of U-shaped curve to GSK123456 and GSK654321 94](#_Toc489029542)

[6.2.6. Retrospective Analysis using NDLM Model 94](#_Toc489029543)

[6.2.7. Bayesian Normal Dynamic Linear Model (NDLM) 95](#_Toc489029544)

[6.2.8. Results of Retrospective Analysis Using NDLM 96](#_Toc489029545)

[6.3. Dose response models in the evaluation 99](#_Toc489029546)

[6.3.1. Design of the Simulation Study 101](#_Toc489029547)

[6.3.2. Decision Criteria in Adaptive Design Simulation 105](#_Toc489029548)

[6.3.2.1. Simulation Results 106](#_Toc489029549)

[6.3.3. Adaptive Design Simulation 107](#_Toc489029550)

[6.3.3.1. Minimal effective dose 108](#_Toc489029551)

[6.3.3.2. Clinical significant dose (CSD) 109](#_Toc489029552)

[6.3.3.3. Scenario 2 – No Adaptive Allocation 109](#_Toc489029553)

[6.3.3.4. Scenario 3 – Half Adaptive 116](#_Toc489029554)

[6.3.3.5. Scenario 4 - Adaptive Allocation 122](#_Toc489029555)

[6.3.4. Simulation results with informative prior 129](#_Toc489029556)

[6.3.5. Simulation results with Calibrated type I error rate 131](#_Toc489029557)

[6.3.6. Assessment of Bias 136](#_Toc489029558)

[6.3.7. Summary of Model Comparison: *Emax* Model versus NDLM Model 140](#_Toc489029559)

[6.4. Discussion 142](#_Toc489029560)

[7. Improving the RA proof of concept desigN: Relationship of early response and late response 147](#_Toc489029561)

[7.1. The Aims of the Chapter 149](#_Toc489029562)

[7.2. Relationship of DAS28 at Early Time Point and Late Time Point in the PoC Study 149](#_Toc489029563)

[7.2.1. Summary of Late and Early Time Point Relationship 158](#_Toc489029564)

[7.3. Meta-analysis to investigate the Relationship of Early Time Point and Late Time Point 158](#_Toc489029565)

[7.3.1. Identification of Trials 159](#_Toc489029566)

[7.3.2. List of Literature Included in the Meta-Analysis 160](#_Toc489029567)

[7.4. Relationship of DAS28 at Early and Late Time Point based on Meta-Analysis 160](#_Toc489029568)

[7.4.1. Correlation and Simple Linear Regression 161](#_Toc489029569)

[7.4.2. Summary of Late and Early Time Point Relationship 162](#_Toc489029570)

[7.5. Relationship of early response and late response in change from baseline 166](#_Toc489029571)

[7.5.1. DAS28 Change at Early Time Point and Late Time Point in PoC Study 167](#_Toc489029572)

[7.5.2. DAS28 Change at Early Time Point and Late Time Point in meta-analysis 169](#_Toc489029573)

[7.6. Structural correlation of early response and late response at change from baseline 174](#_Toc489029574)

[7.6.1. Relationship of early response and late response in Change from baseline and original value 175](#_Toc489029575)

[7.6.2. Simulation to investigate Change from baseline and original value at meta-analysis 183](#_Toc489029576)

[7.6.3. Summary of Structural Correlation 186](#_Toc489029577)

[7.7. Correlation and causality in validation of surrogate Endpoint 188](#_Toc489029578)

[7.7.1. Validation of surrogate endpoints in Single-trial setting 190](#_Toc489029579)

[7.7.2. Validation of surrogate endpoints in multiple-trial setting 194](#_Toc489029580)

[7.7.3. Validation of DAS28 at Day 14 as Surrogate Endpoint 197](#_Toc489029581)

[7.7.4. Discussion of DAS28 at Day 14 as Surrogate Endpoint 200](#_Toc489029582)

[7.7.5. Model Parameters from Covariate Model with DAS28 Early Response as Covariate 201](#_Toc489029583)

[7.7.6. Other Models 204](#_Toc489029584)

[7.7.7. Model Comparisons 204](#_Toc489029585)

[7.8. Discussion 204](#_Toc489029586)

[8. Improving the RA proof of concept desigN USING Early time points to speed UP decisions 206](#_Toc489029587)

[8.1. The Aims of the Chapter 206](#_Toc489029588)

[8.2. Retrospectively Fitting Between Cohort Data 206](#_Toc489029589)

[8.2.1. Retrospective Model Fitting in PoC Study 207](#_Toc489029590)

[8.3. Delayed Response Predictive Model 209](#_Toc489029591)

[8.3.1. Illustrated Delayed Response Predictive Models 210](#_Toc489029592)

[8.3.2. Delayed Response Predictive Model: Mathematical Details 212](#_Toc489029593)

[8.3.2.1. Last Observation Carry Forward (LOCF) method 213](#_Toc489029594)

[8.3.2.2. Delayed Response Predictive Model 1 (DRPM1) 214](#_Toc489029595)

[8.3.2.3. Delayed Response Predictive Model 2 (DRPM2) 214](#_Toc489029596)

[8.3.2.4. Delayed Response Predictive Model 3 (DRPM3) 215](#_Toc489029597)

[8.3.3. Summary of Approaches 215](#_Toc489029598)

[8.4. Illustration of Delayed Response Predictive Models 216](#_Toc489029599)

[8.4.1. Summary of Retrospective Model Fitting in PoC Study 219](#_Toc489029600)

[8.4.2. Delayed Response Predictive Model: Simulation 220](#_Toc489029601)

[8.4.2.1. Data Simulation 220](#_Toc489029602)

[8.4.2.2. Design Scenarios and Decisions 223](#_Toc489029603)

[8.4.2.3. Simulation Results 226](#_Toc489029604)

[8.4.2.4. Type I error rate 227](#_Toc489029605)

[8.4.2.5. Sensitivity analysis 231](#_Toc489029606)

[8.4.3. Summary of Simulations 238](#_Toc489029607)

[8.5. Time-saving with the Predictive model 239](#_Toc489029608)

[8.6. Summary 240](#_Toc489029609)

[9. Discussions and Proposed Phase 2a Design for GSK654321 243](#_Toc489029610)

[9.1. The Aims of the Chapter 244](#_Toc489029611)

[9.2. Review of Research Aims 244](#_Toc489029612)

[9.3. A Better Design for the PoC Study of GSK654321 251](#_Toc489029613)

[9.4. Comparisons of Phase 2a PoC Design in RA 252](#_Toc489029614)

[9.5. Comparisons of Bayesian NDLM model 253](#_Toc489029615)

[9.6. Comparisons of Delayed Response Models 254](#_Toc489029616)

[9.7. Generalization of Delayed Response Models 255](#_Toc489029617)

[9.8. Limitations of the Research 255](#_Toc489029618)

[9.8.1. Limitations of Proposed Phase 2a Design 255](#_Toc489029619)

[9.8.2. Limitation of NDLM model and Predictive Model 255](#_Toc489029620)

[9.9. Recommendations for the Phase 2a PoC Study of GSK654321 256](#_Toc489029621)

[9.10. Extension and Future Work: 257](#_Toc489029622)

[9.10.1. Extending the NDLM Model and Prediction Model to Other Disease Areas 257](#_Toc489029623)

[10. Contribution to clinical research 258](#_Toc489029624)

[11. Appendix 262](#_Toc489029625)

[11.1. Programming codes 262](#_Toc489029626)

[11.1.1. WinBUGS Codes for Bayesian *Emax* Model 262](#_Toc489029627)

[11.1.2. WinBUGS Codes for Bayesian NDLM Model 264](#_Toc489029628)

[11.1.3. SAS Code for Permutation Test 266](#_Toc489029629)

[11.1.4. SAS Codes for DRPM Simulation 268](#_Toc489029630)

[11.1.5. R Codes for Structural Correlation at Meta-Analysis Simulation 271](#_Toc489029631)

[11.2. List of literature in the Meta-analysis 273](#_Toc489029632)

[11.3. Weighted Linear Regression ANCOVA Model Fitting and Diagnostics 274](#_Toc489029633)

[11.3.1. Weighted Linear Regression 274](#_Toc489029634)

[11.3.2. Model Fitting 277](#_Toc489029635)

[11.3.3. Model Diagnostics 279](#_Toc489029636)

[11.4. Supplemental Results for comparison of DRPM 1, 2, 3 and LOCF method in Chapter 9 283](#_Toc489029637)

[11.4.1. Supplemental Tables for Sensitivity Analysis of DRPM 1, 2, 3 and LOCF method 283](#_Toc489029638)

[11.4.2. Sensitivity Analysis on Proportion of Data to Impute 298](#_Toc489029639)

[11.5. Simulation Scenarios Based on Desired Response or Placebo Like Response 301](#_Toc489029640)

[12. Reference 304](#_Toc489029641)

**Abstract**

**Introduction:** Rheumatoid Arthritis (RA) is a chronic inflammatory disorder that typically affects joints of the body such as knees, hands and feet. The prevalence of RA is around 1% in the general population and the incidence rate is about 41 per 100,000 persons per year.

The use of adaptive designs in drug development is of increasing interests to medical researchers since adaptive design potentially delivers faster and better decisions compared to the traditional fixed sample-size designs.

In RA clinical trials, after treatment starts, follow-up and outcome assessments usually take a number of months, with some outcomes collected up to two or more years after the start of study treatment. For this reason, adaptive designs are rarely undertaken in RA clinical trials as the timing of the endpoints make it difficult to make decisions from the interim analyses.

**Aims:** The overall aim of this thesis is to assess whether or not we can apply an adaptive dose finding design, using Bayesian methods, to a Phase 2a Proof of concept (PoC) trials for the development of a new drug or pharmaceutical treatments for patients with RA disease when the dose response is likely to be non-monotonic.

**Data:** The primary data was a PoC study which was re-analysed retrospectively using an adaptive design procedure to investigate the aspects of the study design.

**Method:** In this dissertation, a Bayesian dose response model was applied to the adaptive design of a Phase 2a PoC clinical trial in RA patients. The first part of the design was to “learn” the dose response in multiple dose cohorts and the second part was to “confirm” the selected dose in a group sequential design with an O’Brien Fleming alpha spending correction.

A delayed response predictive model using outcomes collected short-term (Day 14 post-randomisation) for the prediction of longer term outcomes (Day 56 post-randomisation time points) is proposed, based on retrospective analysis and a literature based meta-analysis which was undertaken to investigate the use of early time points to predict Day 56 responses with the intent of reducing the time taken to make decisions at interim analyses of efficient design. Simulations are used to evaluate the models in the desired settings.

**Results:** It was shown that the two-part adaptive design with “learn” and “confirm” could be implemented in the Phase 2a PoC study with two potential improvements:

1) The delayed response predictive model can be used to predict Disease Activity Score (DAS28) Day 56 outcome based on Day 14 outcome to expedite the time to interim decisions in the context of a Phase 2a PoC design;

2) The Bayesian Normal Dynamic Linear Model (NDLM) can be used in the dose response analysis to handle both monotonic and non-monotonic dose responses without sacrificing statistical power or design performance.

**Conclusion:** This thesis demonstrates that it is possible to apply an adaptive design to a PoC study in the treatment of RA. It is recommended that the dose response design with Bayesian NDLM model using predicted DAS28 outcomes at Day 56 based on Day 14 data can expedite the interim decision making. In most cases the Bayesian *Emax* model works effectively and efficiently, with low bias and a good probability of success in the case of a monotonic dose response. However, if there is a belief that the dose response could be non-monotonic based on prior knowledge then the NDLM is the superior model to assess the dose response. Based on the trial design proposed if the predictive model can be applied to a future adaptive trial, there is potential for a significant time-saving in Phase 2a study.

# ****Introduction****

Drug development is a complicated process which involves clinical scientists, drug manufacturers, and developers to evaluate a test compound over several years before it reaches the market and is delivered to physicians and patients. Drug development in the pharmaceutical industry has gone through several challenges in recent years including lack of productivity, patent expiration, and increasing drug development cost (Woodcock, 2005). To improve productivity and reduce the failure rate in the drug development process, the use of innovative or adaptive clinical trial designs is of special interest due to their flexibility, and their potential to mitigate the developmental risk and to make earlier decisions about whether or not to stop or continue with further phases of development for a new compound (Pong, 2010).

There is a strong interest in the drug development of Rheumatoid Arthritis (RA) drugs, due to high unmet need in the patients (Newman, 1996; Choy, 2013). In this dissertation, we focus on statistical issues in study design of Phase 2a early phase drug development in the patients with active RA disease.

This chapter starts with an introduction to the drug development process and describes the background to RA. The chapter also highlights the roles of disease progression and their impact on the design and analysis of RA drug trials, followed by the need for new medicines to address the disease burden of patients. Finally, the rationale for the research, the research questions, and the content of the dissertation are introduced.

In summary, the aims of this chapter are to:

1. Define what RA is and introduce the background of RA disease;
2. Summarize the research questions of this dissertation, the objective of the dissertation, and the rationale for this research work;
3. Provide a synopsis of the content of all chapters that follow.

## Introduction to the Drug Development Process

Drug development can generally be divided into three stages - the non-clinical or preclinical stage, the clinical stage, and the post-approval or post-marketing stage (Adams, 2010). The first stage is the non-clinical or preclinical phase which focuses on the medicinal chemistry, drug manufacture, safety assessment, and evaluation of whether a new drug is safe in animal models. If the drug is effective in *in vitro* experiment (lab data) and has no safety concerns based on animal data (*in vivo* evidence of the drug effect in animals), an application to regulatory agencies as an investigational new drug (IND) will be filed. After the IND is approved, the drug will go through Phase 1, 2, and 3 of clinical developments which are part of the clinical stage (Adams, 2010).

During the Phase 1 trials, in most disease areas, the new drug will be tested in healthy volunteers to investigate how the drug interacts with human body, especially how much dosage the body can tolerate and what the side effects are (FDA guidance, 1999; EMEA, 2007).

Phase 2 studies are sometimes divided into Phase 2a and Phase 2b. Phase 2a is specifically designed to assess dose requirements and drug effect, i.e. researchers administer the drug to patients with active disease and evaluate whether the drug can show benefits in its target population, and this is known as Proof of Concept (PoC). Many of these Phase 2a clinical trials use biomarkers, surrogate clinical endpoints as primary endpoints. However, these Phase 2a trials are not large enough to show whether the drug will be beneficial in routine clinical use (Adams, 2010). Phase 2b is specifically designed to test how well the drug works at the prescribed dose(s) in routine clinical practice with demonstration of medicine’s efficacy. The Phase 2b trials usually use registration endpoints.

Phase 3a studies are designed to demonstrate whether the drug is efficacious compared to either placebo or other comparators. If the studies are successful, the drug company will file a new drug application (NDA) to a regulatory agency for approval. After the careful consideration of the benefits and risks of the test drug by the regulatory agency, the NDA application can either be approved for market access, rejected, or the regulatory agency might request further study or information before making a decision. Even after market access has been granted, the regulatory agency can request that the drug company conducts additional post-marketing studies, such as Phase 3b or even Phase 4 studies, for safety surveillance or additional efficacy data (FDA guidance, 1999).

A typical clinical drug-development programme (phase 1, 2, and 3 clinical trials) takes about 7 years (Sherman, 2013). For certain clinical conditions, the Food and Drug Administration (FDA) has expedited pathways available (i.e. breakthrough therapy) particularly for drugs being studied for serious or life-threatening illnesses (Darrow, 2014).

## Introduction to Rheumatoid Arthritis and the Treatment Being Evaluated

RA is a chronic inflammatory disease that typically affects joints of the body such as knees, hands, and feet (Newman, 1996). In addition to causing joint damage, RA is a leading cause of disability. The prevalence of RA is about 1% of the world’s population (Newman, 1996). Despite recent advances in the RA treatments, many RA patients still suffer from disease symptoms and disability (Newman, 1996; Choy, 2013). Therefore, there is a strong unmet medical need for new effective treatments.

GlaxoSmithKline Inc. (GSK) has developed an anti-rheumatic compound, called GSK123456, which is an anti-Oncostatin M (OSM) monoclonal antibody using a new mechanism of biological action (Choy, 2013). This new compound could potentially benefit millions of RA patients around the world if it demonstrates clinical effectiveness in reducing rheumatic symptoms in clinical trials.

A two-part adaptive Phase 2a clinical trial was designed to demonstrate whether the study compound can show “efficacy” using surrogate clinical endpoint – DAS28. The study has two-part. Part A of the Phase 2a trial uses a dose finding design to explore the dose responses to search for the optimal dose for next part of the study. Since it is also the first time that patients are treated with study drug, although safety data would have been collected from early pre-clinical and clinical studies in animals and healthy volunteers, the safety profile in RA patients was still unknown, hence the trial is also a dose escalation trial. Part B of the Phase 2a study is to confirm the effectiveness of the chosen dose.

This two-part adaptive design was successfully implemented in the Phase 2a PoC study. It was shown that GSK123456 failed to demonstrate effectiveness at the study dose in the Part B of the study (Choy, 2013). The compound GSK123456 was discontinued due to lack of efficacy. A follow-on compound of the same drug class, GSK654321, is now under development by GSK.

This dissertation is to re-visit the study design from the Phase 2a PoC trial of GSK123456 and to improve the study design based on the lessons learned from the Phase 2a study. This learning will help to reshape the study design of the follow-on compound GSK654321 in the context of Phase 2a PoC study for patients with RA disease.

## Research Questions

The overall aim of the dissertation is to determine whether or not we can apply a Bayesian dose finding design to a Phase 2a PoC trial in the development of a new drug or pharmaceutical treatment for patients with RA disease. The specific research questions are:

* Can we apply adaptive Bayesian dose finding design to Phase 2a PoC trials in the clinical development of RA drugs?
* What are the limitations of the traditional fixed sample size design in Phase 2a clinical trials in RAs?
* How does the proposed Bayesian dose finding design compare with the traditional fixed sample size design, in terms of statistical power and operating characteristics, in a Phase 2a clinical trial design in the treatment of RA disease?
* What are the statistical properties of the Bayesian dose finding design, with a Normal Dynamic Linear Model (NDLM), as compared to the Bayesian *Emax* model?
* How can we reduce the drug development time by applying a predictive model using early time point outcomes to predict later time point outcomes?

## Rationale for the Research

Drug development in the pharmaceutical industry has gone through several challenges in recent years including a lack of productivity, patent expiration, and increasing costs of developing a drug (Woodcock, 2008). It was shown in a recently published review paper (Hay, 2014) that the success rate for a drug to advance to the next phase is 64% from Phase 1 to Phase 2, and only 32% from Phase 2 to Phase 3. Therefore, there is potentially a high failure rate in drug development and increasing costs to R&D budgets, especially for drugs with novel mechanisms.

In 2004, the United States Food and Drug Administration (FDA) began a Critical Path Initiative to assist drug developers and pharmaceutical companies identify the *“scientific challenges underlying medical product pipeline problems”*, with the intent of improving the drug development process (Woodcock, 2008). Among the possible challenges, the use of innovative or adaptive clinical trial designs is of special interest to the drug developers due to the flexibility to allocate more subjects to effective drugs, allow for stopping a failed compound for futility or increasing the sample size, therefore improving the efficiency of clinical trials and reducing uncertainty (Chow, 2010).

Despite intensive research and recent developments in RA, the precise cause of RA remains elusive and there is still unmet need. Most Phase 2 (including both Phase 2a and Phase 2b) and phase 3 studies of RA diseases commonly employ a traditional design with fixed sample sizes (van Vollenhoven, 2003). One of the reasons is that the outcomes are long term endpoints (Hochberg, 2009) which make interim analysis and adaptation difficult. It may take months for decision-making at the interim analysis of adaptive design, by that time most of the patients would have already been recruited into the trial.

In clinical development of RA drug, selecting the correct dose is an important step. If the dose is too high, the dose may have unnecessary adverse events for no gain in efficacy and it is also unethical to treat patients with such high dose. If the dose is too low and though it may have a good safety profile, there is little chance of showing an effect. Phase 2 is therefore a critical to bridge the data on safety and tolerability established in Phase 1 to select a dose to carry forward to Phase 3.

A Phase 2a study would initially be undertaken to assess a wide range of doses in usually a (relatively) small number of people, and then doses are chosen from Phase 2a study to be assessed in 2b. In Phase 2b clinical trial, the sample size is relatively large and will be used to select the dose to carry forward to Phase 3. Apart from the sample size a main difference between 2a and 2b is the range of doses and the anticipated dose response. In Phase 2a, there are usually a wide range of doses so there is little anticipation that the dose response will be linear across the full spectrum of dose range and there could have higher chance to see non-monotonic dose response. While in Phase 2b trial, the doses are mainly selected from linear portion of Phase 2a dose range with the anticipation that the doses are linear. Therefore, more attention is needed to explore dose response model that can handle both monotonic and non-monotonic dose response curves. Research has been done to select correct doses in Phase 2b to Phase 3 (Temple, 2012; Bornkamp, 2007), however, little research is available to compare statistical models in Phase 2a.

Given the limitations of clinical research in RA, there is a strong need to develop new clinical study designs in Phase 2a, such as adaptive designs to expedite the decision making in the clinical development of new treatment. There are two areas of potential improvements in the clinical study design, the first area is how to the choose correct doses to inform the Phase 2a trial when there is the possibility of a non-monotonic dose response curve and the other is how to reduce decision times.

## The Objective of the Dissertation

As discussed earlier, the traditional design has its limitations. In contrast to the traditional design, an adaptive trial may be modified based on interim data i.e. stopping for futility or success or sample size re-estimation (Hung, 2005; Hung, 2006). It is well documented in the literature that the adaptive design with in-stream decision-making, not only increases the efficiency of identifying treatment benefits, but also increases the flexibility in the decision making of clinical development (Hung, 2005; Bretz, 2009; Chow, 2010).

The first objective of this dissertation is to discuss the statistical aspects of a novel adaptive dose finding design in Phase 2a of RA drug development. Recent statistical developments (Bretz, 2009; Chow, 2010) using adaptations based on treatment differences shows potential benefits of delivering drugs to market, but there are some concerns in the implementation of adaptive design. The concerns include statistical methodology to control the Type I error rate or false positive rate, time to execute the interim analysis, and the challenge of clinical operation.

The Phase 2a design was applied to patients with RA successfully in a double-blind randomized multiple centre trial of GSK123456. Although it is a failed study, there are two main learnings: the first learning is that Bayesian *Emax* model performs poorly if the dose response follows a non-monotonic curve - as was observed in the Phase 2a study of GSK123456, i.e. a U-shaped curve; the second learning is that there is a strong need to develop methods to reduce the decision and development times at interim analyses.

The second objective of this dissertation is to compare the *Emax* model and the NDLM model in the Phase 2a design with different dose response profiles, especially a dose response profile with non-monotonic changes as the dose increases. The original phase 2a PoC design was anchored with a Bayesian *Emax* model prior to my starting work on the study. An observed U-shaped dose response curve motivated the research into identifying a flexible model to select the best doses in the PoC trial. Other dose response models such as MCP-Mod methods were not considered in the comparison since the NDLM model has been reported to have higher probabilities in identifying the clinical relevant doses if the dose response followed U-shaped curve (Bornkamp 2007). The focus will be on applying the dose response to support future development of GSK654321 by understanding the design issues and developing a better statistical methodology.

The third objective of the dissertation is to develop a statistical method to predict the late time point outcome data using outcome data collected earlier to address the issue of prolonged time to interim decision-making. The method of using outcome data collected at early time points to inform the adaptation decisions will help to make quicker decisions.

Lastly, based on what are learned from the Phase 2a trial of GSK123456, a recommendation is made for the Phase 2a study design of the follow-on compound GSK654321 which is in the same drug class and currently under-development for arthritis treatment.

## Outline of Dissertation

This section introduces the background of RA disease. It starts with the definition of the disease, followed by the disease symptoms, disease disposition, and research questions and objectives of the dissertation.

The beginning of Chapter 2 briefly discusses the epidemiology of RA in the context of recent observational studies. The chapter discusses the assessment of disease activity including biomarkers, clinical activity score, patient assessment, and a physician assessment. The chapter discusses the limitations of the current drug treatments and the need for developing safer and more effective drugs for treating RA.

Chapter 2 continues to perform literature review on standard disease activity reporting and study outcomes/endpoints in different phases of clinical development, such as biomarkers, individual and composite disease activity scores, followed by the experimental designs that are commonly used in clinical trials of RA patients. RA is a chronic disease and RA clinical trials usually take more than 6 months, with some taking more than 2 years for follow-up. It can take months or even years to observe the clinical responses.

Various statistical methods in adaptive designs are reviewed in Chapter 3, including the concept of the adaptive design, and the benefits and challenges of each statistical method in the adaptive design. Chapter 3 starts with the most commonly used adaptive design – a group sequential design. In a group sequential design, the researcher makes decisions at planned interim analyses, and the test statistics are compared with critical values based on alpha spending function so that the decisions to stop for success or futility, or continue are made on the accumulated data (Pocock, 1977).In addition, this chapter highlights the usage of conditional power or predictive power in interim decision-making such as stopping for futility, stopping for success or sample size re-estimation. Bayesian methods such as the posterior probability and predictive probability are introduced in this chapter, since Bayesian methods integrate prior knowledge and data into the decision-making process.

Chapter 3 also introduces the model based design, which is used to guide the dose selection, sample size re-estimation, and study stopping. There are three types of model based designs, parametric, semi-parametric, and non-parametric models. Parametric models make assumptions on the data distribution and an example of this design is the 4-parametric logistic model, *Emax* model and polynomial regression model. Example of semi-parametric models is the NDLM model. Non-parametric models make no assumptions on the data distribution. Examples of non-parametric models are the spline model etc. A review of the literature that compare *Emax* model and NDLM model highlights no work within a Phase 2a study setting to assess dose response with small sample sizes across a wide dose response range.

Chapter 4 introduces the Bayesian adaptive dose finding method and its application in the RA trial of GSK123456. The chapter starts with the experimental design using a Bayesian dose response model to determine the doses for each cohort. Part A is a cohort randomization design with adaptive dose finding. A Bayesian adaptive *Emax* dose-finding method is used between cohorts to target a dose that provides 90% of maximal benefit (ED90) which guides the dose selection for the subsequent cohort. Part B is a confirmatory group sequential design and expands the number of subjects at a specified dose which is chosen after a review of the results at the end of Part A.

Chapter 5 discusses the proposed design and simulation to support the Phase 2a PoC trial of GSK654321. Similar to GSK123456, the proposed PoC design is also a cohort randomization design in Part A and group sequential design in Part B. There are two potential improvement based on the learnings from PoC trial of GSK123456.

The first learning and improvement is to use a more flexible dose response method – Bayesian NDLM model. Both the Bayesian *Emax* model and NDLM are used to inform the study dose for the subsequent cohorts and to investigate the dose-response relationship. Multiple scenarios with different dose response and data fitting methods are evaluated through simulation.

Chapter 7 discusses the retrospective analysis of DAS28 data at Day 14 and Day 56 and a systematic literature review from RA randomized clinical trials is performed to investigate the relationship of DAS28 at Day 14 and Day 56 in a meta-analysis. A structural correlation between DAS28 change from baseline of Day 14 and Day 15 iss compared with that of observed DAS28 between day 14 and day 56 values. In addition, the validation of Day 14 as surrogate endpoint of Day 56 is discussed.

A delayed response predictive model (DRPM) is proposed. The model is based on the meta-analysis is discussed in Chapter 8 and uses predicted Day 56 response based on Day 14 data in the PoC design to expedite decision-making. Chapter 9 summarizes the whole dissertation and discusses research extensions. It gives design recommendations for the future PoC trial of the follow-on compound GSK654321 as well as discussing their limitations.

# Common study designS in TRIALS of PATIENTS WITH Rheumatoid Arthritis

Statistics plays a fundamental role in clinical trials for the evaluation of the benefits and risks of new treatments or drugs under development (FDA E9 1998) including designing clinical trials, maintaining the integrity of the trials and performing data analysis (ICH E9, 1998). This chapter discusses standard disease activity endpoints in different phases of clinical trials and common study designs used in the drug development for RA.

## The Aims of the Chapter

The chapter highlights the roles of disease progression and its impact on the design and analysis in RA clinical trials. It discusses the background information and explains the need for new medicine to address the patients’ disease burden.

The chapter will aim to:

1. Review common practice in the reporting of disease activity of RA;
2. Discuss the primary composite endpoint of most RA clinical trials;
3. Review the traditional experimental designs of Phase 2a clinical trials;
4. Highlight the most common adaptive design in dose selection in RA trials.

## Reporting Disease Activity in Clinical Trials for Rheumatoid Arthritis

There are general consensuses on reporting disease activities (Felson, 1993; 1995). International societies introduce a core set of results that need to be reported to create consistency in presenting trial results (Felson, 1993) including counts of swollen and tender joints, patient assessment of pain, global assessment of disease activity etc. An evidence-based consensus paper recommended that each trial (Felson, 1993; 1995) should report the following items:

(1) Disease activity score, time to onset, and sustainability of the primary outcome

(2) Descriptive statistics for the primary or secondary endpoints;

(3) Summary of baseline disease activity;

(4) Patients’ percentage of disease activity state or remission;

(5) Patients’ self-reported fatigue.

### Composite Assessment

RA is a complex disease and it is difficult to use a single response to explain the disease activity, so a common practice is to use a composite endpoint (van der Heijde, 1992). Additionally, composite endpoints potentially have higher statistical power than single endpoint alone if the study treatment has similar effects on individual endpoints in the composite (Fransen, 2005). The commonly used composite endpoints in RA clinical trial are DAS28 and ACR which will be described in the next section.

#### Disease Activity Score based on 28 Joint Count (DAS28)

The disease activity score (DAS) was developed in 1990 and simplified in 2003 (Fransen, 2003; 2005). The DAS is a linear combination of 28-joint assessments for tenderness and swelling, ESR the erythrocyte sedimentation rate (or C-reactive protein CRP), and the patients’ global assessment of disease activity. The DAS is a clinical disease index of RA disease activity combining information regarding the number of swollen and tender joints, inflammation biomarker and the patient’s global assessment of disease activity. The DAS provides an absolute value which assesses rheumatoid inflammation (Fransen, 2005).

The DAS is a well validated outcome/endpoint for clinical studies and has been modified as a DAS28 index consisting of 28 tender joint count, 28 swollen joint count, ESR (or CRP) and general health assessment on a visual analogue scale (VAS) (Fransen, 2005). The DAS28 has a continuous scale ranging from 0 to 9.4 with disease activity considered low (DAS28<3.2), moderate (3.2≤DAS28≤5.1) or high (DAS28≥5.1) (Fransen, 2005). DAS28 can be derived using ESR or CRP. A significant correlation between DAS28-ESR and DAS28-CRP was reported, while DAS28-CRP was significantly underestimated compared with DAS28-ESR in a Japanese population (Matsui, 2007). The formula (Fransen, 2005) for calculating the DAS28 score for an individual patient using ESR is

2.1

where tender and swollen joints are a count of the number (up to 28) of joints affected; ESR is the erythrocyte sedimentation rate and patient global assessment of disease (measured on a 0 to 100 scale).

One important advantage of using DAS28 is that it provides a continuous scale, as opposed to the categorical American College of Rheumatology ACR20 which reflects the extent of underlying inflammation and disease activity. Also, since DAS28 provides an absolute value of disease activity, responses to treatment can be compared across clinical trials and trial results can be expressed as clinically relevant outcomes (Fransen, 2005). In addition, DAS28 has been validated as “*a measure of disease activity*” in RA patients undergoing treatment with biological treatments (Fransen, 2005), a similar drug class to GSK compounds GSK123456 and GSK654321.

In this dissertation, DAS28 is the primary endpoint of the Phase 2a PoC clinical trial to evaluate safety and efficacy of GSK123456 and GSK654321.

## Common Designs in Dose Selection of RA trials

In early RA clinical trials, no model-based dose selection was undertaken in the literature being reviewed (Hochberg 2009). There were empirical designs with dose selection algorithms being used, such as dose escalating sequential design and step-up, step-down and saw-tooth designs (Boers, 1997; Grigor, 2004). Step-up and step-down designs have flexible schedules so that drugs can be added or dropped, and dosage increased or decreased (Pincus, 2004). Saw-tooth design is another strategy that starts with a step-down model, and the regimen repeated when disease recurs (Mottonen, 1996; 1998). Through an intensive literature search, no other adaptive designs with application in Phase 2a dose finding in RA patients have been seen in the literature.

In summary, though clinical trials have been increasingly popular for comparative multiple arm clinical trials (Pocock, 1977; 1982; Spiegelhalter, 2004; Grieve, 2005), most of RA clinical trials reported in the literature employ traditional fixed design and there is limited application of adaptive designs in the patients with RA disease. This dissertation will discuss the application of adaptive design in Phase 2a PoC design and potential improvement to expedite the decision making using a GSK compound as an example.

## Summary

International Rheumatoid societies introduced a core set of disease reporting measures to create consistency in the presentation of clinical trial results (Felson, 1993). This core set includes the assessment of disease activity based on joint measurement, for example tender or swollen joints, the patient assessment of pain and disease activity, the physician’s assessment of pain and physical function and the acute-phase reactant levels. The disease assessment uses composite variables including continuous variables (i.e. DAS28) and binary variables (i.e. ACR20).

Appropriate clinical trial design is fundamental to drug development, for evaluating the effectiveness and assessing the benefit-risk ratio. Several trial designs, including sequential monotherapy, fixed combination, step-up, step-down (Boers, 1997) and Sawtooth designs (Mottonen, 1996), have been used to examine the effectiveness and safety of study drugs in RA trials. With the recent development of statistical theory in adaptive design, statistical evidence and decision-making have the potential to be useful in assessing benefit and risk in drug development. The next chapter introduces the concept of adaptive design, including group sequential design and model based designs, and is followed by the study design to support Phase 2a PoC study in Chapters 5 and 6.

# Literature review OF Adaptive designS

As discussed in the previous chapters, RA is a chronic disease affecting 1% of the general population, and the clinical trial designs reported in the literature are mainly limited to the traditional (fixed sample size without interim analysis) design with adaptive designs in studies of patients with RA not commonly undertaken (Boers, 1997; Grigor, 2004; Pincus, 2004). This chapter discusses the various statistical methods which can be applied to adaptive designs and the analysis of Phase 2a clinical trials.

This chapter starts with the definition of what is an adaptive design followed by popular adaptive designs such as Group Sequential Design and model base design etc. Finally, the chapter introduces the basic ideas of model based design where parametric, semi-parametric, or non-parametric models are used to guide the dose selection, followed by existing literature in the comparison of *Emax* model and NDLM model.

## The Aims of the Chapter

The aims of this chapter are to:

1. Discuss the reason and motivation for using adaptive designs in drug development;
2. Introduce the different types of adaptive study designs in the literature (not just limited to RA trials);
3. Outline the statistical methodology used in adaptive designs;
4. With a focus on dose response studies, discuss the *Emax* model, NDLM, and their implementation in clinical trials;

## Motivation for using Adaptive Design in Drug development

Traditionally, most RA studies, especially in the early stages, use fixed arms and fixed sample sizes (Cruyssen, 2005; Fransen, 2005). The compound only moves to the next phase of study after the previous study is completed and the success criteria are met. A decision-making process only occurs after all the data are available in a traditional design, which may delay the development of the drug (Gallo, 2006). At the same time, if the drug is not working or there are safety concerns in the conduct of clinical trials, it potentially poses a risk to patients. Therefore, there is a great need for better designs which are more flexible and have more power for key decision-making via the accumulating data.

There are several challenges facing the pharmaceutical industry including the expiration of drug patents including Research and development spending has doubled but productivity has decreased more than 33% (Woodstock, 2005). Inherent uncertainty in decision-making in drug development creates a risk which results in a decrease in productivity (Woodstock, 2005). To mitigate the risk to drug discovery and development, innovative or adaptive study designs incorporating in-stream decision-making, prospective planning, and engaging clinical operation are needed (Woodcock, 2005).

In 2006, FDA issued the Critical Path Opportunities Listthat calls for the usage of prior experience or accumulated data in the interim decision making of novel trial design (Lionberger, 2008). In 2010, FDA (FDA, 2010) produced draft guidance on adaptive design which provided a clear definition of adaptive design as “*changes in design or analyses guided by examination of the accumulated data at an interim point in the trial, that may make the studies more efficient (e.g., shorter duration, fewer patients), more likely to demonstrate an effect of the drug if one exists, or more informative (e.g., by providing broader dose-response information).”*

The terms adaptive design or flexible design are increasing being used in drug development (Chow, 2008). An adaptive design has the options to modify the trial, such as adding or dropping a dose or a sample size reassessment. This flexibility of design is of great interest to pharmaceutical companies and medical researchers since it provides quality evidence in a timely manner (EMEA, 2006; Chow, 2008).

Adaptive designs have issues as well as benefits (Chow, 2008; Kairalla, 2012). Firstly, the clinical operation, of such a design, can be challenging since additional resources are needed for data management, programming and prospectively planning interim analysis. Secondly, to maintain the study integrity, all the interim decisions must be made using an independent data monitoring committee (IDMC). Thirdly, the adaptive design may inflate the Type I or Type II error rate in certain circumstances. Lastly, adaptive designs need to be planned well; they are not a remedy for bad planning (Chow, 2010). Therefore, only careful logistical consideration of the application of the adaptive design in the clinical trial will drive its success.

### Efficient Design

Efficient study design is a relatively new term (Julious, 2015) and was used in the specification document of the national institute for health research (NIHR) in the UK which encouraged the increase of efficiency through better methods of recruitment, use of adaptive designs, and the exploration of existing cohort and pilot studies (NHS document, 2014).

The term efficient design is thus an umbrella term which incorporates adaptive designs as well as, as the name implies, approaches which make designs more efficient through for example utilising routine data to derive outcomes. An efficient design is defined therefore as one that is statistically robust and facilitates quicker decisions and/or lower cost decisions compared to conventional pragmatic fixed trials and encompass both non-pharmaceutical and pharmaceutical trials.

Efficient trial designs, in the context of adaptive trial designs, can be classified into two main types of designs. The first type is the adaptive design or flexible design, as outlined in Section 3.3, which is the focus of this dissertation. If we classify designs based on the type of adaptation, there are multiple types, including prospective, concurrent, and retrospective adaptive designs (Chow, 2008). If we classify designs based on the phases of clinical trials, they can be classified into the learning phase, seamless phase and confirmatory phase. The group sequential design, the adaptive randomization, enrichment and sample size re-estimation methods can all be implemented in confirmatory trials. The learning phase has adaptive dose response methods and there are multiple types of seamless design, i.e. phase 1/2a and Phase 2b/3 seamless design.

The second type of efficient design is platform trials, in which multiple pharmaceutical companies collaborate to test multiple compounds in the same platform trials, sharing the same control. In addition, there are other trial designs to adapt to non-planned scenario, which is defined as *ad hoc* modifications during the trial. Those unplanned adaptations may include modifications to inclusion/exclusion criteria, treatment administration and duration, or study endpoints (Chow, 2008). In practice these ad *hoc* adaptations are implemented by amendments to the study protocol.

In addition, there are other trial designs that adapt in an unplanned manner, defined as *ad hoc* modifications during the trial. These unplanned adaptations may include modifications to inclusion/exclusion criteria, treatment administration and duration, or study endpoints (Chow, 2008). In practice these *ad hoc* adaptations are usually implemented by amendments to the study protocol and may be at the recommendation of a data monitoring committee.

Figure 3.1 Summary of types of efficient designs for clinical trials. The efficient designs are categorised by trial classification and phases of clinical trial. (Adapted from Kairalla, et al. Trials 2012, 13:145)

## Types of Adaptation in Clinical Trials

Based on the investment sources, clinical trials can be classified as publicly funded clinical trials and privately funded trials. Publically funded trials are those where the direct cost or funds that finalise the trial are from governmental programmes, for example, from the National Institute of Health (NIH), the National Science Foundation (NSF), or other government sponsored programmes in the US, or the National Institute for Health Research (NIHR) or the Medical Research Council (MRC) in the UK. Privately funded clinical trials are those supported and funded by private corporations or associations such as pharmaceutical companies. There are a number of benefits from publicly financed drug trials and one of the most important benefits is the elimination of conflicts of interest (Baker, 2008). However, under the publicly financed system, most funding is awarded to prescription drug trials after the drug is approved in the market. New drug development is usually done in privately funded trials by pharmaceutical or biotech companies (Baker, 2008).

Adaptive trial designs can be classified into three types based on the methods of adaptation: prospective, concurrent, and retrospective adaptive designs (Chow, 2008).

* Prospective adaptations are pre-planned in the study protocol (Gallo, 2006).
* Concurrent adaptations are those with *ad hoc* modifications during the trial. Based on the accumulated trial data or external information, the adaptations may include modifications to inclusion/exclusion criteria, treatment administration and duration, or study endpoints (Chow, 2008). In practice these ad *hoc* adaptations are implemented by amendments to the study protocol.
* Retrospective adaptations are those with modifications prior to database release which usually occurs prior to the database unblinding (Chow, 2008).

In addition, according to FDA guidelines (2011), adaptive clinical trials have the following types:

1) Adopting interim analysis to stop or to adjust patient accrual;

2) Stopping the clinical trial for early success or futility at interim analysis;

3) Changing the hypothesis;

4) Adding or dropping arms or doses adjustment;

5) Modification of allocation rate;

6) Sample size re-estimation.

In the next sections, we will discuss the statistical literature related to the above adaptations as outlined in the FDA guidance, including drop-the-loser design, group sequential design, sample size re-estimation based on conditional power, predictive power, posterior probability, and adaptive dose finding design.

## Drop-the-loser or Pick-the-Winner design

A drop-the-loser or keep-the-winner design has been used in Phase 2 dose finding clinical development due to its flexibility and allowing the inferior treatment arm(s) to drop out of the study (Hills, 2011).

Typically, a drop-the-loser or keep-the-winner design has two stages. In the first stage, the dose responses in patients receiving multiple dose level of study medicine are compared. The treatment arms which have inferior response or have safety issues will be dropped from the study (Chow, 2008). The treatment arm(s) which have the highest responses will proceed to the next stage of study. Additional treatment arm(s) may be allowed to be added to the drop-the-loser or keep-the-winner design if the additional arm(s) is believed to add benefits to the current dose/regimen (Chow, 2008).

## Group Sequential Design

The group sequential design (GSD) allows the study to stop for efficacy or futility, based on interim analyses results. Armitage proposed a repeated significance test based on accumulating data (Armitage, 1969). As there is repeated significance testing in group sequential tests, it increases the chance of getting a false positive result (also called inflation of the Type I error) (Jennison, 1999) of rejecting the null hypothesis (when it is true). To control Type I error, the P-values based on the alpha spending, for accepting or rejecting the hypothesis test, is usually less than the corresponding P-values from a fixed-sample test (Jennison, 1990; 1993; 1999).

The group sequential design was described by Pocock (1977; 1982) and later Lan and DeMets (1983) showed that the group sequential method could be applied by using an alpha spending function. The group sequential test procedures are widely used in clinical research due to their statistical simplicity and flexibility (Pocock 1977, 1982; O’Brien and Fleming 1979; DeMets, 1983).

Pocock (1977, 1982) gave clear guidelines, defining the critical value requirement in the implementation of GSD. Figure 3.2 below (blue dashed curve) is the cumulative alpha spending curve for percentage of time or information (t) in the interim analysis using Pocock’s method. For example, if we assume the total alpha (Type I error rate) is 0.05 and we plan to have 3 interim analyses and 1 final analysis, the cumulative alpha spending is roughly 0.0178, 0.0310, 0.0414, 0.05 respectively for 25%, 50%, 75% and 100% of the information which can be derived from the alpha spending function. Nominal alpha spending is roughly 0.0178 at all analyses. If the trial goes to the final analysis without interim stopping, the final nominal significance level is 0.0178 at the final analysis.

The alpha spending function using Pocock’s method (Pocock, 1983; Pinheiro, 1997) is

, 3.1

where t is the proportion of interim analysis time and α is Type I error rate.

Figure 3.2 Cumulative alpha spending curve for percentage of time or information (t) in the interim analysis using Pocock (blue) and O’Brien and Fleming’s (red) method in group sequential design with alpha fixed at a level of 0.05 or 5%.

O’Brien and Fleming (1979) proposed different stopping boundaries which gave a more conservative approach to early stopping. The same method using nominal significant levels from repeated significant tests could be extended to normal, binary and time to event responses (Jennison, 1999). Figure 3.2 (red line curve) is the cumulative alpha spending curve for percentage of time or information (t) in the interim analysis using O’Brien and Fleming’s method. For example, if we assume the total alpha (Type I error rate) is 0.05, the cumulative alpha spending is 0.000015, 0.003051, 0.019299 and 0.05 respectively for 25%, 50%, 75% and 100% of the information. Nominal alpha spending is roughly 0.000015, 0.003045, 0.018323 and 0.044003 at 25%, 50%, 75% and 100% of the information which can be derived from the alpha spending function.

The formula for the O'Brian Fleming approach that was given is the same as in DeMets (1994) which is

,

where Φ denotes the cumulative standard Normal distribution with mean of 0 and variance of 1, Z1-α/2 denotes the critical value of 1-α/2 quantile of the standard Normal distribution which Φ(z1‑α/2)=1-α/2, α denotes the overall significance level in the hypothesis test. If α = 0.05; Z1-α/2 = Z0.975 and the value is 1.96 from the standard cumulative Normal distribution i.e. 0.975 of the area under the standard Normal distribution curve is to the left of a Z-value of 1.96 and 0.025 of the area is to the right of a Z-value of 1.96.

The alpha-spending function approach was proposed by Lan and DeMets (1983). The type I error was controlled using alpha-spending function approach as a function of time or information (1983).

In medical research, Jennison & Turnbull reported (1999) that group sequential designs will deliver the benefits of early efficacy stopping, early futility stopping, and sample size re-estimation. In traditional designs, there is only one analysis when all the data are collected, statistical Type I error is strictly controlled and the data are reviewed only when the study is completed and the database is frozen. There are certain levels of risks that relate to the traditional design, if the intervention is supra-efficacious or non-efficacious, a trial will not be able to know until all subjects have been recruited at the completion of study, which may mean that time, resources and money are wasted.

Another question of efficient or adaptive design is how to control the risks of wasting time, money and resources in a drug development whilst at the same time maintaining the integrity of the study. GSD is one type of adaptive designs and can be a potential solution, if a new study intervention is promising, we could use a design that can stop early so time and resources are saved (Jennison, 1999). Alternatively, if we are not sure about the performance of the new study intervention, we can use futility assessments, and early stopping for efficacy and futility can be considered as well.

In double blind clinical trial, the interim analyses are led by a data monitoring committee (DMC) with independent statistical centres (EMEA guideline, 2005). A data monitoring committee can make recommendations using ongoing data and recommend study stopping or design modifications based on pre-specified criteria. In Phase 2a RA trials, an internal safety review committee (iSRC) is formed to review the safety and efficacy data and make recommendations on the design modifications. Details will be discussed in the next Chapter.

## Design Modification Based on Conditional and Predictive Power

At the planning stage of a clinical trial, the effect size is rarely known accurately. If the meaningful clinical effect is too small or the variability is too big, statistically significant efficacy cannot be adequately achieved with a small sample size. One hedge against this is to have an interim sample size re-estimation (Shih, 1997; Whitehead, 2001; Li, 2002). Wittes and Brittain (1990) proposed an adjustment by breaking the treatment code or blind estimation and updating the variance in the middle of the trial. Other literature suggests updating treatment difference together with group sequential early stopping (Proschan, 1995). Li, et al. (2002) proposed a modified conditional power calculated using interim results from the first stage to determine the final sample size.

In this section, the motivation and technical details of conditional power and predictive power are discussed.

### Conditional Power and Predictive Power

Traditional power is dependent on the pre-specified treatment effect and uncertainty such as assumptions on the variability of the outcome which are usually based on prior information (Shuster, 1990). The pre-specified treatment effect and variability may not hold in the current study. When interim results are available, we can refine the classical power calculation to as to incorporate the estimation of effects and the uncertainty observed from interim data (Spiegelhalter, 2004).

The formula for conditional power is

. 3.3

Finally, z1-α/2 is the 1-α/2 quantile of the standard Normal distribution (Spiegelhalter, 2004) and here Φ(zα/2)=1-α/2. The formula assumes data Normally distributed and *m* is the total number of subjects at the interim analysis, *n* is the number of subjects remaining to be collected after interim analysis, ym is the observed treatment difference at the interim analysis, σ is the known standard deviation at interim analysis (assumed to be the same at the interim and final analysis), σ is standard deviation accumulating at interim analysis and is also assumed to be the standard deviation for the remainder of trial data to be collected, θ is the assumed treatment difference in the subjects remaining to be collected. Conditional power will depend amongst other factors on the assumed treatment difference in the subjects remaining to be collected. This could be under the null hypothesis: no treatment difference; alternative hypothesis or the current trend in the data.

Bayesian power is the probability of detecting a treatment difference assuming a true effect, θ, where the prior information on this parameter is incorporated into the model (Spiegelhalter, 2004). It is also the predictive probability of obtaining a significant Bayesian result when testing the null hypothesis against an alternative. A prior distribution of θ, p(θ), is needed before one can obtain a posterior distribution for θ conditional upon the interim observed data. If we know nothing about the parameter of interest, θ, the true treatment effect, we can use a non-informative prior or a very vague prior for θ, for example a flat curve or a Normal distribution with a very large variance.

Predictive power can be seen as a possible extension of conditional power. Predictive power involves averaging conditional power values over various treatment effect θ (Spiegelhalter, 2004). When a non-informative prior, for treatment effect θ, is utilized the formula for predictive power (Spiegelhalter, 2004) is

. 3.4

Other predictive power formulas with informative prior can be found in Spiegelhalter (2004) or also be obtained using simulation.

## Drug Response Model in Drug Development

Dose response is defined as, “*knowledge of the relationships among dose, drug concentration in blood, and clinical response (effectiveness and undesirable effects.)*” (ICH E4, 1994). The dose response relationship in the target population should be planned carefully at each stage of the development process. In support of this, International Conference of Harmonization (ICH) E4 guidance states that, ‘*the entire database should be examined intensively for possible dose-response effects’* and further notes that such meta- or multivariate analyses should be conducted, ‘*even if the analyses can yield principally hypotheses, not definitive conclusions*’ (ICH E4, 1994).

The ICH-E4 guidance (1994) notes that ‘*it is prudent to carry out dose-ranging or concentration-response studies early in development as well as in later stages in order to avoid failed phase 3 studies or accumulation of a database that consists largely of exposures at ineffective or excessive doses’*. As such, it is of great interest to characterize dose-response as early as possible in drug development. Knowledge of the relationship is important in determining the optimal dose for late phase study and providing effective and safe use of the drugs in individuals (FDA Guidance, 1994).

There are various dose response models available in the Phase 2a phase of clinical evaluation of novel pharmaceutical products due to the study population, research objectives, and the nature of questions arising at each phase of clinical trials. The dose response models in dose response analysis include multiple comparisons, trend test, linear contrast test, model based modelling, and non-parametric models, which are described in below sections respectively.

### Multiple Comparison Methods in Dose Response Analysis

Historically, statistical analyses are performed on various dose levels in clinical and pre-clinical experiments (Sheskin, 2007). Subjects are assigned to either treatment groups or a control group. Different treatment groups receive incremental levels of the test drug. The control group is given a volume of placebo or a positive control. The null hypothesis of no drug effect is tested first against any possible drug effect between various dose groups and the placebo. In statistical terms, the null and alternative hypotheses are:

H0: No difference among various dose levels

H1: There is a difference between various dose levels.

Multiple statistical tests are currently used to test the null against the alternative hypothesis (Sheskin, 2007). Once H0 is rejected, a Type I error rate of alpha, pairwise comparisons (also called post-hoc tests) between each of the dose levels and the control are performed to assess the potential effect of one or more dose levels compared to the control subjects (Lehmann, 2005).

#### Multiplicity Adjustments

Multiple test procedures using closed tests which control the family wise error rate (FWE) have been discussed in the literature (Hsu, 1999; Marcus, 1976). The multiple testing procedures are to control the overall type I error probability which usually requires a closed family of hypotheses (Hsu, 1999; Dmitrienko, 2003 and 2006).

In additional, the P-values used to assess statistical significance can be adjusted with a multiplicity adjustment; popular adjustments include: Bonferroni adjustment (Abdi, 2007); Dunnett’s (1955) method, and Williams’ adjustment by permutation procedure (Westfall, 2000)*.*

### Model-based Dose Finding

The dose response models are used to estimate MED, ED90, and/or MTD which are usually identified in early or middle phase clinical development using an adaptive dose finding design including dose escalation or searching for optimal dose (Bornkamp, 2007).  The dose level and duration of the study treatment of next phase clinical trial may be determined by the model.

We will cover several popular model based designs in this section. The NDLM model makes no assumptions about the dose response curve, and fits a normal linear regression ateach dose, allowing changes in the regression parametersbetween neighbouring doses. Other model based designs use linear, log-linear, 4-parameter sigmoidal, *Emax*, quadratic, exponential models etc.

### Linear and polynomial regression model

Statistical analysis in dose response normally starts with linear or nonlinear regression of a response to a dose (Ting, 2006). In its simplest form, a dose-response model can be assumed to be a polynomial regression model (Kutner, 2005). A simple linear regression model is a special case of a polynomial regression model.

Polynomial regression is a form of linear regression and the relationship between the independent variable x and the dependent variable y is modelled as nth degree polynomial, the model

, 3.5

is used, where ai (i=0 to p) are regression coefficients and ε is random error with mean zero. Further discussion of polynomial regression method will be provided in Chapter 6.

### Parametric Emax Model

The 3-parameter logistic model, 4-parameter logistic model, and *Emax* models are examples of nonlinear regression models, which have been used in bioassay validation and nonlinear dose response (Ting, 2006).

*The Emax* model is one of the key nonlinear regression models which may reflect the underlying biological response (Ting, 2006). In its 4-parameter form this model is usually written as:

3.6

where response is the response of interest, E0 is the basal effect or placebo response when Dose is equal to 0, *Emax* is the maximum achievable effect a drug can have over the placebo response, ED50 is the dose which produces 50% of the effect, and S is the slope.

It is not uncommon to fix the slope as 1 in the analysis of *Emax* model. Ting (2006) discussed the *Emax* model in (3.2) as a particular case of *Emax* model in dose response modelling which is referred to as the hyperbolic *Emax* model. The *Emax* model with the slope factor (slope is not fixed at 1) is referred as sigmoidal *Emax* model. The steepness of dose response relationship is determined by the slope parameter in *Emax* model. Although both models can be justified biologically in the relationship of drug receptor, the use of hyperbolic model is primarily for empirical reasons. Both a hyperbolic and sigmoidal model can be used to fit dose response, the choice of hyperbolic is primarily for empirical reason. The study with hyperbolic *Emax* model (Slope =1) was designed prior to my joining the study team and was the protocol specified analysis.

In the model fitting of proposed Bayesian *Emax* model with the slope parameter (sigmoidal Emax mode) from simulated data, computationally errors often occur in WinBUGS, regardless of the choices of non-informative or informative prior (for example the slope follows a uniform distribution between -2 and 2 or Normal distribution with mean of -1 and standard deviation of 0.5). While with the slope parameter fixed at 1, there were little computational errors. The bias of the model parameters ED90 were also small (the estimated ED90 is within 20% of true ED90). Therefore, an *Emax* model with slope parameter fixed at 1 is used in the model and analysis of PoC study of GSK123456 as well GSK654321. As a result, the previous result (3.6) is converted to a 3-parameter *Emax* model if we set S to 1, then the model becomes

3.7

The dose response data can be fitted using an *Emax* model with a Gauss-Newton iterative algorithm for non-linear least square (Ting, 2006). Alternatively, the parameters of interest for the *Emax* model can be estimated by maximum likelihood estimation (MLE) and Bayesian methods. We chose the latter approach as Bayesian statistics (Carlin and Louis, 2008) integrate information into the computation of the posterior probability of parameters, using the accumulated data observed so far for later doses and prior information for the early doses. In addition, the parameters from the Bayesian method are displayed as distributional profile - which can be useful to illustrate uncertainty - and offer a robust estimation of parameters in complicated model (Carlin and Louis, 2008). The *Emax* model will be further discussed in Chapter 4 and Chapter 6.

### Contrast Test

Unlike multiple comparisons with a pairwise comparison between treatments, this section describes a linear contrast method in constructing a linear combination of parameters whose coefficients add up to zero, allowing comparison made to dose response setting (Bretz, 2005). Linear contrast methodology is relatively simple to use and the hypothesis testing can be constructed accordingly. If multiple hypothesises of linear contrast are constructed concurrently, the family-wise error rate of combining modelling procedures and multiple comparisons in dose ranging studies should be controlled (Bretz, 2005).

#### Emax model using Linear Contrast Method

Thomas (2006) extended the linear contrast to the sigmoid *Emax* model and designed a set of linear contrast to describe difference shapes of *Emax* dose-response curve. *Emax* model is a flexible model, with varying the model parameters E0, Emax, ED50 or slope or Hill parameters, to approximate many families of parametric dose response models with monotonic response (Thomas 2006), such as the linear, log-linear, or exponential model. The linear contrasts are then generated from the dose response of different curves of possible *Emax* models, example *Emax* dose curves are displayed in Figure 3.3. The six dose response profiles in the dose response plots are then described by 6 linear contrasts which are pre-specified in the analysis. A statistical test is constructed on the smallest P-value from a set of contrasts. Bayesian methods are recommended and weighting of the contrasts can be implied by prior distribution (Thomas, 2006).

Figure 3.3 Example dose response curves for building linear contrasts.

#### Trend Test in Dose Response Analysis

If an assumption is made that an increase in dose will result in an increase in a response variable, possible methodologies in dose response analysis are trend-based or monotonic response analyses. A trend test can be considered as a special case of linear contract method, where the weight is chosen so the trend test is more powerful to detect associations when the assumed trend is true. It is often expected that response increases or decreases monotonically with dose, but not necessarily linear.

For different types of response variables and depending on whether the data are Normally distributed or not, the following statistical tests can be used:

* Cochran-Armitage trend test (Agresti, 2002) (nonparametric for categorical or binary data),
* Jonckheere's statistic (Williams, 1971) (nonparametric for continuous or count data),
* Williams' test (Williams, 1972) (parametric for normally distributed data),
* Shirley's test (Shirley, 1977) (nonparametric for continuous or count data).

A statistically significant result for any of the above tests shows that there is a trend for changes in response as dose changes.

### MCP-MOD Method

The MCP-Mod method is a hybrid approach that combines multiple comparison techniques with hypothesis testing and parametric modelling or model averaging method to find suitable dose(s) for confirmatory Phase 3 trials (Bretz, 2005; Pinheiro, 2007). In 2005, Bretz et al discussed a dose response testing method based on several non-nested linear contrasts of the dose levels from a selection of broad candidate dose response models of which are used to test the null hypothesis. Appropriate contrasts are then built for each candidate model and is tested employing multiple comparison techniques to preserve the familywise type I error. The model is significant when at least one of the tests is significant and the “best” dose is chosen from the model after fitting the data (Bretz, 2005; Pinheiro, 2007). Since the optimal dose is chosen out of broad candidate models, the MCP-MOD method reduces the chance of model misspecification and model uncertainty at design stage.

### Semi-Parametric Normal Dynamic Linear Model

A NDLM can be used to fit to estimate the dose-response relationship. The NDLM model or state space model is a very general class of non-stationary time series models which include a term to model trends, seasonality, covariates and autoregressive components (Petris, 2009). It is a flexible model and does not assume monotonic dose response. A description of the NDLM used in the analysis is shown below (Petric 2009),

jk = θj + εjk, , where j =0,...,J, k=1,...,K, εjk ∼ i.i.d. N(0, σ2 ), 3.7

θj = θj−1 + δj−1 + ωj where ωj ∼ N(0, σθ2 ) ,

δj = δj−1 + ǫj where ǫj ∼ N(0, σδ2 ).

The likelihood of Yjk follows a Normal distribution with mean of θj for each j dose θ (Dosej) and k subject and with variance of σθ2. The Dosej is assumed to be spaced equally. θj is the estimated treatment effect at Dosej. Furthermore, θj has a linear relationship with neighbouring θj−1 with slope of δj−1. ωj is the smooth coefficient which allows smooth changes. θ1 is the untreated or placebo response when the drug dose is equal to 0 and both θ1 andδ1 follow Normal distributions. The coefficients for the NDLM model can be estimated from maximum likelihood methods (Newman, 2014) and Bayesian methods – we used the Bayesian NDLM method because Bayesian methods offer robust estimation of parameters with complicated models and provides better model fitting in both monotonic and non-monotonic dose response (Carlin, 2008). The prior distributions on σ2, σθ2, and σδ2 are inverse-gamma distributions. Further detail of NDLM model will be discussed in Chapter 6. The basic idea of NDLM model is to fit response of adjacent dose θj as a straight line with neighbouring θj−1, as intercept and δj-1 as slope (Weir 2007). Therefore, any predication of response between θj and θj will be linearly extrapolated between doses.

NDLM model is semi-parametric model so unlike a parametric model, such as the *Emax* dose –repose model which assumes a monotonic dose response relationship, the parameters determine the location and shape of the dose response curve. The parameter used in the semi-parametric NDLM model are the treatment effect at each dose θj where j =1,...,J; the slope of neighbouring doses δj−1 where j =2,...,J; the evolution noise σ2, σθ2, and σδ2. So there are minimum 2J+3 parameters for J dose levels.

Bayesian NDLM model was applied in an Acute Stroke Therapy by Inhibition of Neutrophils (ASTIN) trial by Krams and Grieve (2003). The Phase 2b ASTIN trials allocated patients to 1 of 15 doses, or a placebo, adaptively based on the response and allowed for early terminationfor efficacy or futility based on posterior probability using the Bayesian NDLM model.  The model used the Markov chain Monte Carlo to derive joint posterior distribution of parameters which informed the estimation of the ED90.

## Existing Literature in Comparison of Emax and NDLM model

To investigate the use of *Emax* and NDLM models a systematic literature search was conducted. The databases Google scholar, PubMED line and web of science (WoS) were searched.

* PubMedline comprises citations for biomedical literature life science journals, and online books. <http://www.ncbi.nlm.nih.gov/pubmed> [Date last accessed 10th July 2017].
* Google Scholar is a web search engine that indexes the full text or metadata of scholarly literature  <https://scholar.google.com/> [Date last accessed 10th July 2017].
* Web of Science (WoS previously known as Web of Knowledge) is an online subscription-based scientific citation indexing service maintained that provides a comprehensive citation search. [Date last accessed 25th May 2016]

The focus of PubMed and WoS databases is peer reviewed publications. While Google scholar also searches the grey literature/documents as well peer-reviewed and non-peer-reviewed articles.

A total of 2548, 95000 and 8400 return citations were found by searching “Bayesian *Emax*” model and the number of returned citations were 5, 87, and 13 repetitively for “Bayesian normal dynamic linear model” alone. However, there were 14 records found from Google Scholar and none from PubMed and Web of Science when searching for both *Emax* and NDLM model combined.

Of the total of 14 returned records that had both *Emax* and NDLM models, there were two papers that were relevant. The first one is the comparisons of NDLM model and other models such as ANOVA, D-optimal to analyze adaptive dose ranging trials (Bornkamp, 2007). This paper had no direct comparison of the Bayesian *Emax* model and the NDLM model.

The other paper was a poster (Temple 2011) and the author was contacted for their unpublished PhD dissertation on which the work was based (Temple, 2012). In the dissertation, the comparisons of NDLM model and *Emax* was discussed in adaptive and cohort randomized trial assuming dose response follow linear, difference shapes of *Emax* and Umbrella curve (a class of U-shape dose response where high dose response is a good response). It was shown through simulation that NDLM method has better power in correctly identifying a dose in the target dose interval than Bayesian *Emax* model. The focus of Temple’s research was to investigate models through simulation to select the best dose to carry from Phase 2b to Phase 3, with a limited dose range (0-8 mg) and a comparatively large sample size (n=250-280 in total and ~30 each arm). The context of this dissertation in this paper is dose selection in first time in patients study i.e. Phase 2a, with a much wider dose range (0-30 mg/kg), and a small sample size (n=64 and 6-8 each arm except placebo). Our work also had a specific interest in an investigating study where there is pronounced U-shaped curve as this had previously occurred in a Phase 2a trial and the anticipation was the dose response may be similar in the study being planned (Choy, 2013). Therefore, although the research of Temple was of interest the work could not be generalised to the study design in this dissertation. Thus, there is no comparison of methods in the literature within a Phase 2a study setting to assess dose response with small sample sizes across a wide dose response range. Following on from this there is thus no comparison of methods to assess dose response in patients with rheumatoid arthritis (RA), which is the main interest of this dissertation (and the specific therapeutic area).

The aim of proceeding chapter will be firstly to compare the two main statistical models (*Emax* and NDLM) for estimating a dose response relationship, between a drug and outcome, in a Phase 2a trial in patients with rheumatoid arthritis (RA). Also to use extensive simulations show how the two models perform under a fixed and adaptive design under a variety of assumed dose-response profiles with a focus on U-shaped response curve, a plausible dose response curve in GSK654321, being developed for treatment of RA. The *Emax* model and NDLM model as well as the comparison of *Emax* and NDLM model will be further discussed in Chapter 6.

## Design based Dose Response Model

Many dose response models are available in describing the study drug and drug response relationship in the drug development. The popular models are linear model, 4-parameter sigmoidal, *Emax*, log-linear, quadratic and exponential models (Bretz, 2005).

Due to the nature of biopharmaceutical receptor response, the *Emax* model, a parametric model, is considered a better model, suitable for larger molecular drugs and the model parameter has biological and clinical meanings. Dragalin et. Al. (2007) applied a D-optimal design in selecting the ED90 of the *Emax* model to a dose range study. There are also other methods like the Bayesian adaptive *Emax* model covered in the literature (Leonov, 2009), in which the number of patients enrolled into the clinical trial is based on the dose response and centred at the point where the patients have the optimal response.

## Statistical Issues in Implementing the Adaptive Design

Most interim analyses use unblinded data in the adaptive design. If not handled appropriately, unblinded interim data and design modifications may introduce bias into the interpretation of the final trial results (FDA Guidance, 2010). For example, there can be issues related to sample size re-estimation (FDA Guidance, 2010). The Type I error can be inflated if the sample size re-estimation is not handled properly. It was advised by FDA that sample size modification, should be pre-planned and performed by an independent party to avoid the conflict of interests with the drug developer (FDA guidance, 2010). A well planned design is the key to answer the research question of interest and success of the adaptive design.

## Operational Issues in Implementing the Adaptive Design

There are several challenges in the study operation and process in successfully implementing an adaptive design including the randomization system, maintain the study integrity during design changes, the drug supply and rapid access to primary or key analysis data (Pong, 2001; Gallo, 2006). FDA draft guidance on adaptive designs (2011) suggests operational bias or risk resulting from study conduct can affect the validity of statistician conclusions in an adequate and well controlled (A&WC) trial. Maintaining blinded data access by shielding the investigators from knowledge of decision in adaptive is important since the decision may impact the investigators to select, manage and care the patients. To minimize the operational bias, it is suggested that an independent data monitoring committee (IDMC) is used to implement the adaptive design decision.

## Summary

This chapter has reviewed the statistical methodology on adaptive trial designs and applications that are related to this dissertation. Adaptive design methodologies are of great interest to researchers due to its ability to aid decision-making without compromising the clinical trials’ integrity at interim analysis, if handled properly, thus enabling faster decisions of stopping a failed drug, and speeding the development of effective drugs to market.

The PoC study to investigate the efficacy of GSK123456 in RA patients (Chapter 4), is a two-stage seamless design which combines the dose finding phase with an 8-cohort design and confirming phase with group sequential design, using the O’Brian Fleming (OBF) boundary. The Bayesian *Emax* model is applied to fit the dose response and allocate the patients to the dose which is nearest to the ED90, the dose to achieve 90% of maximal response at next cohort. In Chapter 6, a more flexible dose response model, the Bayesian NDLM model, is proposed to re-fit the data retrospectively and to provide a path for the development of GSK654321, the follow-on compound to GSK123456.

In summary, the statistical methodologies and techniques for selected adaptive designs, such as the drop-the-loser design, the group sequential design and the dose response model, that may be used in the design and analysis of RA clinical trials, have been reviewed. The decision criteria are usually pre-defined in the study document prior to the study. The decision criteria are determined by the operating characteristics, recruitment, primary analysis, safety analysis, clinical significant difference, and the risk to investment.

The next chapters will discuss the Phase 2a RA study as part of the clinical development of the GSK compounds using the Bayesian dose response model for futility and success stopping. Then, the proposed work, which will apply statistical methodology in a Phase 2a clinical drug trial for RA patients, will be discussed. Multiple design options to support compound progression will be discussed in Chapter 6.

# ADaptive Design in Proof of concept study of GSK123456– a case study

In clinical development Phase 2a clinical trials are undertaken both for finding the right drug dose and demonstrating the efficacy of the compound using a surrogate endpoint such as DAS28. The Phase 2a clinical trial is one of the key studies that provide data for future investment decisions in the drug development for a pharmaceutical company. It is also a major decision point to determine whether the compound is efficacious in patients with the disease by concept, the benefit and risk ratio of the compound, whether the pharmaceutical company would like to pursue a Phase 2b and Phase 3 development, which normally requires a large investment by the pharmaceutical company (Milligan, 2013).

In Chapter 2 it was highlighted how the application of adaptive design in RA clinical trials is limited. In the treatment of RA disease, most study designs have a traditional fixed parallel study design (Kimpel, 2007; Nishimoto, 2007; Edwards, 2004) with long term study endpoints. This may lead long development time in multiple cohort dose response study especially with biological treatment that has a long half-life. In addition, in Chapter 3, it was also highlighted that potential benefit using adaptive design in evaluating dose response efficiently. In a Phase 2a dose finding study, a wide range of doses are explored and the optimal doses are determined. A dose higher than the optimal dose could mean side effects may be unduly high. If the dose is too low, suboptimal efficacy may be observed. Therefore, there is increasing interest in clinical trials using adaptive design features, which may make the studies more efficient (Chevret, 2006). As discuss in earlier Chapters, the adaptive designs have potential to include stopping for futility or for efficacy and sample size re-estimation. By comparison to a traditional design, an adaptive design has the potential to more likely to allocate more patients to efficacious doses and minimize the exposure of patients to ineffective doses and this increase the chances that a study could demonstrate the effectiveness of the drug (Chevret, 2006).

Similar designs have been used in clinical trials to determine the optimal dose, the range of dose response, the maximum tolerated dose, and the minimal effective dose (Haines, 2003), for example, a Bayesian optimal design has been used in oncology Phase 1 clinical trials to estimate the maximum tolerated dose (Haines, 2003) with Bayesian optimal design use for a pilot study to allocate patients in the latter part of the study design (Haines, 2003).

Two-stage designs were undertaken in non-RA Phase 2a or Phase 2b clinical trials, especially in oncology clinical trials (Simon, 1989). In a two-stage design, the sample size at the second stage is dependent on the results of the first stage. However, the application of such design is relatively rare in RA disease areas. Therefore, in this chapter we will introduce a two-part adaptive design and apply the design into a Phase 2a RA clinical trial.

## The Aim of the Chapter

The aim of the chapter is to review PoC study design and apply it to a RA clinical trial using GSK123456.

## Study Design

### Design Considerations

There are a number of factors that impact on the design of this study. Firstly, the compound had just completed a Phase 1 study and was in an early phase of development. Limited information was available to inform the efficacy of the study drug. Secondly, the study was relatively small due to the limited budget available. It is estimated that a minimum sample size of 270 would be needed for a traditional fixed parallel design with 8 proposed study doses (0.03, 0.1, 0.3, 1, 3, 10, 20 and 30 mg/kg) and placebo, to achieve 90% statistical power to achieve a target decrease of treatment effect of -0.95 in DAS28 change from baseline with pairwise comparisons without multiplicity adjustment in Part A alone. Secondly, it is a first time the compound was assessed in RA patient population and so there was a wish within the team to monitor safety. The plan was to start the doses at low dose (0.03 mg/kg), followed by careful dose escalation to the highest dose (30 mg/kg).

A traditional fixed group design with multiple arms is not a viable option for the Phase 2a study as there is not adequate statistical power to assess all the possible doses in the dose ranges (a total of 10 planned dose levels). A novel study design, combining both “learn” and “confirm” and providing an efficient design in exploring the dose response and investigating the target drug effect, is needed to address the challenges, which will be discussed in next section.

### Study Design

The aim of this section is to review study design of the PoC study and to investigate whether GSK123456 is efficacious and safe/tolerated in RA patients, which will help to design the next study for follow-on compound GSK654321.

The study was a two-part (A and B), multicentre study to investigate the safety, tolerability, dose response, efficacy and pharmacodynamics of intravenous (IV) GSK123456 in patients with RA. Part C is a bioequivalence study, and was added after the completion of Part A and will not be covered in this dissertation.

A Bayesian *Emax* dose finding design (Part A) and group sequential design (Part B) were proposed to accommodate the “learn” and “confirm” scheme of the study. Part A had a Bayesian dose finding design to learn the dose response and find the optimal dose, and Part B used the optimal dose from Part A to confirm the efficacy using a group sequential design. Figure 4.1 is a graphical schematic of the study design (Part A and Part B) originally proposed by the GlaxoSmithKline study team. I was the statistician leading all statistical activities on the Phase 2a PoC study to apply the Bayesian *Emax* model to the two-part Phase 2a PoC clinical trial in patients with active disease RA including between cohort dose escalation, interim analysis and final analysis in this dissertation (Choy, 2013).



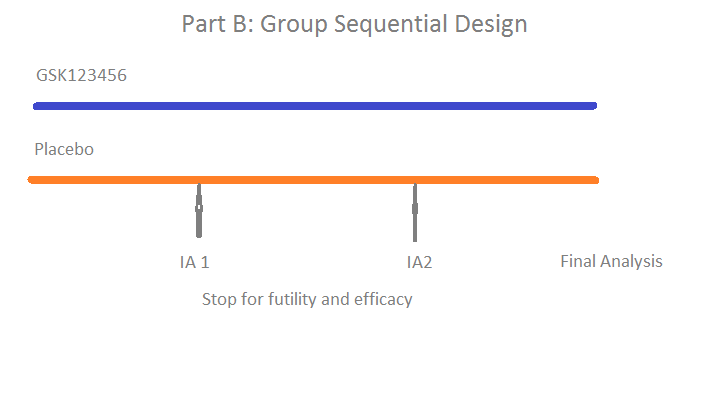


Figure 4.1. A graphical schematic of the study design (Part A and Part B). Part A is learning phase and has 8 cohorts (6 patients received GSK123456 and 2 received placebo, the ratio of GSK123456: placebo is 3:1) whose doses would be decided by the *Emax* dose response model. Part B is confirmation phase and is a group sequential design the dose for which would be selected based on Part A (Choy, 2013).

Part A is an adaptive dose finding phase. Each cohort consists of 8 subjects (6 are randomized to receive active doses and 2 to receive placebo). An analysis with a Bayesian *Emax* is undertaken between the cohorts with the aim to target dose that provides 90% of maximal benefit (ED90) across all doses and thus guide the dose for the next cohort – subject to no additional safety concerns.

Part B is a confirmatory repeat dose phase using a group sequential design. Part B expands the number of subjects at a specified dose which is chosen after the review of the results at the end of Part A. Subjects are randomized on a 2:1 basis to the selected dose of GSK123456 or placebo. There was two interim analysis planned at ~33% (18 patients) and ~66% (36 patients) of patients respectively completing the Day 56 assessment in DAS28.

The sample size for Part B is estimated based on the estimated treatment difference and variability observed between the optimal dose and placebo in Part A. The optimal dose is defined as the dose which is the ED90: which is the predicted lowest dose level based from the Bayesian model that would achieve 90% of maximal effect at each dose escalation review. The doses were half-log spaced at design stage. The sample size is also capped with maximal sample size of 54.

Part B will expand the number of subjects at a specified dose which will be chosen after the review of the results at the end of Part A. Subjects will be randomized on a 2:1 basis to the selected dose of GSK654321 or placebo. A minimum of 18 subjects will be randomized in Part B in order to assess the safety and tolerability of repeat intravenous doses. The maximum sample size needed in the study is based on sample size calculation assuming variability of Part A interim analysis (standard deviation of 1.1). Overall; a maximum of 54 subjects will be enrolled in Part B. The design diagram is presented is Figure 4.1 (bottom plot).

At the interim analysis, the nominal significance level for efficacy given by O’Brien-Fleming error-spending function was used for decision-making. For example, if the observed P-value is less than or equal to the nominal significance level for efficacy and the difference is positive, we can conclude that the GSK123456 dose is superior to placebo. If the observed P-value is greater than the nominal significance level for efficacy we conclude that the GSK123456 dose is neither superior nor harmful compared to placebo nor we do not have enough information to make a decision and must continue to evaluate that treatment arm. A repeated measure analysis was used to analyse the primary endpoint, mean changes from baseline in DAS28 scores at 56 days. A mixed effect model was used, including fixed effects for treatment, visit as a categorical variable, treatment by visit, baseline. Subject was fitted as a random effect. The sample size of Part B is re-assessed based on the results observed from Part A and the sample size is 56.

This study was run at multiple sites in multiple countries. A central randomization schedule was generated and investigators, nurses, and patients remained blinded throughout the study. The bulk drug supply was sent to each investigator’s site and site pharmacists diluted the drug to the correct dose level for each patient.

The follow-on compound GSK654321, which is the main focus of this dissertation, will use the similar two-part study design.

### Endpoints and Hypothesis

The primary efficacy endpoint was the clinical endpoint of the Disease Activity Score based on 28 joints (DAS28) at Day 56 (Month 2). DAS28 was measured at multiple time points during the study treatment and follow-up phase, i.e. screening and pre-dose (for study eligibility criteria), Day 14, Day 28, Day 56, Day 84 and follow-up visits, additional measurements may be taken if deemed necessary.

The null hypothesis (H0) and alternative (H1) hypotheses for this endpoint in Part A are

H0: There is no relationship between study doses and weighted mean DAS28 changes.

H1: There is a relationship between study doses and weighted mean DAS28 changes.

The null hypothesis and alternative (H1) hypotheses for this endpoint in Part B are

H0: There is no difference between GSK123456 and placebo in DAS28 Day 56 changes

H1: There is a difference between GSK123456 and placebo in DAS28 Day 56 changes

### Adaptive Dose Escalation Strategies

The study design is a within cohort randomized design which all patients in each cohort except placebo patients receive the same dose level. A Bayesian dose response analysis was performed by an unblinded independent GSK analyst and a dose recommendation was provided to the study team by an independent safety review committee (iSRC) based on estimation of efficacious ED90. The iSRC reviewed all blinded efficacy as well as safety data from participating patients.

All doses are nominal doses. The ED90 is defined as the dose to achieve 90% of maximum DAS28 response with the lower dose chosen if there are multiple values. In this calculation the maximum response is estimated from the maximal DAS28 effect at all doses. The 90% (ED90) of maximum response is then calculated as the lowest nominal dose in which is closest to the estimated dose that achieves 90% of maximal efficacy. A study limitation was that the maximal dose for a given cohort could not be more than 10-times (two dose levels) the dose in the previous cohort (for safety reasons). The maximum dose level is 30 mg/kg. If the posterior ED90 exceeded 30mg/kg (Level 10), the maximum planned dose of 30mg/kg would be used.

### Bayesian *Emax* model

The *Emax* model is a widely applied model relating drug concentrations to effects (Pinheiro 2006) and was planned for the analysis of dose response*.*

The *Emax* model is written as

4.1,

where is the change in DAS28 score from baseline to day 56 post-randomisation, E0 is the untreated or placebo response when the drug dose is equal to 0, *Emax* is the maximum achievable increase response (or decrease as appropriate) over placebo, ED50 is the dose which produces 50% of the effect, and S is the slope of the dose response curve.

The parameters of interest for the *Emax* model can be estimated by maximum likelihood estimation (MLE) and Bayesian methods – in the dissertation the latter method was chosen as Bayesian methods integrate information into the computation of the posterior probability of parameters, using the accumulated data observed so far for later doses and prior information for the early doses and offers robust estimation of parameters in complicated model (Carlin, 2008). A further advantage is the parameters from Bayesian method are displayed as distributional profile - which can be useful to illustrate uncertainty to a study team.

The estimation of ED90 from Bayesian *Emax* model is of interest to the dose escalation therefore, result 4.1 was re-parameterized using ED90 in the place of ED50 (ED90=9xED50) when the slope is 1, the ratio of ED90 and ED50 is 9 (Ting 2006) when the slope is 1.

Where ED90 is defined as the dose to achieve 90% of maximum DAS28 response. In this calculation the maximum response is estimated from the maximal DAS28 effect at all doses. The 90% (ED90) of maximum response is then calculated as the lowest nominal dose in which is closest to the estimated dose that achieves 90% of maximal efficacy.

A factor of 1.1513 maps the dose level (dose level 1-10) to dose in mg/kg on the scale of interest (0.001–30 mg/kg). The slope of the *Emax* model has been fixed at 1 so the slope is removed. The planned clinical doses are 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, and 30, which are the nominal doses with a rate of half log increase. Therefore, a factor of “*exp*(1.1513)” converts clinical doses to dose levels with 1 to 1 mapping as illustrated in below table.

|  |  |
| --- | --- |
| β3: Dose level in *Eqn* 4.x | Clinical doses= *Exp*(1.1513\*( β3-7) |
| 1 | 0.001 |
| 2 | 0.003 |
| 3 | 0.010 |
| 4 | 0.030 |
| 5 | 0.100 |
| 6 | 0.300 |
| 7 | 1.000 |
| 8 | 3.000 |
| 9 | 10.000 |
| 10 | 30.000 |

The model parameter β3 covers a broader range of doses, it also prevents the clinical dose from being negative by taking the exponential transformation. The estimate of β3 is then approximated at the lowest nominal dose level that is close to ED90. If the estimate of β3 is outside range, i.e. the maximal estimate of β3 in the dose range is used i.e. the patients will be assigned to maximal dose at 30 mg/kg. The Bayesian dose response model for the GSK123456 PoC study was

,

where DAS28(0) denotes baseline DAS28, *β*2 maximal DAS28, *β*3 the dose level that produces 90% of the maximal change in DAS28, and *ε* denotes an error term that follows a Normal distribution. The posterior predicted ED90 (β3) is selected as a target dose for the next cohort.

Table 4.1 Bayesian priors used for updating information in dose-finding simulations. The slope of the *Emax* model has been fixed at 1. All the priors are non-informative priors with continuous endpoints to follow Normal distribution of large variance.

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Interpretation** | **Prior distribution** |
| β0 | Intercept | N(0, 1E6) |
| β1 | Baseline of DAS28 | N(0, 1E6) |
| β2 | *Emax* maximal effect | N(0.3, 100) |
| β3 | ED90, dose level generating 90% of maximal response | N(8, 1E6) |

The factor 1.1513 maps the dose level (dose level 1-11) to dose in mg/kg on the scale of interest (0.003–30 mg/kg). The slope of the *Emax* model has been fixed at 1 so the slope is removed from the initial modelling due to the computational difficulty in the model fitting. The priors of model parameters *β*0 and *β1* follow a Normal distribution with large variance i.e. *N*(0,1E6) and the prior distribution of *β*2 and *β3* are N(0.3,100) and N(8, 1E6) respectively.

Markov Chain Monte Carlo (MCMC) with Gibbs sampling was used to simulate the posteriors distribution from the prior distribution and likelihood. 1000 samples were used to estimate the model parameters after burn-in of 500 samples. The estimation of mode of posterior predicted ED90 (β3) is selected as the target dose for the next cohort. The SAS software (SAS version 9.2) was used to call WinBUGS (Spiegelhalter, 2000) using WinBUGSio macro (Smith 2007) during the Bayesian analysis of the clinical trial.

In part B, a repeated measures analysis was used to analyse the primary endpoint DAS28 change from baseline at Day 56. A mixed effect model was used in the repeat measure analysis, including treatment, time point, and treatment by time point interaction as fixed effects and subject as a random effect. If the P-value for active treatment and placebo comparison was less than O’Brian Fleming (OBF) normal significance level (e.g. 0.049 at final analysis), it would be taken that the study has observed a statistically significant difference. The mathematical model of the repeated measure analysis is described below in (4.3),

*yijt = μ + Ti + mjt + Ti mjt +eit* 4.3

where *yit* = DAS28 for *i*-th dose j-th subject at *t*-th time; *μ* = general mean; *mjt* = effect of *t*-th time with j-th subject; *Ti* = effect of *i-th treatment and eijt* = residual. Residuals, *eijt*, from the same subject are potentially correlated due to repeated measurements on the same patient. Here, an autoregressive covariance structure of AR(1) is used in the model, which states that errors on the same subject are correlated by Corr(*eijt*, *eijt’*) = *ρ*|*t*-*t*’|, t and t` refers to the different endpoints on the same subject. All time points (Day 14, month 1, month 2, month 3 and month 6) were included in the analysis.

## Results of Proof of Concept Study of GSK123456

Subjects in Cohorts 1 through 6 receive 0.03mg/kg, 0.3mg/kg, 3mg/kg (2 cohorts of subjects, Table 4.2), 10mg/kg and 30mg/kg of GSK123456. A starting dose of 0.03mg/kg GSK123456 was used for Cohort 1, and all the doses in subsequent cohort in Part A are identified, as described earlier, using a Bayesian *Emax* dose-finding model to provide 90% of maximal benefit (ED90). The dose escalation was performed between cohorts to assess any potential safety issues. If there were no obvious safety concerns, the subjects in the next cohort would receive the dose predicted by Bayesian *Emax* modelling.

The Bayesian *Emax* model used a non-informative prior and integrated the accumulated information or data and prior into posterior distribution, in which a random draw generated a distribution of parameters of interest, for example, the estimated dose level to achieve 90% of maximal effect also called ED90.

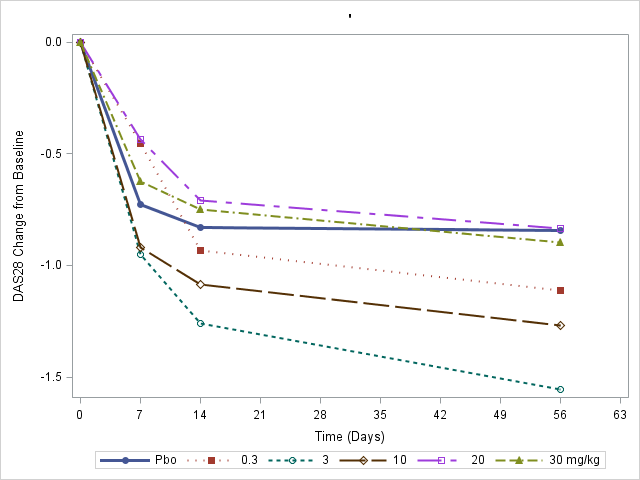
The formal interim analysis (IA1) was carried out when the first 6 cohorts completed up to Day 56 of all assessments. After first 6 cohorts, the data were unblinded and an interim analysis was undertaken. It was discovered that there was a possible U-shape dose response and the peak decrease of DAS28 at change from baseline was at 10 mg/kg. There was a small sample size at this dose. Thus a dose 10 mg/kg was also assigned to cohort 7 - to repeat the apparently efficacious dose - as well as intermediate dose of 20 mg/kg (median dose between 20 mg/kg and 30 mg/kg) which was assigned to cohort 8 to investigate whether there was any treatment effect at 20 mg/kg. Therefore, following this, two more cohorts were enrolled to confirm the dose response and there were total 12 patients who received 10 mg/kg and 20 mg/kg (cohort 7 and 8 in Table 4.2). There are 3 patients who are mis-dosed in the first cohort, one of them receives 0.06mg/kg and the other two receive 0.3mg/kg, the 10 times exposure of originally planned dose. The data were analysed using the treatment dose level they actually received.

Table 4.2 Number of subjects and treatment received at each cohort in Part A. In the first cohort, the planned dose level is 0.03 mg/kg, three subjects receive the wrong doses, two at 0.3 mg/kg and one at 0.06 mg/kg level. The 0.3 mg/kg dose of cohort 1 has been combined with the corresponding dose in cohort 2.

|  |  |  |
| --- | --- | --- |
| Cohort(s) | Treatment | Number of Subjects |
| 1 | 0.03 mg/kg | 3 |
| 1 | 0.06 mg/kg | 1 |
| 2 | 0.3 mg/kg | 8 |
| 3 and 4 | 3 mg/kg | 12 |
| 5 and 7 | 10 mg/kg | 12 |
| 8 | 20 mg/kg | 6 |
| 6 | 30 mg/kg | 6 |
| 1-8 | Pooled placebo | 16 |

## Summary of Study Results

In Part A, 3mg/kg group of GSK123456 on Day 56 showed a statistically significant clinical response, on mean DAS28 score change from baseline (P-value=0.0378; Figure 4.2). The significant result was based on a comparison of 12 patients receiving 3mg/kg and 16 patients receiving placebo combining all 8 cohorts. However, in the second part of the trial (Part B) no significant results are shown in the repeated dosing of 6mg/kg group (n=37) in comparison with placebo (n=17) and it was thus concluded that that there no evidence of efficacy for the dose selected (P-value=0.5962; Part B). All analyses were performed in SAS 9.2.



\*

Figure 4.2 Time profile of DAS28 change from baseline for placebo, dose 0.3mg/kg, 3mg/kg, 10mg/kg, 20mg/kg and 30mg/kg. The means of each dose group are displayed at Day 0, Day 7, Day 14, and Day 56 time points. The placebo is the blue line. No error bar is displayed for demonstration purpose.

## Outcome of Compound GSK123456

As started in the previous section Part B of the PoC study is the “confirm” part of the two-part design and the study results failed to demonstrate the effect of GSK123456 in comparison to placebo as the P-value for the treatment comparison was greater than 5% (P-value=0.5962). A decision was therefore made to discontinue the study compound due the lack of efficacy.

A decision by the company was been made to pursue follow-on compound (GSK654321) in the same drug class as GSK123456. The follow-on compound GSK654321 binds to the same molecular site and *in vitro* assay showed that the follow-on compound is more potent than the GSK123456.

## Adaptive Design Discussion

Although the compound GSK123456 failed to show efficacy compared to placebo and thus was a “failed study”, the original design of the PoC study for GSK123456 had many merits. The dose finding section (Part A) of the design accommodated two distinct functions in this trial, the dose escalation and dose finding. The dose escalation carefully escalated to higher doses to allow for an assessment of safety between cohorts. The dose finding selected an efficacious and optimal dose for the Part B of the trial. The study design combined two functions (learning and confirming) thus presented an attractive study design option for GSK654321.

In addition, the adaptive design applied a Bayesian dose response adaptive design to a clinical trial in a RA patient population. The experience of the adaptive design will help in the design of the Phase 2a study for the follow-on compound GSK654321.

The dose level for the next cohort was estimated based on the Bayesian *Emax* model using the prior accumulated cohort data. Safety data such as AE, lab data and ECG vital sign data were also presented for review. The project team used all the available information to determine the most appropriate dose level for the next cohort. During the next dose escalation, cumulative data from Cohort 1 and Cohort 2 (up to month 2) were used to inform the decision making for the next cohort.

Although there are many merits, this study design has its disadvantages. The *Emax* model was used to guide the selection of the dose for the next cohorts. This analysis approach assumes the dose response increases monotonically. When the fundamental monotonic assumption is violated, it makes the estimation of ED90 difficult.

It took more than two years to complete the Part A of the study. The primary endpoint is DAS28 at Day 56 and Part A of the study was run by cohorts which meant that the next cohort only started after the dose escalation meeting. So the time to obtain the primary endpoint and decision was long.

In the next chapter, we will explore simulation work to further support the study design of the follow-on compound GSK654321. In chapter 6, an alternative dose response model, the NDLM model, is proposed to compare the performances of the *Emax* model under various assumptions about the shape of the dose response curves - flat curve, *Emax* like curve, Log-linear curve and U-shaped curve. In Chapters 7 and 8 a delayed response model is proposed to predict the Day 56 response using Day 14 data, so the time to decision at between cohort escalations can be reduced.

# Simulation to support Adaptive Design in Proof of concept Study in follow-on compound GSK654321

In the previous chapter, we used a RA compound – GSK123456 as an example, to illustrate the application of the Bayesian adaptive design in order to examine whether GSK123456 was efficacious in patients with RA disease.

There were some learnings from this study even though the compound failed to show efficacy in the clinical trial. Firstly, it was demonstrated that an adaptive dose finding design could be applied successfully in a Phase 2a PoC study. Secondly, in Part A, the *Emax* model analysis was used guide the Bayesian selection of dose levels for the next cohorts, assuming a monotonic dose-response relationship.

The PoC study design of GSK123456 forms the basis of the Phase 2a PoC design of follow-on compound GSK654321.The proposed a study for GSK654321will also have still two parts. Part A is a “learning” phase which will have a total of 8 cohorts of RA patients with a dose finding design to find the optimal dose. Part B is a “confirming” phase using a group sequential design with two interim analyses. Computer simulations will be used to evaluate the adaptive study design under various hypothetical scenarios.

## The Aims of the Chapter

The aims of this chapter are to introduce the background to the Phase 2a PoC design, show why we use such a design, and demonstrate the operating characteristics of the proposed Bayesian dose response design using simulations.

### **Recommendation on PoC design of GSK654321 Based on Learnings** from GSK123456

Similar to GSK123456, the follow-on compound GSK654321 is a humanized monoclonal antibody (mAb) and is under development for the treatment of RA. GSK654321 is currently under pre-clinical and early clinical development. It binds to the same receptor as GSK123456 and is in the same class drug as GSK123456 with more potency, so the follow-on compound is expected to achieve the equal or higher efficacy.

A two-part Phase 2a PoC design of GSK654321 is also proposed based on the learnings from GSK123456, in Part A, cohort randomization design will be used to allocate more patients to the optimal dose (around ED90). In Part B, a group sequential design with O’Brien Fleming stopping rules will be used to evaluate efficacy.

### Part A of Phase 2a PoC study of GSK654321

The proposed Bayesian dose response model including *Emax* model and NDLM model and the comparison of *Emax* and NLDM models under difference dose responses (i.e. U-shaped curve) in the context of cohort randomization design of Phase 2a PoC study will be discussed in Chapter 6.

A similar two-part Phase 2a PoC design of GSK654321 is also proposed based on the learnings from GSK123456, in Part A, a (within) cohort randomized design will be used to allocate more patients to the optimal dose (ED90). In this cohort randomization design, the subjects are allocated to all available doses since there is no dose limitation resulting from safety. There are options of fixed design and adaptive design with non-adaptive, half adaptive and adaptive allocation. Taking half-adaptive design as an example, the first 50% of subjects are fixed allocated using pre-defined allocation ratio of treatments and placebo followed by adaptive allocation for the rest of the subjects based on the posterior distribution of dose near ED90 estimated which is the dose to achieve 90% of maximum DAS28 response as defined in section 4.2.4, based on either *Emax* model or NDLM model.

The placebo is given to a fixed proportion of the sample size allocation to ensure we have enough power for treatment comparisons vs. placebo. The fixed proportion is 25% of the total sample size. For each study dose (0.03, 0.3, 3, 10, 20 and 30 mg/kg), 4 patients (50% of the planned sample size) will be randomized first, prior to any interim analysis. The dose response curve will then be fitted using the dose response model and ED90 is estimated. For each subject randomized into the study afterwards, the dose level will be randomized to the dose close to the ED90 dose response. The iSRC will review all the efficacy and safety data during the interim analysis.

### Part B Design of GSK654321

The proposed study design of Part B of GSK654321 is similar to the study design for GSK123456. The primary analysis includes two-sided O’Brien-Fleming (OBF) boundaries for early stopping for efficacy or futility. Part B expands the number of subjects at a specified dose after the review of the results at the end of Part A. Subjects will be randomized on a 2:1 basis to a selected dose of GSK654321 or placebo. A minimum of 18 subjects will be randomized in Part B in order to assess the safety and tolerability of repeat intravenous doses. The maximum sample size needed in the study is based on sample size calculation assuming variability from the Part A interim analysis (standard deviation of 1.1). Overall; a maximum of 54 subjects will be enrolled in Part B.

In order to understand the decisions in predicting the probability of success (PoS) and mitigate the developmental risk in Part B of this study, an assurance calculation, based in part on the results observed in Part A, may be used to predicting the PoS in Part B of the PoC design of GSK654321. Assurance is referred as the unconditional power and was first introduced into clinical study design by O’Hagan et al. (2005). Both power and assurance calculations make assumptions about the true treatment difference (θ). The difference between power and assurance comes in the underlying assumptions regarding this unobservable true treatment difference, and therefore the inference we can make from them.

Power assumes that the treatment difference is exactly θ and assurance is the PoS accounting for uncertainty around the true treatment difference. A power calculation enables the probability of success to be determined for a given, fixed value, of the true treatment difference. Assurance calculations take account of sampling variability in the data but alsoacknowledge that there is uncertainty in what we know and more importantly what we do not know about the true treatment difference.

Therefore, the choice of priors over the assumption about the true treatment difference will have an impact on the assurance calculation, which is an unconditional PoS incorporating uncertainty given what we know and what we do not know about the true treatment difference. In case of a linear or *Emax* like dose response observed in Part A of the PoC trial, an informative prior around treatment difference which will result in similar estimate of PoS as statistical power calculation, while in case of U-shaped dose response curve, a pessimistic prior could be used in the assurance calculation so it is predicated to have a lower PoS in Part B of the study.

#### Simulation in Group Sequential design (Part B)

The dose in Part B uses the optimal dose selected by Bayesian dose escalation design in Part A. A repeated measure analysis is used to analyse the primary endpoint DAS28 change from baseline at Day 56. A mixed effect model was used in the repeat measure analysis, including treatment, time point, and treatment by time point interaction as fixed effects and subject as a random effect. All time points (Day 14, month 1, month 2, month 3 and month 6) were included in the analysis. If P-value for treatment comparison is less than OBF nominal significance level (e.g. 0.049 at final analysis), a statistically significant difference will be declared.

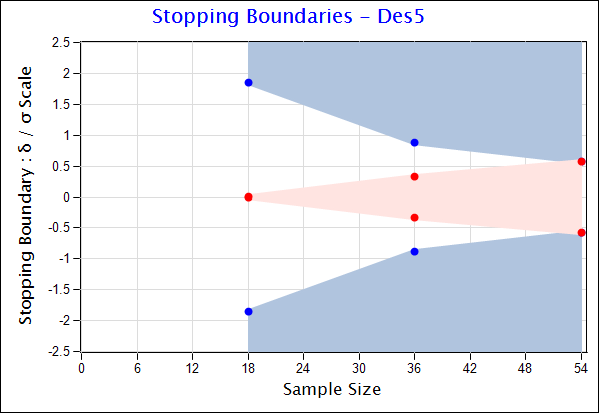
There are two planned interim analyses and one final analysis. Table 5.1 is the stopping boundary for both efficacy and futility. The nominal critical value (Z‑value), P-value, and delta/sigma are reported in the table below for interim 1, interim 2 and final analysis. The first interim analysis takes place when 33% of subjects completed the Month 2 (Day 56) measurements, and the second interim takes place when 66% of subjects completed the Month 2 (Day 56) measurement. The trial will stop for success if Z-value is greater than 3.71 and 2.51 at first interim and second interim respectively. If the Z-value is greater than 1.97, the significance is established at final analysis. The nominal critical value (Z value), P-value, and delta/sigma are interchangeable. 

Figure 5.1 Graphical view of OBF boundary for success (blue region) and futility (red region) in Part B of PoC study, the nominal critical point (delta/sigma) is displayed for 33%, 66% and final analysis at 100% patients enrolled. The graph was generated by *East* version 6.0.

Table 5.1 The stopping boundary of OBF with two interim analyses and one final analysis

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| % of Information | Z-value | | P-value | | Delta/sigma ratio | |
| Efficacy | Futility | Efficacy | Futility | Efficacy | Futility |
| Interim 1 (33%) | 3.71 | 0.01 | 0.00 | 0.99 | 1.69 | 0.00 |
| Interim 2 (66%) | 2.51 | 0.83 | 0.01 | 0.41 | 0.81 | 0.27 |
| Final Analysis (100%) | 1.97 | 1.97 | 0.049 | 0.049 | 0.52 | 0.52 |

Table 5.2 Operating characteristics of group sequential design in Part B at treatment effect (under null hypothesis of effect = 0, under alternative hypothesis of effect = -0.95, lower than expected effect (-0.3, -0.6) and greater than expected effects (-1.3,-1.6)) based on 10,000 simulations.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Expected Treatment effect (ΔDas28) | First Interim | | Second Interim | | Final Analyses | | Average Sample Size | Overall Power | % of Failed Trials |
| Suc  Stop | Fut  Stop | Suc  Stop | Fut  Stop | Suc  Stop | Fut  Stop |
| 0 | 0.2% | 0.8% | 1.6% | 58.1% | 3.0% | 36.2% | 46 | 4.9% | 95.0% |
| -0.3 | 0.5% | 0.6% | 5.5% | 43.9% | 12.9% | 36.4% | 48 | 19.1% | 80.9% |
| -0.6 | 1.9% | 0.5% | 20.5% | 18.8% | 31.6% | 26.9% | 50 | 53.4% | 46.1% |
| -0.95 | 6.9% | 0.0% | 51.3% | 2.7% | 32.4% | 6.6% | 45 | 90.7% | 9.3% |
| -1.3 | 18.2% | 0.0% | 69.8% | 0.1% | 11.4% | 0.4% | 38 | 99.4% | 0.6% |
| -1.6 | 34.5% | 0.0% | 63.4% | 0.0% | 2.2% | 0.0% | 33 | 100.0% | 0.0% |

\*Suc: Success; Fut: Futility.

We used simulation as a tool to investigate power, type I error, percentage of efficacy stopping and futility stopping in the proposed group sequential design with OBF alpha spending function via simulations. In the design, a maximum of 54 subjects will be enrolled in Part B. A clinically meaningful treatment difference of -0.95 between the selected dose and placebo in DAS28 scores at 56-day post dose will be detected with approximately 90% power if the alternative hypothesis that there is a treatment difference is true. Additional simulations have also been undertaken to investigate the operating characteristic such as power, type I error, percentage of efficacy stopping and futility stopping when treatment difference of 0, -0.3, -0.6, -1.3, and -1.6. These assume a standard deviation of 1.1 in the active dose group and in placebo with allocation ratio of treatment and placebo is 2:1, and a two-sided test and an overall alpha (α) of 5%.

Table 5.2 shows the probability of success and futility at first interim, second interim, and final analysis based on simulations. The timing of interim analysis is evenly spaced so the first interim analysis occurs when 18 patients’ follow-up data are available, the second interim analysis occurs when 36 patients have follow-up outcome data. The success rate is 6.9%, 51.3%, and 32.4% under the desired treatment effect of -0.95 at first interim, second interim, and final analysis respectively. The average sample size is calculated based on the sample size from 10,000 simulations. The average sample size is 45 under desired effect and overall statistical power of 90.7%. The type I error is 4.9% so the type I error is maintained below 5%.

The statistical power graph is shown in Figure 5.2. A sample size of 54 is reasonable to detect at least 90% power with treatment effect of 0.95 and a standard deviation of 1.0 in the treatment and placebo groups (allocation ratio of treatment and placebo is 2:1).

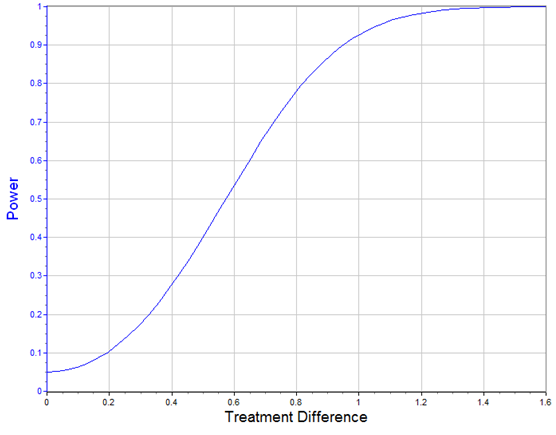


Figure 5.2 Probability of success curve also known as statistical power displayed as the expected treatment difference assuming fixed standard deviation of 1.0 in group sequential design. The treatment difference is defined as the absolute DAS28 difference of GSK654321 and placebo. Under alternative hypothesis, the desired absolute treatment difference is 0.95. The graph was generated by *East* version 5.4.

### Planned Statistical method at final analysis

For primary endpoint of DAS28 at final analysis, the repeated measure analysis using a mixed effect model will be used, as described in Chapter 4. In part B, a repeated measures analysis was used to analyze the primary endpoint DAS28 change from baseline at Day 56. A mixed effect model was used in the repeat measure analysis, including treatment, time point, and treatment by time point interaction as fixed effects and subject as a random effect. If the P-value for active treatment and placebo comparison was less than O’Brian Fleming (OBF) normal significance level, it would be taken that the study has observed a statistically significant difference.

## Summary

In this chapter, a similar 2-part adaptive design is planned for Phase 2a PoC trial of GSK654321. The group sequential design with the O’Brien Fleming method demonstrated that the type I error was controlled and a proposed sample size in Part B was able to detect the desired treatment difference of -0.95 in DAS28 change from baseline, with at least 90% power.

In the next chapter, a retrospective analysis will be performed to model the clinical trial data using Bayesian NDLM model and further use simulation to compare the *Emax* model and NDLM model in the context of fixed and adaptive Phase 2a PoC clinical trial setting.

# Improving RA Proof of concept trial with A better statistical model

In Chapter 4, a U-shaped dose response was observed in the Phase 2a PoC study of GSK123456. Therefore, in this chapter, an alternative dose response model, the NDLM model, is explored to compare the performances between the *Emax* model and the NDLM model under various assumptions (about the true dose-response) in the retrospective analysis of Part A and to compare them in fixed design and adaptive design settings using simulations.

## The Aims of the Chapter

There are four aims in this chapter:

1. Perform literature review on Frequency of U-shaped Curve and provide possible explanation
2. Perform a retrospective analysis of the GSK123456 PoC study;
3. Compare the Bayesian *Emax* model and Bayesian NDLM model in the context of GSK654321 Phase 2a PoC clinical trial setting.

## Frequency of U-shaped Curve and possible explanation

It was shown from Phase 2a results that the dose response may follow a U-shaped curve of DAS28 in patients receiving the study drug GSK123456 (Figure 6.1). Significant difference was shown in primary efficacy endpoints at 3 mg/kg in Part A; however, no significant statistical difference was shown in Part B. The study was therefore concluded to be a negative study (P-value=0.5962), however, there are some remaining questions to answer.

1) Why was the Bayesian *Emax* modelling not predictive of the dose response curve?

2) Can other models accommodate the non-proportional dose responses and improve the model fitting and selection of ED90 as the targeted dose?

3) What are the operating characteristics of the *Emax* model and NDLM model under fixed design and adaptive design scenarios?

4) How can the modelling fitting and performance be improved for future studies, and do any baseline characteristics explain the variability?

The *Emax* model is one of the most widely applied models relating drug concentrations to effects (Pinheiro, 2006; Thomas 2014). Much biological activity follows a 4-parameter or 3-parameter logistic model. The *Emax* model is commonly used in the modelling of biological effects and it assumes the drug response follows monotonic increase or decrease. In practice, the *Emax* model assumes the drug effect is proportional to the dose, i.e. the bigger the dose, the bigger the effect. The Bayesian *Emax* model failed to select the optimal dose since the concentration–effect relationships deviated from a simple proportional relationship without reaching a clear maximum (Figure 6.1, blue line is observed data and red dashed line is Emax model fitting).

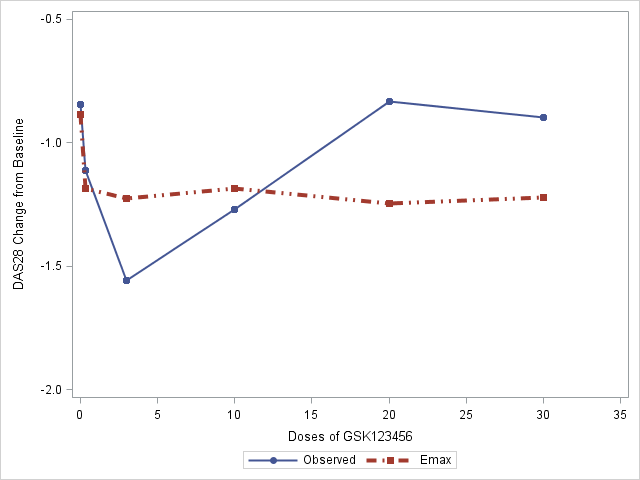


Figure 6.1 Mean and estimated dose response of mean change in DAS28 scores from baseline to day 56 post-randomisation of observed data (solid blue line) and modelling fitting with *Emax* model (red dashed line).

The statistical power will be much lower if the responses are not monotonic. An alternative model, the NDLM model, was introduced in Chapter 3. This model does not assume a monotonic relationship between increasing drug dose and effect. In this chapter, the questions are attempted to be addressed by reviewing the literature on U-shaped dose-response curves and by performing additional post-hoc analyses such as the permutation test and retrospective analysis to confirm whether or not the true dose response relationship is U-shaped or just a chance occurrence.

In the following sections, the frequency of U-shaped dose-response curves in general and in new compounds and drugs to treat RA patients are discussed respectively. A plausible biological explanation of the U-shaped dose-response curve observed for GSK123456 is given and the implication of this U-shaped dose-response curve for the follow-on compound on GSK654321 are discussed. The DAS28 study data were then analysed using quadratic model in comparison to linear model.

### Literature review on the frequency of the U-shaped dose response curve

To further understand the U shape dose response a further literature review was undertaken to assess the likelihood - based on the literature - of seeing this dose response relationship. This review was undertaken to assess if the dose response observed was plausible.

Reynolds (2010) reported that there was reference to a “U-shaped” dose response curves in 1000 molecules. The prevalence of a U-shaped dose response occurrence has been commented on in a number of review papers (Calabrese, 1997; Calabrese, 2001; Reynolds, 2010; Owen, 2014; Almstrup, 2002, Almstrup 2014). Additionally, it was shown (Reynolds, 2001) that evidences of hormesis, which was shown as a similar to U-shaped curve (Calabrese, 2014), have been found in 245 (37% of 668) dose-response relationships from 86 articles between 1966-1998. Complementary to this research in a comprehensive review undertaken to identify articles demonstrating chemical hormesis, one in 6 studies (17% out of 6000 studies) were shown to have the evidence of hormesis (Calabrese, 1997; Calabrese, 1999). Other reasons of U-shaped curves include angiogenesis – which was defined as growth of new blood vessels into tumor, facilitating tumor growth (Reynolds, 2009) and the therapeutic effect inhibited by immunogenicity.

### The frequency of U-shaped curve in biological to treat RA patients

A biological therapy is a type of biopharmaceutical drugs that are manufactured in or extracted from biological sources (Rader, 2008). A monoclonal antibody is a common type of biological investigative drug and may result in therapeutic effects in patients with certain disease conditions (Rader, 2008). A recent publication (Wu, 2017) summarized the clinical dose response for biological product and showed that 7% of biologics had non-monotonic dose response. Both GSK123456 and GSK654321 are monoclonal antibody.

Literature searches in PubMedline (pubmed.gov) using the term of ‘U-shaped dose response biological” and ‘U-shaped dose response antibody” resulted in 109 and 69 papers respectively (Table 6.1). There have been several reports of U-shaped dose response in Phase 2a studies in RA patients. One of examples is a test compound antibody – MOR103 (Behrens, 2015), which is a human monoclonal antibody to granulocyte-macrophage colony-stimulating factor (GM-CSF). In this study, the patients were randomized to placebo or to one of three doses of the drug. A U-shape curve was observed in the middle dose at 1.0 mg/kg which had the most pronounced effects.

Another example is the biological drug belimumab, a U-shape curve was shown in the primary analysis is at 24 weeks while low dose was shown to have the highest response in patients with RA who had failed previous therapies. For example, ACR20 responses with placebo and belimumab 1, 4, and 10 mg/kg were 15.9% (n=69), 34.7% (P = 0.010, n=72), 25.4% (p = 0.168, n=71), and 28.2% (p = 0.080, n=71), respectively (Stohl, 2013).  A dose of 1 mg/kg had the highest ACR20 response rate. Similarly, a humanized monoclonal antibody that was developed to treat RA patients, gevokizumab showed a U-shaped dose response in which moderate doses have more favourable anti-inflammatory effects in animal model (Owyang, 2010).

Table 6.1 Number of literature in PubMed, web of science and Google scholar with the search term (last access on July 2016)

|  |  |  |  |
| --- | --- | --- | --- |
|  | PubMed | Web of Science | Google Scholar |
| U-Shaped | 5847 | 257,960 |  |
| “U-Shaped dose response” | 762 | 1135 |  |
| “U-Shaped dose response biological” | 109 | 145 |  |
| U-shaped antibody | 69 | 132 |  |
| U-shaped dose response antibody | 11 | 19 |  |
| U-shaped dose response monoclonal antibody | 1 | 2 |  |
| Bell U-shaped dose response NDLM | 0 | 0 | 0 |
| Bell U-shaped dose response Emax | 0 | 0 | 60 |
| Bell U-shaped dose response NDLM Emax | 0 | 0 | 0 |
| Bell U-shaped dose response NDLM Emax RA | 0 | 0 | 0 |

### Possible Explanation of U-shaped curve to GSK123456

A dose response in a new class of compound or target is generally unknown and not well understood; especially if the drug has never been tested in healthy volunteers or patients. In part A of Phase 2a PoC study which GSK123456 is a humanised anti-Oncostatin M(OSM) monoclonal antibody, Choy (2013) suggested that the U-shaped curve may be due to a protein carrier effect which GSK123456 has moderate binding affinity and rapid off-rate of as compared to the higher affinity OSM receptor. At low doses of study drug, GSK123456 was unable to neutralise OSM. At moderate dose, GSK123456 effectively neutralised the activity of OSM therefore a drug effect is observed. While at high doses of GSK123456, all OSM will form complex with the study drug so the there was less drug effect; therefore, the protein carrier effect may be the cause of U-shaped curve (Choy, 2013).

### Retrospectively Analysis of DAS28 data using Permutation Test on Polynomial Regression model

A linear regression model was initially used to fit the DAS28 change from baseline data at Day 56 (primary endpoint) and Day 84 (follow-up time-point) with dose as explanatory variable. The residual diagnosis showed a pattern of curvature. Therefore, a quadratic polynomial regression model (formula 3.5 when n=2) is used to fit the DAS28 change from baseline data with dose, dose square, and interaction between dose and dose square as explanatory variables. Interaction term was dropped due to insignificant effects. Observed data and modelling fitting with linear quadratic model are presented in Figure 6.2 (blue line for observed data and green square dotted line for model fitting of linear quadratic model).

There were total 60 patients and 59 patients who have record of Day 56 and Day 84 were included in the model fitting. Only the groups whose sample size is greater than 3 are included in the analysis. The number of patients in dose 0 (placebo), 0.3, 3, 10, 20 and 30 mg/kg are 18, 8, 12, 12, 6, and 6 patients respectively. The subject received dose at a half-log dose scale, therefore, dose variable was log-transformed in the analysis.

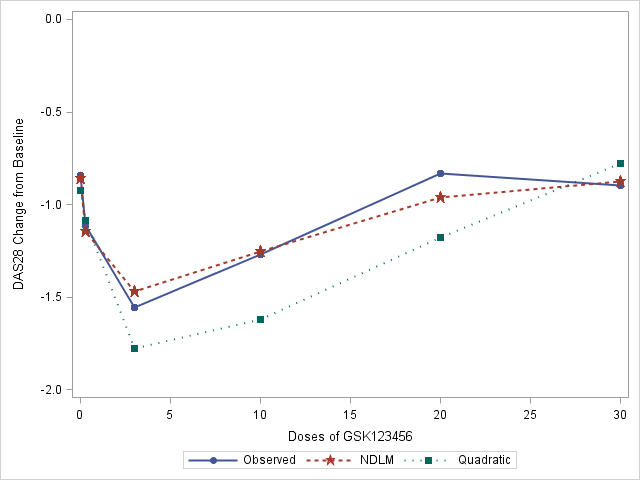


Figure 6.2 Mean and estimated dose response of mean change in DAS28 scores from baseline to day 56 post-randomisation with observed data (solid blue line) and modelling fitting with NDLM model (red star dashed line) and linear quadratic model (green square dotted line).

The model fitting parameters are presented in Table 6.2 for linear regression model (a) and quadratic model (b). There was no significant effect of baseline DAS28 (P-value > 0.5) and thus baseline DAS28 was removed from model fitting.

A permutation test was used to calculate the P-values. Permutation tests are a class of widely-applicable non-parametric tests. The key concept of permutation tests was first described by Fisher (1935), followed by further development of exact permutation by Dwass (1957) and permutation using resampling methods (Good 1994). A permutation test randomly shuffles the data across all dose groups to get the distribution of a test statistic, the P-value in our example under a null hypothesis. Generally, the permutation test is more computationally intensive. The null and alternative hypotheses for the permutation tests for the U-shaped curve is

Hypothesis of permutation test for linear regression are

H0: There is no linear relationship between DAS28 change from baseline and dose in logarithm scale (the slope of β for independent variable of dose is zero);

H1: There is a linear relationship between DAS28 change from baseline and dose in logarithm scale (the slope of β for independent variable of dose is not zero).

Hypothesis of permutation test for quadratic polynomial regression are

H0: A quadratic equation does not fit the data significantly better than a linear equation (the slope of β2 for quadratic term is zero);

H1: There is a quadratic relationship between DAS28 change from baseline and dose in logarithm scale (the slope of β2 for quadratic term is not zero).

There is no analytical derivation of test statistics in the permutation test. Through a large number of permutations, a distribution of test statistics is constructed under null hypothesis. One of the assumptions is that all possible test statistics are equally likely to occur, where the test statistics in the observed data is one of the possible. The distribution of the test statistics of the quadratic polynomial model and an illustration of with the permutation test the for testing β2 part of the quadratic term is presented in Figure 6.3. In this graph, the test statistics is /SE() and percentage of test statistics is plotted using histogram plot with a Normal distribution curve overlaying. It was shown that distribution of test statistics of DAS28, /SE(), are likely to be Normally distributed. Vertical dotted lines denote the observed test-statistics (-2.82) from phase 2a PoC of GSK123456. The proportion of permutations less than -2.82 is the P-value, which are shown in Table 6.3.

The sample size is too large for exact permutation and so a permutation test is based on approximation method to approximate the P-value, which is illustrated in below steps.

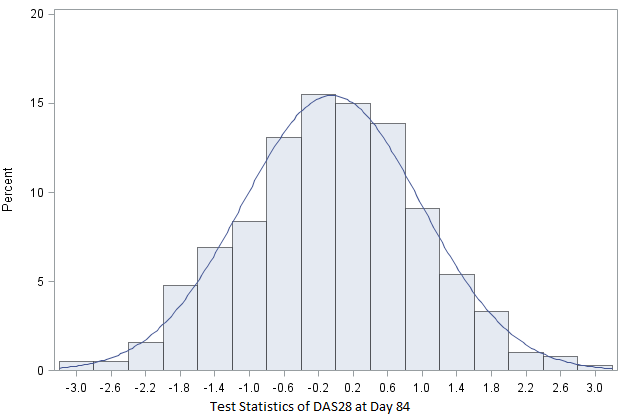
For linear regression,

1. Fit a polynomial regression model and calculate the test statistics t-value of slope b, tobs , from the observed data.

2.  Assuming all dose groups have similar treatment effects under null hypothesis, so without replacement, randomly select observations for respective dose groups (the sample size are 15, 8, 12, 12, 6, 6 for 0, 0.3, 3, 10, 20, 30 mg/kg respectively), calculate the t-value of slope b, for that permutation.

3.  Repeat Step 2 at 5000 times and then we have 5000 t-values, the p-value of permutation test is calculated as percentage of these t-values great than *t*obs.

This is the approximation method, the larger the number of permutation is, the more precision the approximation is. Repeat the above step 1-3 for linear quadratic polynomial regression model. The permutation analysis is performed in SAS 9.3 and the results are presented in Table 6.2. SAS code is available in Appendix 11.1.3.



Test statistics of observed

PoC data t =-2.82

Figure 6.3 Histogram of t-test statistics of DAS28 at Day 84 with permutation test for testing β2 of quartic term is zero or not. Vertical dotted lines denote the t-test of observed data from phase 2a PoC of GSK123456.

Table 6.2. Permutation test on analysis of DAS28 using linear and polynomial regression model.

|  |  |
| --- | --- |
| Variable: DAS28 at Day 84 | Permutation P-value |
| Y (response) = a + b\*log\_dose. |  |
| Dose (log scale) | 0.206 |
|  |  |
| Y (response) = a + b\*log\_dose+ c\*(Log\_dose\*log\_dose). | |
|  |  |
| LogDose\*LogDose (log scale) | 0.006 |

There is no linear relationship or trend when the DAS28 is fitting using linear regression (P-value of 0.206), however, a trend of negative quadratic curve was observed at DAS28 and the P-value for quadratic variable (LogDose\*LogDose) is 0.006.Statistical significances of the quadratic effects may be indicative of a non-monotonic curve in the DAS28 data.

### Implication of U-shaped curve to GSK123456 and GSK654321

Overall, it seems highly plausible that DAS28 followed a U-shaped dose response curve in the PoC study of GSK123456 which impacts on considerations for GSK654321. The follow-on compound GSK654321 is in the same drug class as GSK123456. It binds to the same binding site as GSK123456 and it is believed to have therapeutic properties but with higher potency. Therefore, the chance of U-shaped dose response cannot be ruled out.

Given the literature on U-shaped curves, the results observed and the rationale within GSK123456 for the dose response, it was proposed to adopt a more flexible model in the development of the follow-on compound to handle both monotonic and non-monotonic dose response in the design and analysis consideration for the follow on compound GSK654321.

### Retrospective Analysis using NDLM Model

In this section, a retrospective analysis of DAS28 data from the GSK123456 PoC trial is undertaken using a NDLM model. The NDLM or state space model is a very general class of non-stationary time series models which include a term to model trends, seasonality, covariates and autoregressive components (Weir, 2007). It is a flexible model and does not assume monotonic dose response.

The U-shaped dose response curve was an unexpected finding from the GSK123456 PoC clinical trial. To investigate whether NDLM model can improve the model fitting and statistical power, the data from the GSK123456 study are retrospectively analysed using a NDLM approach. The NDLM modelis applied to DAS28 change from baseline data at 56. Baseline DAS28, as a covariate, is adjusted regressing to mean. The NDLM model makes no monotonic assumption, and is therefore a flexible approach to dose response analysis.

### Bayesian Normal Dynamic Linear Model (NDLM)

A NDLM can be used to fit to estimate the dose-response relationship. Similar to Section 3.7.8, a description of the NDLM used in the analysis is shown below,

jk = θj + βXk +ε, where ε ~ N(0, σ2 ) , 6.1

θj = θj−1 + δj−1 + ωj where j>= 2 and ωj ∼ N(0, σθ2 ) ,

δj = δj−1 + ǫj where ǫj ∼ N(0, σδ2 ), with

θ1 ~ N(0, σ12) and δ1 ~ N(0, σ12)

where DAS28jk is the observed individual change in DAS28 score from baseline to day 56 post-randomisation at Dosej for kth subject (Dose is placebo when j=1). DAS28 at Day 56 change from baseline follows a Normal distribution with mean (θj + βXk)for eachDosej and with variance of σ2, the Dosej is assumed to be spaced equally. θj is the estimated treatment effect at Dosej, Xk is the baseline DAS28 score fitted as covariate and β is the coefficient for covariate. Furthermore, θj has a linear relationship with neighbouring θj−1 with slope of δj−1(slope parameter). θ1 is the untreated or placebo response when the drug dose is equal to 0 and δ1 is the slope between placebo and the first dose. Both θ1 andδ1 follow Normal distributions. Bayesian NDLM method was used in the model estimation and it calculated the estimate based on sampled data from posterior distribution of multivariate normal distribution. Bayesian methods (Weir, 2007) offer a robust estimation of parameters in complicated model (Carlin, 2010).

IN the retrospectively analysis, the priors of the effect of both the first dose θ1 and δ1 follow a Normal distribution with the mean following a vague prior of *N*(0, 100) and variance of inverse-gamma distribution *IG*(0.001,0.001)). The prior distributions on σ2, σθ2, and σδ2 are vague inverse-gamma distribution *IG*(0.001,0.001)). WinBUGS code is available in Appendix 11.1.2. The results of modelling fitting of NDLM model is presented as red dashed line in Figure 6.2.

### Results of Retrospective Analysis Using NDLM

U-shaped dose response curve was an unexpected finding from the Phase 2a PoC clinical trial of GSK123456. To investigate whether NDLM model can improve the model fitting and statistical power, the data from the GSK123456 study are retrospectively analysed using a NDLM approach. The NDLM modelis applied to DAS28 change from baseline data at 56. Baseline DAS28, as a covariate, is adjusted regressing to mean. The SAS software (SAS version 9.2) was used to call WinBUGS (Spiegelhalter, 2000) using WinBUGSio macro (Smith 2007) to run NDLM dose response analysis. The number of MCMC iterations as 100,000 with burn-in of 1,000 are used to achieve desired MCMC precision. The study futility and study efficacy (negative effect in DAS28 improvement) are defined as,

*P(futility)=P(effect over placebo at x dose level < 0) < 20%*

*P(efficacy) = P(effect over placebo at x dose level < 0) > 80%*

If there is high chance of efficacy based on the above pre-specified criteria, the probability of success in the study dose is higher and is worthy of further development. If the chance of dose effect over placebo is lower than 20%, there is a high probability of futility. The probability of effect over placebo for each dose larger than 0.8, 0.95 and 1.2 is calculated from Bayesian posterior distribution. The Box-whisker plot of posteriors samples of DAS28 change from baseline is presented in Figure 6.3 and descriptive statistics of effect of dose 0.3, 3, 10, 20 and 30 mg/kg over placebo at Day 56 from MCMC samples after Bayesian NDLM model fitting is presented in Table 6.3.

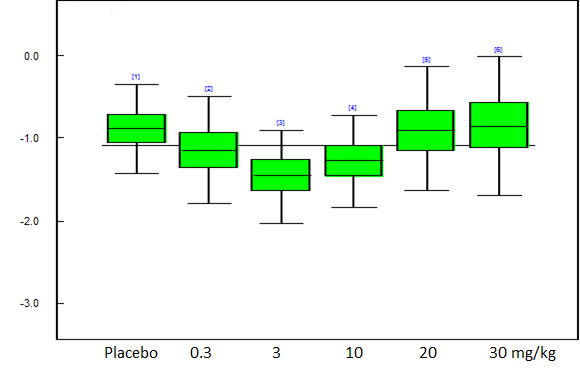


Figure 6.3 Box-whisker plot of DAS28 change adjusted for the baseline plot at Day 56 at multiple planned doses after NDLM fitting, the middle line in the box is the median and upper and lower whisker represent 2.5% and 97.5% percentile sampling from posterior distribution with NDLM framework.

Table 6.3 Descriptive statistics of effect of dose 0.3, 3, 10, 20 and 30 mg/kg over placebo at Day 56 from MCMC samples from posterior distribution of NDLM model fitting. Treatment effect is difference of placebo and treatment. The data from 0.03mg/kg was not included in the analysis due to the sample size of less than 4.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Treatment Effect** | **Mean** | **SD** | **MC error** | **2.50%** | **Median** | **97.50%** | **# of simulations** |
| 0.3 mg/kg | 0.2570 | 0.3914 | 0.0012 | -0.5170 | 0.2578 | 1.0310 | 100000 |
| 3 mg/kg | 0.5617 | 0.4000 | 0.0013 | -0.2199 | 0.5601 | 1.3540 | 100000 |
| 10 mg/kg | 0.3876 | 0.3940 | 0.0012 | -0.3888 | 0.3873 | 1.1630 | 100000 |
| 20 mg/kg | 0.0323 | 0.4654 | 0.0016 | -0.8926 | 0.0338 | 0.9406 | 100000 |
| 30 mg/kg | -0.0269 | 0.5115 | 0.0018 | -1.0340 | -0.0272 | 0.9844 | 100000 |

Table 6.4 Probability of effects (absolute value of treatment effect over placebo) greater than 0, 0.8, 0.95 or 1.2 at Day 56 using Bayesian NDLM model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Doses | Pr(effect>0) | Pr(effect>0.8) | Pr(effect>0.95) | Pr(effect>1.2) |
| 0.3 mg/kg | 75.3% | 8.0% | 3.8% | 1.0% |
| 3 mg/kg | 92.3% | 27.3% | 16.3% | 5.5% |
| 10 mg/kg | 84.2% | 14.5% | 7.5% | 2.0% |
| 20 mg/kg | 52.3% | 4.8% | 2.4% | 0.6% |
| 30 mg/kg | 47.3% | 5.3% | 2.9% | 0.9% |

The probability of effect over placebo for each dose larger than 0 and 0.95 was calculated from Bayesian posterior distribution. Adjusting the baseline DAS28 as a covariate shows that the 3mg/kg and 10mg/kg dose levels have 92.3% and 84.2% probability of achieving the effect of treatment over placebo greater than zero respectively, and have more than 16.3% and 7.5% chance of achieving an effect of treatment over placebo greater than 0.95 (Table 6.4).

In summary, DAS28 data were fitted using the NDLM model retrospectively to estimate the dose-response relationship. Adjusting for the baseline DAS28 as a covariate, it was shown that the 3mg/kg and 10mg/kg doses are optimal for achieving the effect of treatment over placebo greater than zero. The NDLM analysis for dose response model seems to be a good model to support the dose selection and to handle both monotonic and non-monotonic to dose response.

In the observed data we have shown how a *Emax* model was sub-optimal in modelling the dose response. We have applied a NDLM model and demonstrated it to be superior. It is important to highlight however that *Emax* was the pre-specified analysis. NDLM was retrospectively applied. For NDLM to be prospectively planned for GSK654321 there is needed first to do some evaluation of its properties in the context of a Phase 2a PoC study design.

In next sections, we will explore the NDLM model, to compare the performances of the *Emax* model and the NDLM under various assumptions about the shape of the dose response curves - flat curve, *Emax* like curve, Log-linear curve and U-shaped curve.

## Dose response models in the evaluation

As highlighted in the previous section in a retrospective analysis the NDLM seemed to model the dose response better than an *Emax* model for a Phase 2a study of the new proposed treatment for RA – GSK123456. The NDLM seemed to be a superior approach for the analysis of the dose response. Before a NDLM could be prospectively planned for GSK654321 we need to make an evaluation of the NDLM through simulation of a number of scenarios we anticipate as being likely for the dose response profile for the new drug. The mathematical details of *Emax* model and NDLM model are described in section 4.2.5 and 6.2.7 respectively.

For the simulations, four true dose response profiles (Figure 6.4) are used for the primary endpoint, change in DAS28 score from baseline to Day 56, to mimic the wide range of dose response scenarios likely to be observed and be analysed as dose response methods in clinical practice. In all models, the placebo effect (on the background of MTX) was set to be -0.5. That is a change in DAS28 score from baseline to day 56 post-randomisation of -0.5 points i.e. a small decline/improvement is disease activity. The error term ε was assumed to be independently Normally distributed with a mean of 0 and a variance of 1.44 for *Emax*, Log linear and U shaped curve, which was the estimated variance from Phase 2a study of GSK123456 (Figure 6.4), the error term has variance 0.25 for profile 4 -- placebo like response.

|  |  |
| --- | --- |
| Profile 1 | Flat curve: DAS28=-0.5+ε |
| Profile 2 | *Emax* curve: y DAS28=-0.5-1.7\*Dose/(2.5+Dose)+ε, ED50 is 2.5 |
| Profile 3 | Log linear curve:DAS28=-0.5 -log(Dose+1) +ε |
| Profile 4 | U-Shaped curve: DAS28 follows a predefined U shape curve explicit with:DAS28 = (-0.5, -0.7, -1.6, -1.8, -1.2, -1, -0.6) for dose 0, 0.03, 0.3, 3, 10, 20, and 30 mg/kg respectively. |

These four profiles were chosen as plausible dose responses for the new compound in development GSK654321 ranging from a null effect (Profile 1) to what was previously observed with GSK123456 (Profile 4).

Figure 6.4 Dose-response profiles used in the simulation. The model profiles include a placebo like flat curve which is denoted in blue and is fixed at -0.5 for all dose levels, a dose proportional *Emax* model in red, a Loglinear model in green, and a U-shaped model in purple. The label for the vertical axis is the change in DAS28 score from predose to day 56 post-randomisation.

The scenarios of fixed design simulation and adaptive design simulation are discussed in the next section. The two basic designs set up are a fixed design and an adaptive design. The fixed design assumes that all 6 doses and placebo are allocated to a fixed number of patients. No adaptations are applied in this design. In the adaptive design, the subjects are allocated according to the dose responses of all the subjects enrolled in the study.

### Design of the Simulation Study

The range of doses is between 0.03mg/kg to 30mg/kg. The design is a parallel design and the total target sample size is 64. The available doses in the evaluations are 0.03, 0.3, 3, 10 ,20 and 30 mg/kg. The goal of the trial is to characterize the dose-response curve at various doses. The fixed design assumes that all 6 doses are allocated to a fixed number of patients with no interim analysis or adaptation of the dose. In the adaptive simulation, the subjects are allocated due to the subjects’ response in the study at the end of each cohort.

Decisions regarding success and futility of the trial at completion are made based on the probability of DAS28 relative to control greater than clinically significant difference (a decrease of 0.95 as measured by DAS28 change from baseline between placebo and treatment). The positive difference of placebo and treatment is used to facilitate the positive effect and probability calculation. All the designs except fixed scenarios include 8 cohorts, with 8 patients in each cohort (2 on placebos and 6 on active treatment).

An adaptive design was used in the PoC design of GSK123456 and is considered as a better option than fixed design since it increases the chance of stopping a failed compound and expediting a good one as well as potentially maximizing the information on the doses which are most interest to carry forward for later development. For GSK654321 the study design has not been finalised. The wish therefore was to evaluate modelling the dose response using NDLM or *Emax* for different options for the study design which we have detailed. The follow-on compound GSK654321 is in the same drug class as GSK123456 which demonstrated good safety and tolerability in the PoC study (Choy 2013)., so there is no single dose escalation planned for the PoC study in GSK654321. ED90 is defined as the dose to achieve 90% of maximum DAS28 response with the lower dose chosen if there are multiple values. In this calculation the maximum response is estimated from the maximal DAS28 effect at all doses. The 90% (ED90) of maximum response is then calculated as the lowest nominal dose in which is closest to the estimated dose that achieves 90% of maximal efficacy. The following fixed design as well as adaptive design scenarios are considered in the design options and evaluations.

|  |  |
| --- | --- |
| Scenario 1: | Fixed design; the design is non-adaptive, the study allocates 8 patients to receive doses of GSK654321 (0.03, 0.3, 3, 10, 20 and 30 mg/kg) and 16 patients to receive placebo. There is no interim stopping and adaptation in the fixed design. The evaluation of final success will occur at the end of the study. |
| Scenario 2 | Cohort randomization design with no adaptive allocation; the ratio of patients (100% of the planned sample size) randomized into each study dose level (placebo, 0.03, 0.3, 3, 10, 20 and 30 mg/kg) are 2:1:1:1:1:1:1. The placebo is given to a fixed proportion of the sample size allocation to ensure there is enough power for treatment comparisons vs. placebo. There are a total of 8 cohorts (6 treated + 2 placebo) and the interim analysis will occur between cohorts, for example, at 8 patients, 16 patients, 24 patients, 32 patients (50%), 40 patients (62.5%) and 48 patients (75%) enrolled and complete the primary endpoint assessment (Day 56 post-randomisation DAS28 score). The study is evaluated with the interim study success and interim study futility. |
| Scenario 3 | Half adaptive, the first 50% of subjects are fixed allocated using pre-defined allocation ratio of treatments and placebo followed by adaptive allocation for the rest of the subjects based on the posterior distribution of dose around ED90; the placebo is given to a fixed proportion of the sample size allocation to ensure we have enough power for treatment comparisons vs. placebo. The fixed proportion is 25% of the total sample size. For each study dose (0.03, 0.3, 3, 10, 20 and 30 mg/kg), the 4 patients (50% of the planned sample size) will be randomized first, prior to any interim analysis. The dose response curve will then be fitted using the dose response model and ED90 is estimated. For each subject randomized into the study afterwards, the dose level will be randomized to the dose close to the ED90 dose response. The interim analysis will occur at 32 patients (50%), 40 patients (62.5%) and 48 patients (75%) that complete the primary endpoint assessment at Day 56. The maximal allowable dose is 30 mg/kg. If the posterior ED90 exceeded 30mg/kg, the maximum planned dose of 30mg/kg would be used. The study is evaluated for interim study success and interim study futility. |
| Scenario 4 | Cohort randomization design with adaptive allocation. In the fully adaptive simulation, the placebo is given a fixed proportion of the sample size allocation to ensure there is enough power for treatment comparisons vs. placebo. The fixed proportion is 25% of the total sample size. The dose response curve will be fitted using the dose response model and ED90 is estimated. For each subject randomized into the study afterwards, the dose level will be randomized to the dose close to ED90 dose response. the interim analysis will occur between cohorts, for example, at 8 patients, 16 patients, 24 patients, 32 patients (50%), 40 patients (62.5%) and 48 patients (75%) enrolled and complete the primary endpoint assessment. The maximal allowable dose is 30 mg/kg. If the posterior ED90 exceeded 30mg/kg, the maximum planned dose of 30mg/kg would be used. The study is evaluated for interim study success and interim study futility. |

Comparison of dose allocation algorithm in simulation scenarios.

|  |  |  |  |
| --- | --- | --- | --- |
| Fixed (S1) | No adaptive (S2) | Half adaptive (S3) | Fully adaptive (S4) |
| Equal allocation  No stopping  Ratio is are 2:1:1:1:1:1:1 for placebo, 0.03, 0.3, 3, 10, 20 and 30 mg/kg | Ratio is are 2:1:1:1:1:1:1 for placebo, 0.03, 0.3, 3, 10, 20 and 30 mg/kg.  Interim analysis is planned between each cohort at 8 patients, 16 patients, 24 patients, 32 patients (50%), 40 patients (62.5%) and 48 patients (75%) | First 32 patients (50%) of subject, the ratio of allocation is 2:1:1:1:1:1:1 for placebo, 0.03, 0.3, 3, 10, 20 and 30 mg/kg.  Interim analysis is conducted after 32 patients complete all assessment.  Subjects will be assigned receive the dose based on ED90 for the remaining 32 subjects.  Interim analysis is conducted after 40 and 48 patients complete all assessment. | Placebo is assigned as 25% of sample size and subjects will be assigned receive the dose based on ED90.  the interim analysis will occur between cohorts, for example, at 8 patients, 16 patients, 24 patients, 32 patients (50%), 40 patients (62.5%) and 48 patients (75%) enrolled and complete the primary endpoint assessment. |

A summary of dose and patient allocation algorithm among four design scenarios is discussed in Table 6.5. The simulation and analysis are performed using a data simulation and analysis software - FACTs (Fixed and Adaptive Clinical Trial Simulator) version 2.1 developed by Tessella and Berry Consultant. Simulated data are fitted using similar Bayesian *Emax* model and NDLM models as described in Section 4.2.5 and 6.2.7 respectively. All the doses were half-log spaced at design stage with the exception of 20 mg/kg. The maximum dose level across the study cohorts is 30 mg/kg. The 30mg/kg dose is the maximum tolerated dose for the study based on prior studies. If the posterior mean of ED90 exceeded 30mg/kg, the maximum planned dose of 30mg/kg is used. It is possible that the choice of informative prior impacts the simulation results, for consistency and comparison purpose, a non-informative prior is chosen in the calculation and simulation. For the Bayesian Emax model, the priors of model parameters *E0* and *Emax* follow a Normal distribution with large variance i.e. *N*(0,1E4) and the prior distribution of ED50 follows normal distribution *N*(3,1E2). The prior on σ2 is an inverse gamma distribution (*IG*(0.5,0.7). For the Bayesian NDLM model, the covariates of baseline DAS28 are not included in the model. The prior distribution on θhas a vague Normal distribution with a large variance estimated from inverse-gamma distribution (*IG*(0.5, 72). The prior distributions on σ2 and the evolution variance σθ2 and σδ2 are inverse-gamma distribution (*IG*(0.5, 72). Additionally, selected informative priors are explored in the simulations. The simulation starts with fixed seed and all results are based on 5000 simulations. The number of simulations and number of MCMC simulations as 2,500 with burn-in of 500 are chosen based on the estimated minimum precision.

### Decision Criteria in Adaptive Design Simulation

Decision criteria for interim success, interim futility, final success and final futility in the adaptive design simulation are displayed in below table. In the Bayesian analysis, the parameter of interest is modelled as a distribution and investigates the probability of the difference of any dose treatments and placebo greater than 0 which is analogous to a typical frequentist test with cut-off P-value of 5%. Therefore, the interim success criteria are the chance is greater than 70% in the probability of the difference of any dose treatment and placebo greater than 0.95 (a clinically meaningful difference) and 95% in the in the probability of the difference of any dose treatment and placebo greater than 0. The final success is defined as 95% in the in the probability of the difference of any dose treatment and placebo greater than 0.

For the fixed design, the final success is based on at least 95% posterior probability that the dED90 dose achieves a drug effect greater than the control or placebo, otherwise it is final futility. For all other adaptive design (scenario 2, 3, and 4), the decision criteria of the interim success, interim futility, final success and futility are presented in Table 6.6.

Table 6.5 Decision criteria at the interim analysis and final analysis in the proposed design scenarios.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Decision Criteria** | InterimSuccess | InterimFutility | FinalSuccess | FinalFutility |
| Pr(|RED90 –Ctrl| > 0) | >95% | <20% |  |  |
| Pr(|RED90 –Ctrl| > 0.95) | >70% |  |  |  |
| Pr(|Rdmax –Ctrl| > 0) > 0.95 or Pr(|RED90 –Ctrl| > 0) >95% |  |  | Yes | No |

Pr(|RED90 –Ctrl| > 0): The probability of dose response near ED90 dose level achieves a drug effect greater than the control or placebo.

Pr(|RED90 –Ctrl| > 0.95): The probability of dose response near ED90 dose level achieves a drug effect greater than the control or placebo and 0.95 is the clinical significant difference.

Pr(|Rdmax –Ctrl| > 0): The probability of any dose with maximal effect achieves a drug effect greater than the control or placebo.

Only final success and futility are accessed in fixed design.

When there is truly is no effect or a placebo like effect, the Type I error rate is calculated based on the chance of rejecting the null hypothesis (when it is true). In the context of this simulation it would also be the chance of incorrectly accepting that the drug has a dose response - or the false positive rate.

#### Simulation Results

Scenario 1 is non-adaptive, therefore the study allocates 8 patients to receive doses of GSK123456 (0.03, 0.3, 3, 10, 20 and 30 mg/kg) and 16 patients to receive placebo. All doses including placebo are top of standard first-line therapy, therefore a placebo like response is a -0.5 as a measure of DAS28 change from baseline. Figure 6.5 is the subject allocation for the proposed fixed design, the subject allocations are the same for all four types of dose response model.

When there is no effect or a placebo like effect, the type I error rates are 6% and 17% when analysed using the *Emax* model and the NDLM model respectively (Table 6.6). A type I error rate of 17% signals that the NDLM model is over-sensitive and inflate the false positive so appropriate adjustment to the test statistics or decision criteria may be needed.

The analysis of the *Emax* and NDLM models on the *Emax* curve and linear curve also show the same trend. The probability of success is 98% for both *Emax* and NDLM models if the true response curve follows *Emax* data and 96% and 95% for log linear curve for *Emax* and NDLM models respectively. For U-shaped dose response curves, the power measured as the probability of success of the *Emax* model is 26% and power using the NDLM model is 92%. The NDLM model significantly improves the probability of success compared to the *Emax* model in the fixed design model.

Table 6.6 Probability of final success and failure in fixed design using Bayesian *Emax* model and NDLM model in Scenario 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumed True Dose Response** | **Bayesian *Emax* model** | | **Bayesian NDLM model** | |
| **Prob. of final success** | **Prob. of final failure** | **Prob. of final success** | **Prob. of final failure** |
| Placebo like flat curve | 0.06 | 0.94 | 0.17 | 0.83 |
| *Emax* with EC50=3 | 0.98 | 0.02 | 0.98 | 0.02 |
| LogLinear Curve | 0.96 | 0.04 | 0.95 | 0.05 |
| U-Shaped curve | 0.26 | 0.74 | 0.92 | 0.08 |

### Adaptive Design Simulation

Fixed design assumes that all 6 doses are allocated to a fixed number of patients. No adaptations were adopted in this design. In the adaptive simulation, study may stop for futility or success or subjects are allocated to certain dose based on the dose response of the subjects already enrolled in the study.

In this section, there are three scenarios for which we will use both the Bayesian *Emax* model and Bayesian NDLM model to evaluate dose response curves (dose response follows *Emax* model, log linear model, U-shaped, and placebo like flat curve). The three scenarios are:

1) No adaptive allocation (Scenario 2);

2) Half adaptive, the first 50% of subjects are fixed allocated using pre-defined allocation ratio of treatments and placebo followed by adaptive allocation for the rest of the subjects (Scenario 3);

3) Adaptive allocation in 100% of subjects (Scenario 4).

In Bayesian analysis, the prior distribution can be combined with new data to produce the new modified distribution of probabilities, posterior distribution. Adaptive analysis takes advantage of real-time data and has a lower risk of inflating the type I error, so multiple interim analyses can be performed. Weak priors are used in the simulations.

#### Minimal effective dose

As described in section 6.3.1, the future subject is allocated to the dose level of ED90 from a Bayesian dose response model of *Emax* of NDLM model. As well as ED90, the minimal effective dose (MED) and the clinical significant dose are also of interests since they provide a comparison between each dose level of treatment and control or placebo. The minimal effective dose, also called the MED, is defined as the lowest dose that will produce a desired outcome, in this study the lowest dose to achieve 95% posterior probability that the dose achieves a drug effect greater than the control or placebo. The proportion of simulation being selected as MED is presented in the result section. MED dose can be described using below formula, Pr(R > Ctrl) > 0.95, where R is any dose (0.03, 0.3, 3, 10, 20 or 30 mg/kg). The proportion of simulated trials selecting the MED is presented in the result section.

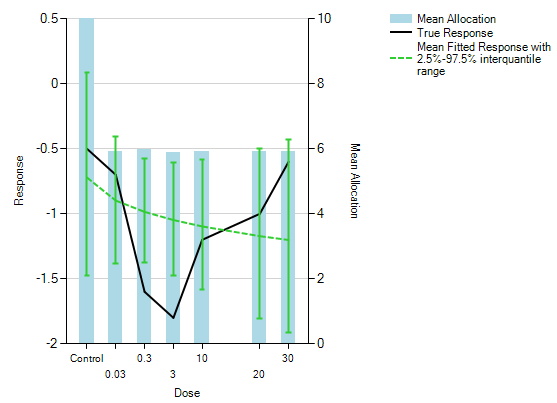
#### Clinical significant dose (CSD)

The clinical significant dose, also called the CSD, is defined by Pr(|R – Ctrl|>0.95) > 0.70, where R is any dose (0.03, 0.3, 3, 10, 20 or 30 mg/kg). The proportion of simulation being selected as CSD is presented in the result section.

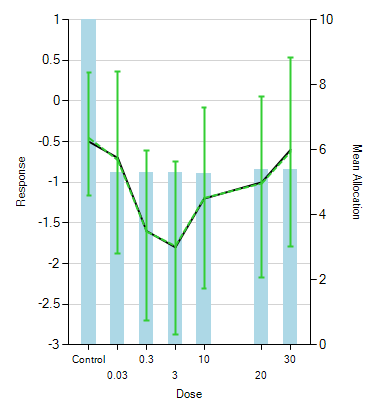
#### Scenario 2 – No Adaptive Allocation

Figure 6.6 shows the predicted mean response and mean dose allocation for each dose level when we assume the data follows a U-shaped curve. For comparison, Bayesian *Emax* model (A) is presented together with Bayesian NDLM model (B) to perform the dose escalation and dose finding.

Both the *Emax* model and NDLM model were able to fit the data well when the true effects are a placebo like response, *Emax* like response, or log linear like response (Data not shown). When the dose response follows a U-shaped curve, the *Emax* model was not able to fit the U-shaped curve, however, a more flexible semi-parametric NDLM model fitted the data well.



A



B

Figure 6.6. True response and simulated mean fitted response plots with 95% interquartile range (Y-left) and mean dose allocation for each dose level (Y-right) assuming dose response following U-shaped curve. A: upper plot: the dose response of DAS28 was analysed with Bayesian *Emax* model. B: lower plot: the dose response of DAS28 was analysed with Bayesian NDLM model. Weak informative priors were used in the calculation.

The NDLM model is semi-parametric and more flexible, so the model fitting is better in all four data curves. The figures show the predicted mean response and mean allocation are similar for the NDLM model and the *Emax* when we assume the dose response is monotonic. In the case of non-monotonic dose response, the NDLM model performed better by fitting the U-shaped curve. The dose response is not dose proportional and more than 80% of simulations choose 3mg/kg as ED90. The predicted means decrease proportionally up to 3mg/kg and then start to increase. The Bayesian NDLM model is able to select the correct dose at 3mg/kg since 3mg/kg is also the dose that achieved the maximal response in this simulation

Table 6.7 Probability of success and failure at interim and final analysis in adaptive design settings, the mean subjects are also summarized in the last column across 5,000 simulations. A) Bayesian *Emax* model in the analysis dose response models; B) Bayesian NDLM model in the analysis dose response models.

A

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| True Dose Response | Bayesian *Emax* model | | | | |
| Prob. of early success | Prob. of early failure | Prob. of final success | Prob. of final failure | Mean subjects |
| Placebo like flat curve | 0.00 | 0.39 | 0.04 | 0.57 | 51 |
| *Emax* with EC50=3 | 0.74 | 0.03 | 0.19 | 0.04 | 38 |
| Log Linear Curve | 0.64 | 0.05 | 0.24 | 0.08 | 40 |
| U-Shaped curve | 0.10 | 0.26 | 0.14 | 0.50 | 47 |

B

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| True Dose Response | Bayesian NDLM model | | | | |
| Prob. of early success | Prob. of early failure | Prob. of final success | Prob. of final failure | Mean subjects |
| Placebo like flat curve | 0.03 | 0.03 | 0.16 | 0.78 | 58 |
| *Emax* with EC50=3 | 0.82 | 0.01 | 0.09 | 0.08 | 37 |
| Log Linear Curve | 0.74 | 0.00 | 0.15 | 0.11 | 40 |
| U-Shaped curve | 0.63 | 0.01 | 0.17 | 0.18 | 43 |

For no allocation design (Scenario 2), the Bayesian *Emax* model and NDLM model have similar probabilities of success (early success + final success: 93% vs. 92%) and probability of futility (7% vs. 8%) in *Emax* like true dose response. A similar probability of success and futility has been shown in the data when the true data follows a linear curve. When using the Bayesian NDLM model, the adaption of dose allocation improves the model performance and the probability of success (24% vs. 80%) when the data change non-monotonically (U-shaped curve). In the case of placebo like flat dose response curve, the probably of success was 4% and 18% when fitting with the *Emax* model and NDLM model respectively. Therefore, a NDLM model is likely to inflate the type I error rate.

Table 6.8 and table 6.9 give the additional operating characteristics of the model fitting to the data that were analysed using the *Emax* model and NDLM model respectively. The mean allocation, proportion of simulation where the dose was selected by ED90/MED/Max, and probability of each being statistical significant compared to placebo are displayed with each of the four curves being the true effect. It was shown that the Bayesian *Emax* model is able to find the correct dose for ED90 (10 mg/kg, 20mg/kg or 30mg/kg) when the true response is either an *Emax* curve or log linear curve, however, it cannot find the correct ED90 dose when the true response follows a U-shaped curve. All data show that the NDLM model is a more flexible model in all four types of model and is able to identify the correct ED90 doses when the true response followed a U-shaped curve. NDLM model is able to identify the correct ED90 doses 82% or 89% of the time when the true response is an *Emax* or log linear curve respectively.

Table 6.8 Comparisons of the operating characterises of four types of dose response curve using Bayesian *Emax* model in the adaptive design settings, weak prior was used in the calculation with adaptive allocation. Placebo allocation is fixed at sample size of 16 (25% of total sample size).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Dose Level (mg/kg)** | | | | | |
| **0.03** | **0.3** | **3** | **10** | **20** | **30** |
| Mean allocation |  |  |  |  |  |  |
| Flat placebo like curve | 6.4 | 6.4 | 6.4 | 6.4 | 6.4 | 6.4 |
| *Emax* like Curve | 4.7 | 4.7 | 4.8 | 4.7 | 4.7 | 4.7 |
| Log Linear Curve | 5.0 | 5.0 | 5.1 | 5.0 | 5.1 | 5.0 |
| U Shape Curve | 5.9 | 6.0 | 5.9 | 5.9 | 5.9 | 5.9 |
| Proportion of simulation being selected as ED90 |  |  |  |  |  |  |
| Flat placebo like curve | 0% | 0% | 0% | 0% | 35% | 0% |
| *Emax* like Curve | 0% | 0% | 0% | 0% | 92% | 8% |
| Log Linear Curve | 0% | 0% | 0% | 0% | 83% | 16% |
| U Shape Curve | 0% | 0% | 0% | 0% | 71% | 0% |
| Proportion of simulation being selected as MED |  |  |  |  |  |  |
| Flat placebo like curve | 0% | 0% | 0% | 0% | 0% | 0% |
| *Emax* like Curve | 1% | 23% | 45% | 5% | 5% | 0% |
| Log Linear Curve | 0% | 11% | 41% | 5% | 9% | 0% |
| U Shape Curve | 2% | 2% | 3% | 0% | 0% | 0% |
| Proportion of simulation being selected as MAX |  |  |  |  |  |  |
| Flat placebo like curve | 51% | 0% | 0% | 0% | 0% | 49% |
| *Emax* like Curve | 0% | 0% | 0% | 0% | 0% | 100% |
| Log Linear Curve | 1% | 0% | 0% | 0% | 0% | 99% |
| U Shape Curve | 19% | 0% | 0% | 0% | 0% | 81% |
| Probability of dose compared to control |  |  |  |  |  |  |
| Flat placebo like curve | 49% | 49% | 49% | 49% | 49% | 49% |
| *Emax* like Curve | 98% | 98% | 98% | 98% | 98% | 98% |
| Log Linear Curve | 97% | 97% | 97% | 97% | 97% | 97% |
| U Shape Curve | 73% | 73% | 73% | 73% | 73% | 73% |
|  |  |  |  |  |  |  |

Table 6.9. Comparisons of the operating characteristics of four types of dose response curve using Bayesian NDLM model in the adaptive design settings. Weak prior was used in the calculation with adaptive allocation.

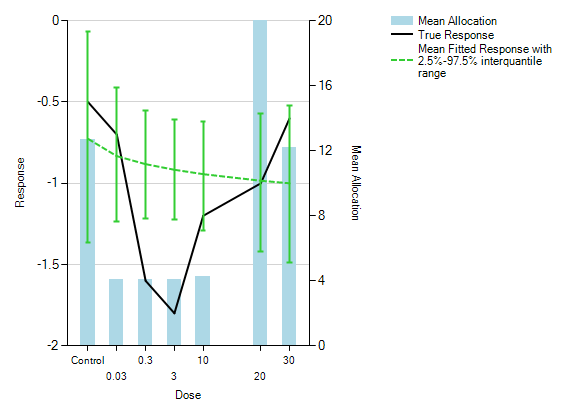
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| True Dose Response | **Dose Level (mg/kg)** | | | | | |
| **0.03** | **0.3** | **3** | **10** | **20** | **30** |
| Mean allocation |  |  |  |  |  |  |
| Flat placebo like curve | 7.3 | 7.3 | 7.3 | 7.3 | 7.3 | 7.3 |
| *Emax* like Curve | 4.6 | 4.7 | 4.7 | 4.6 | 4.7 | 4.7 |
| Log Linear Curve | 4.9 | 4.9 | 4.9 | 4.9 | 4.9 | 5.0 |
| U-Shaped Curve | 5.3 | 5.3 | 5.3 | 5.3 | 5.4 | 5.4 |
| Proportion of simulation being selected as ED90 |  |  |  |  |  |  |
| Flat placebo like curve | 18% | 17% | 18% | 13% | 13% | 14% |
| *Emax* like Curve | 1% | 2% | 15% | 24% | 33% | 25% |
| Log Linear Curve | 2% | 3% | 6% | 13% | 28% | 48% |
| U-Shaped Curve | 4% | 38% | 44% | 7% | 6% | 2% |
| Proportion of simulation being selected as MED |  |  |  |  |  |  |
| Flat placebo like curve | 0% | 0% | 0% | 0% | 0% | 0% |
| *Emax* like Curve | 24% | 18% | 33% | 15% | 6% | 2% |
| Log Linear Curve | 25% | 16% | 16% | 16% | 15% | 6% |
| U-Shaped Curve | 22% | 51% | 16% | 1% | 0% | 0% |
| Proportion of simulation being selected as MAX |  |  |  |  |  |  |
| Flat placebo like curve | 16% | 17% | 19% | 15% | 14% | 18% |
| *Emax* like Curve | 0% | 1% | 11% | 20% | 35% | 33% |
| Log Linear Curve | 1% | 2% | 4% | 10% | 27% | 56% |
| U-Shaped Curve | 2% | 32% | 46% | 10% | 8% | 3% |
| Probability of dose compared to control |  |  |  |  |  |  |
| Flat placebo like curve | 51% | 51% | 52% | 50% | 50% | 51% |
| *Emax* like Curve | 58% | 66% | 87% | 91% | 94% | 93% |
| Log Linear Curve | 58% | 64% | 74% | 83% | 90% | 95% |
| U-Shaped Curve | 61% | 89% | 93% | 79% | 72% | 56% |
|  |  |  |  |  |  |  |

For all simulations, the proportion of simulations being selected as ED90 is 8% vs. 25% using the *Emax* model and NDLM model respectively in analyzing the data when the true model is the *Emax* model with EC50 as 3mg/kg. The true ED90 of the model was 30mg/kg. Adding 20 and 30 mg/kg makes the proportion of simulation being selected as ED90 100% vs. 58% using the *Emax* model and NDLM model respectively.

If the trial data has a U-shaped curve, the proportion of simulation being selected as ED90 is 0% vs. 44% using the *Emax* model and NDLM model respectively in analyzing the data when the true ED50 is around 3mg/kg.

#### Scenario 3 – Half Adaptive

Figure 6.7 shows the predicted mean response and mean dose allocation for each dose level when we assume the data follow a placebo like flat curve, *Emax* curve, log linear curve, and U-shaped curve in the half adaptive design. For comparison, Bayesian *Emax* model (A) was presented together with Bayesian NDLM model (B) to perform the dose escalation and dose finding. The figures show that both the *Emax* model and the NDLM model are able to fit the data well when the true effects are a placebo like response, *Emax* like response, or log linear like response. When the dose response follows a U-shaped curve, the *Emax* model is not able to fit the U-shaped curve, however, a more flexible semi-parametric NDLM model fitted the data well.



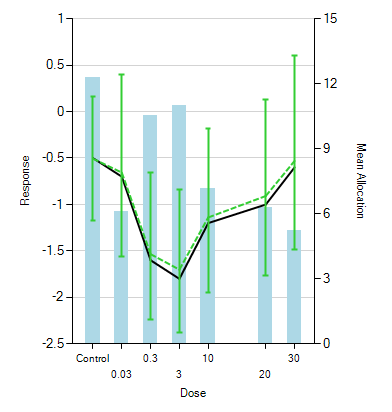


Figure 6.7 True response and simulated mean fitted response plots with 95% interquartile range (Y-left) and mean dose allocation for each dose level (Y-right) assuming dose response following U-shaped curve. Upper plot: the dose response of DAS28 was analysed with the Bayesian *Emax* model. Lower plot: the dose response of DAS28 was analysed with the Bayesian NDLM model. Weak informative prior was used in the calculation.

Table 6.10 Probability of success and failure at interim and final analysis in adaptive design settings, the mean subjects are summarized in the last column across 5,000 simulations. A) Bayesian *Emax* model in the analysis dose response models; B) Bayesian NDLM model in the analysis dose response models

A

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| True Dose Response | Bayesian *Emax* model | | | | |
| Prob. of early success | Prob. of early failure | Prob. of final success | Prob. of final failure | Mean subjects |
| Placebo like flat curve | **0.00** | **0.23** | **0.04** | **0.73** | **61.2** |
| *Emax* with EC50=3 | **0.74** | **0.00** | **0.26** | **0.01** | **54.8** |
| Log Linear Curve | **0.58** | **0.00** | **0.40** | **0.02** | **57.0** |
| U-Shaped curve | **0.06** | **0.09** | **0.10** | **0.75** | **61.5** |

B

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| True Dose Response | Bayesian NDLM model | | | | |
| Prob. of early success | Prob. of early failure | Prob. of final success | Prob. of final failure | Mean subjects |
| Placebo like flat curve | 0.00 | 0.00 | 0.12 | 0.88 | 63.9 |
| *Emax* with EC50=3 | 0.53 | 0.00 | 0.44 | 0.03 | 57.4 |
| Log Linear Curve | 0.45 | 0.00 | 0.47 | 0.08 | 57.9 |
| U-Shaped curve | 0.41 | 0.00 | 0.47 | 0.12 | 58.6 |

Under half adaptive allocation design, the Bayesian *Emax* model and NDLM model have a similar probability of success (early success + final success: 99% vs. 97%) and probability of futility (1% vs. 3%) in *Emax* like true dose response (Table 6.10). A weak or vague prior is given to both models. A similar probability of success and futility has been shown in the data when the true data follows a linear curve. When using the Bayesian NDLM model, adaption of dose allocation improves the model performance and probability of success (15% vs. 88%) when the data changes non-monotonically (U-shaped curve). In the case of a placebo like flat curve, the probably of success was 4% and 12% when fitting with the *Emax* model and NDLM model respectively. Therefore, a NDLM model is likely to inflate the type I error rate.

Table 6.11 and Table 6.12 display the additional operating characteristics of the model fitting to the data that were analysed using the *Emax* model and NDLM model respectively. The mean allocation, proportion of simulation where the dose is selected as ED90/MED/Max, probability being statistical significant compared to placebo are displayed under half adaptive design setting with each of the four curves being the true effect. It is shown that the Bayesian *Emax* model is able to find the correct dose for ED90 (10 mg/kg, 20mg/kg or 30mg/kg) when the true response is either an *Emax* curve or log linear curve, however, it cannot find the correct ED90 dose when the true response follows a U-shaped curve. All data show that the NDLM model is more flexible in all four types of models and is able to identify the correct ED90 doses when the true response followed a U-shaped curve. NDLM is able to identify the correct ED90 doses 86% or 94% of the time when the true response is an *Emax* or log linear curve respectively.

Table 6.11 Comparisons of the operating characterises of four types of dose response curve using Bayesian *Emax* model in the adaptive design settings, weak prior was used in the calculation with adaptive allocation. Placebo allocation is fixed at sample size of 16 (25% of total sample size).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Dose Level (mg/kg)** | | | | | |
| **0.03** | **0.3** | **3** | **10** | **20** | **30** |
| Mean allocation |  |  |  |  |  |  |
| Flat placebo like curve | 4 | 4 | 4 | 4 | 19.5 | 13 |
| *Emax* like Curve | 4 | 4 | 4 | 4.1 | 15.9 | 11.2 |
| Log Linear Curve | 4 | 4 | 4 | 4 | 16.7 | 12.2 |
| U-Shaped Curve | 4 | 4.1 | 4.1 | 4.2 | 20.2 | 12.2 |
| Proportion of simulation being selected as ED90 |  |  |  |  |  |  |
| Flat placebo like curve | 0% | 0% | 0% | 0% | 38% | 0% |
| *Emax* like Curve | 0% | 0% | 0% | 0% | 89% | 11% |
| Log Linear Curve | 0% | 0% | 0% | 0% | 81% | 19% |
| U-Shaped Curve | 0% | 0% | 0% | 0% | 64% | 0% |
| Proportion of simulation being selected as MED |  |  |  |  |  |  |
| Flat placebo like curve | 0% | 0% | 0% | 0% | 0% | 0% |
| *Emax* like Curve | 1% | 8% | 52% | 10% | 16% | 0% |
| Log Linear Curve | 0% | 3% | 37% | 17% | 22% | 0% |
| U-Shaped Curve | 2% | 0% | 0% | 0% | 0% | 0% |
| Proportion of simulation being selected as MAX |  |  |  |  |  |  |
| Flat placebo like curve | 51% | 0% | 0% | 0% | 0% | 49% |
| *Emax* like Curve | 0% | 0% | 0% | 0% | 0% | 100% |
| Log Linear Curve | 0% | 0% | 0% | 0% | 0% | 100% |
| U-Shaped Curve | 27% | 0% | 0% | 0% | 0% | 73% |
| Probability of dose compared to control |  |  |  |  |  |  |
| Flat placebo like curve | 49% | 49% | 49% | 49% | 49% | 49% |
| *Emax* like Curve | 100% | 100% | 100% | 100% | 100% | 100% |
| Log Linear Curve | 100% | 100% | 100% | 100% | 100% | 100% |
| U-Shaped Curve | 67% | 67% | 67% | 67% | 67% | 67% |
|  |  |  |  |  |  |  |

Table 6.12 Comparisons of the operating characteristics of four types of dose response curve using Bayesian NDLM model in the adaptive design settings. Weak prior was used in the calculation with half-adaptive allocation.

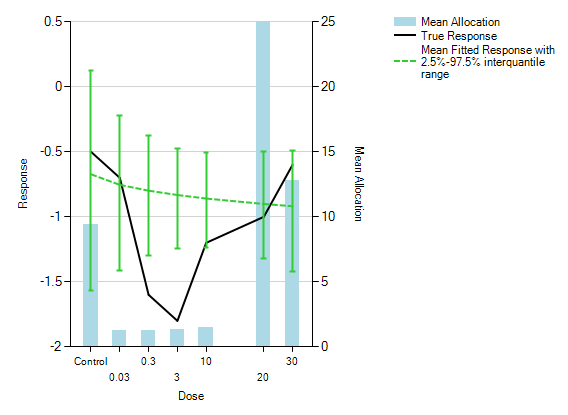
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Dose Level (mg/kg)** | | | | | |
| **0.03** | **0.3** | **3** | **10** | **20** | **30** |
| Mean allocation |  |  |  |  |  |  |
| Flat placebo like curve | 9.4 | 8.8 | 8.3 | 8.4 | 7.9 | 8.0 |
| *Emax* like Curve | 4.9 | 5.4 | 7.6 | 8.9 | 9.3 | 9.1 |
| Log Linear Curve | 5.4 | 5.6 | 6.6 | 7.8 | 9.4 | 11.0 |
| U-Shaped Curve | 6.1 | 10.5 | 11.0 | 7.2 | 6.3 | 5.2 |
| Proportion of simulation being selected as ED90 |  |  |  |  |  |  |
| Flat placebo like curve | 16% | 17% | 14% | 14% | 11% | 12% |
| *Emax* like Curve | 1% | 1% | 14% | 25% | 35% | 26% |
| Log Linear Curve | 1% | 1% | 5% | 11% | 29% | 54% |
| U-Shaped Curve | 2% | 42% | 48% | 5% | 2% | 0% |
| Proportion of simulation being selected as MED |  |  |  |  |  |  |
| Flat placebo like curve | 0% | 0% | 0% | 0% | 0% | 0% |
| *Emax* like Curve | 18% | 15% | 37% | 18% | 6% | 1% |
| Log Linear Curve | 17% | 14% | 17% | 16% | 16% | 10% |
| U-Shaped Curve | 13% | 54% | 17% | 0% | 0% | 0% |
| Proportion of simulation being selected as MAX |  |  |  |  |  |  |
| Flat placebo like curve | 15% | 18% | 16% | 17% | 17% | 18% |
| *Emax* like Curve | 0% | 0% | 7% | 19% | 36% | 38% |
| Log Linear Curve | 0% | 1% | 2% | 8% | 24% | 65% |
| U-Shaped Curve | 1% | 33% | 54% | 8% | 3% | 1% |
| Probability of dose compared to control |  |  |  |  |  |  |
| Flat placebo like curve | 45% | 45% | 46% | 45% | 44% | 45% |
| *Emax* like Curve | 57% | 67% | 89% | 94% | 96% | 96% |
| Log Linear Curve | 55% | 61% | 74% | 84% | 91% | 96% |
| U-Shaped Curve | 58% | 91% | 95% | 79% | 69% | 51% |
|  |  |  |  |  |  |  |

For all the simulations, the proportion of simulation being selected as ED90 is 11% vs. 26% using the *Emax* model and NDLM model respectively in analyzing the data when the true model is the *Emax* model with EC50 as 3mg/kg. The true ED90 of the model was 30mg/kg. Adding 20 and 30 mg/kg makes the proportion of simulation being selected as ED90 100% vs. 61% using the *Emax* model and NDLM model respectively.

If the trial data has a U-shaped curve, the proportion of the simulation being selected as ED90 is 0% vs. 48% using the *Emax* model and NDLM model respectively in analyzing the data when the true ED50 is around 3mg/kg. Other simulation data like the proportion of simulation being selected as ED50 and Max all show a similar trend.

#### Scenario 4 - Adaptive Allocation

The operating characterises of fully adaptive design is shown in Figure 6.8 when we assume the dose response is non-monotonic. When the dose response is not dose proportional and more than 80% of simulations choose 3mg/kg as ED90. The predicted means decrease proportionally up to 3mg/kg then start to increase. The Bayesian NDLM model is able to select the correct dose at 3mg/kg since 3mg/kg is the dose that achieved the maximal response in this simulation.



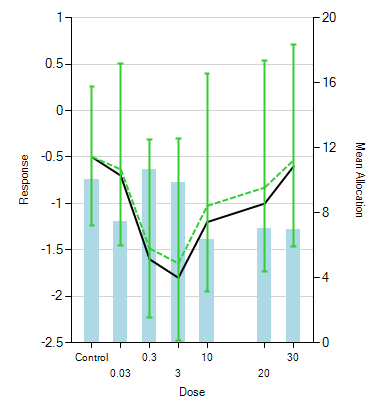


Figure 6.8 True response and simulated mean fitted response plots with 95% interquartile range (Y-left) and mean dose allocation for each dose level (Y-right) assuming dose response following U-shaped curve. Upper plot: the dose response of DAS28 was analysed with Bayesian *Emax* NDLM model. Lower plot: the dose response of DAS28 was analysed with Bayesian NDLM model. Weak informative prior was used in the calculation.

The study design is a fully adaptive design which means that the patients are allocated to the dose nearest ED90. Therefore, the pattern of patients’ allocation to each dose level is similar to the data using the Bayesian *Emax* model. The placebo rate is fixed and there are more patients randomized to 20mg/kg and 30mg/kg. It is shown in the figures that both the *Emax* model and NDLM model are able to fit the data well when the true effects are a placebo like response, *Emax* like response, or log linear like response. When the dose response follows a U-shaped shaped curve, the *Emax* model is not able to fit the U-shaped curve, however, a more flexible semi-parametric NDLM model fitted the data well. Further bias assessment (Figure 6.12) supports this conclusion.

Under adaptive allocation design, the Bayesian *Emax* mode and NDLM model have a similar probability of success (early success + final success: 94% vs. 92%) and probability of futility (6% vs. 8%) in *Emax* like true dose response. The early success is higher in the *Emax* model (77% vs. 51%). A vague prior is given to both models. A similar probability of success and futility has been shown in the data when the true data follows a linear curve. When using the Bayesian NDLM model, adaption of dose allocation improves the model performance and probability of success (17% vs. 80%) when the data change non-monotonically (U-shaped curve). In case of a placebo like flat curve, the probably of success is 6% and 12% when fitting with the *Emax* model and NDLM model respectively. Therefore, a NDLM model is likely to inflate the type I error rate.

Table 6.13 Probability of success and failure at interim and final analysis in adaptive design settings, the mean subjects are summarized in the last column across 5,000 simulations. A) Bayesian *Emax* model in the analysis dose response models; B) Bayesian NDLM model in the analysis dose response models.

A

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| True Dose Response | Bayesian *Emax* model | | | | |
| Prob. of early success | Prob. of early failure | Prob. of final success | Prob. of final failure | Mean subjects |
| Placebo like flat curve | 0.00 | 0.33 | 0.06 | 0.61 | 55 |
| *Emax* with EC50=3 | 0.77 | 0.02 | 0.17 | 0.04 | 42 |
| Log Linear Curve | 0.66 | 0.02 | 0.25 | 0.08 | 45 |
| U-Shaped curve | 0.06 | 0.23 | 0.11 | 0.60 | 54 |

B

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| True Dose Response | Bayesian NDLM model | | | | |
| Prob. of early success | Prob. of early failure | Prob. of final success | Prob. of final failure | Mean subjects |
| Placebo like flat curve | 0.00 | 0.01 | 0.12 | 0.87 | 64 |
| *Emax* with EC50=3 | 0.51 | 0.00 | 0.41 | 0.08 | 56 |
| Log Linear Curve | 0.45 | 0.00 | 0.43 | 0.12 | 56 |
| U-Shaped curve | 0.36 | 0.00 | 0.44 | 0.20 | 58 |

Table 6.14 and table 6.15 display the additional operating characteristics of the model fitting to the data that were analysed using the *Emax* model and NDLM model respectively. The mean allocation, proportion of simulation where the dose is selected by ED90/MED/Max, probability being statistical significant compared to placebo are displayed under fully adaptive design setting with each of the four curves being the true effect. It was shown that the Bayesian *Emax* model was able to find the correct dose for ED90 (either 10mg/kg, 20mg/kg or 30mg/kg) when the true response is either an *Emax* curve or log linear curve, however, it cannot find the correct ED90 dose when the true response followed a U-shaped curve. All data show that the NDLM model was more flexible in all four types of model and was able to identify the correct ED90 doses when the true response followed a U-shaped curve. NDLM was able to identify the correct ED90 doses 85% or 91% of the time when the true response is an *Emax* or log linear curve respectively.

Under adaptive design, the simulated data showed similar results to those in fixed design and half adaptive design. For all the simulations, the proportion of simulation being selected as ED90 is 8% vs. 27% using the *Emax* model and NDLM model respectively in analyzing the data when the true model is the *Emax* model with EC50 as 3mg/kg. The true ED90 of the model was 30mg/kg. Adding 20 and 30 mg/kg makes the proportion of simulation being selected as ED90 100% vs. 58% using the *Emax* model and NDLM model respectively. If the trial data has a U-shaped curve, the proportion of simulation being selected as ED90 is 0% vs. 51% using the *Emax* model and NDLM model respectively in analyzing the data when the true ED50 is around 3mg/kg. Other simulation data like the proportion of simulation being selected as ED50 and *Emax* all show a similar trend.

Table 6.14 Comparisons of the operating characteristics of four types of dose response curve using Bayesian *Emax* model in the adaptive design settings. Weak prior was used in the calculation with adaptive allocation. Placebo allocation is fixed at a sample size of 16 (25% of the total sample size).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Dose Level (mg/kg)** | | | | | |
| **0.03** | **0.3** | **3** | **10** | **20** | **30** |
| Mean allocation |  |  |  |  |  |  |
| Flat placebo like curve | 1.3 | 1.3 | 1.3 | 1.6 | 26.4 | 13.8 |
| *Emax* like Curve | 1.2 | 1.3 | 1.4 | 1.6 | 19.6 | 9.8 |
| Log Linear Curve | 1.2 | 1.3 | 1.4 | 1.5 | 21.0 | 11.1 |
| U-Shaped Curve | 1.2 | 1.3 | 1.3 | 1.5 | 26.0 | 12.8 |
|  |  |  |  |  |  |  |
| Proportion of simulation being selected as ED90 |  |  |  |  |  |  |
| Flat placebo like curve | 0% | 0% | 0% | 0% | 36% | 1% |
| *Emax* like Curve | 0% | 0% | 0% | 0% | 92% | 8% |
| Log Linear Curve | 0% | 0% | 0% | 0% | 91% | 8% |
| U-Shaped Curve | 0% | 0% | 0% | 0% | 59% | 0% |
|  |  |  |  |  |  |  |
| Proportion of simulation being selected as MED |  |  |  |  |  |  |
| Flat placebo like curve | 0% | 0% | 0% | 0% | 0% | 0% |
| *Emax* like Curve | 1% | 16% | 47% | 7% | 9% | 0% |
| Log Linear Curve | 0% | 9% | 41% | 7% | 12% | 0% |
| U-Shaped Curve | 1% | 0% | 1% | 0% | 1% | 0% |
|  |  |  |  |  |  |  |
| Proportion of simulation being selected as MAX |  |  |  |  |  |  |
| Flat placebo like curve | 51% | 0% | 0% | 0% | 0% | 50% |
| *Emax* like Curve | 0% | 0% | 0% | 0% | 0% | 100% |
| Log Linear Curve | 0% | 0% | 0% | 0% | 0% | 100% |
| U-Shaped Curve | 31% | 0% | 0% | 0% | 0% | 69% |
|  |  |  |  |  |  |  |
| Probability of dose compared to control |  |  |  |  |  |  |
| Flat placebo like curve | 49% | 49% | 49% | 49% | 49% | 49% |
| *Emax* like Curve | 99% | 99% | 99% | 99% | 99% | 99% |
| Log Linear Curve | 98% | 98% | 98% | 98% | 98% | 98% |
| U-Shaped Curve | 64% | 64% | 64% | 64% | 64% | 64% |
|  |  |  |  |  |  |  |
| Probability of dose compared to control over CSD (0.95) |  |  |  |  |  |  |
| Flat placebo like curve | 0% | 0% | 0% | 0% | 0% | 0% |
| *Emax* like Curve | 4% | 23% | 49% | 66% | 81% | 84% |
| Log Linear Curve | 2% | 16% | 40% | 57% | 74% | 78% |
| U-Shaped Curve | 1% | 3% | 5% | 8% | 14% | 16% |

Table 6.15 Comparisons of the operating characteristics of four types of dose response curve using Bayesian NDLM model in the adaptive design settings. Weak prior was used in the calculation with adaptive allocation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Dose Level (mg/kg)** | | | | | |
| **0.03** | **0.3** | **3** | **10** | **20** | **30** |
| Mean allocation |  |  |  |  |  |  |
| Flat placebo like curve | 10.1 | 9.2 | 7.8 | 7.6 | 8.7 | 9.4 |
| *Emax* like Curve | 6.3 | 5.7 | 6.3 | 7.9 | 9.7 | 10.4 |
| Log Linear Curve | 6.5 | 5.9 | 5.2 | 6.8 | 9.5 | 12.3 |
| U-Shaped Curve | 7.5 | 10.6 | 9.9 | 6.3 | 7.1 | 7.0 |
|  |  |  |  |  |  |  |
| Proportion of simulation being selected as ED90 |  |  |  |  |  |  |
| Flat placebo like curve | 15% | 15% | 16% | 16% | 11% | 11% |
| *Emax* like Curve | 0% | 2% | 14% | 24% | 34% | 27% |
| Log Linear Curve | 1% | 2% | 6% | 16% | 26% | 49% |
| U-Shaped Curve | 1% | 39% | 52% | 6% | 1% | 0% |
|  |  |  |  |  |  |  |
| Proportion of simulation being selected as MED |  |  |  |  |  |  |
| Flat placebo like curve | 0% | 0% | 0% | 0% | 0% | 0% |
| *Emax* like Curve | 17% | 15% | 33% | 18% | 10% | 2% |
| Log Linear Curve | 19% | 15% | 14% | 17% | 14% | 8% |
| U-Shaped Curve | 13% | 50% | 16% | 0% | 0% | 0% |
|  |  |  |  |  |  |  |
| Proportion of simulation being selected as MAX |  |  |  |  |  |  |
| Flat placebo like curve | 16% | 16% | 19% | 19% | 15% | 16% |
| *Emax* like Curve | 0% | 1% | 8% | 17% | 36% | 39% |
| Log Linear Curve | 0% | 1% | 3% | 10% | 22% | 63% |
| U-Shaped Curve | 1% | 31% | 56% | 9% | 3% | 1% |
|  |  |  |  |  |  |  |
| Probability of dose compared to control |  |  |  |  |  |  |
| Flat placebo like curve | 47% | 46% | 45% | 45% | 46% | 47% |
| *Emax* like Curve | 57% | 65% | 81% | 88% | 94% | 95% |
| Log Linear Curve | 55% | 61% | 68% | 79% | 90% | 95% |
| U-Shaped Curve | 58% | 88% | 90% | 73% | 66% | 53% |
|  |  |  |  |  |  |  |
| Probability of dose compared to control over CSD (0.95) |  |  |  |  |  |  |
| Flat placebo like curve | 0% | 0% | 0% | 0% | 0% | 0% |
| *Emax* like Curve | 17% | 26% | 48% | 61% | 71% | 73% |
| Log Linear Curve | 16% | 23% | 31% | 44% | 58% | 73% |
| U-Shaped Curve | 17% | 54% | 61% | 35% | 26% | 16% |

### Simulation results with informative prior

Amongst all the designs, a hybrid approach of half adaptive design with fixed allocation at 50% subjects before any adaptive allocation seems to have the most reasonable operating characteristics and will be considered to carry forward for GSK654321. To further explore the impact of the analysis methods additional simulations were undertaken to examine the impact of choice of informative prior but anchored in the single half adaptive design (Scenario 3). The results for the *Emax* model are given below in Table 6.16.

Table 6.16. Probability of success and failures at interim and final analysis with Bayesian *Emax* model with informative prior (β1~N(-0.5, 1.2\*1.2) β 2 ~N(-2.9, 1.2\*1.2) and β3~N(3, 2\*2) in the half adaptive design (Scenario 3).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| True Dose Response | Bayesian NDLM model | | | | | |
| Early success | Early failure | Final success | Final failure | Total Success | Mean subjects |
| Placebo like flat Curve | **0.00** | **0.13** | **0.08** | **0.79** | **8%**  **8%** | **63.0** |
| *Emax* Curve | **0.87** | **0.00** | **0.13** | **0.00** | **100%**  **100%** | **50.5** |
| Log Linear Curve | **0.71** | **0.00** | **0.28** | **0.01** | **99%**  **99%** | **53.5** |
| U Shaped Curve | **0.18** | **0.00** | **0.41** | **0.40** | **59%**  **59%** | **61.2** |

The probabilities of success, as a measure of posterior probabilities of treatment effect (difference between treatment and placebo) are greater than zero, increased for all dose response curves with 100%, 99%, 59% success if the dose response follows a *Emax* model, Loglinear model and U-shaped curve respectively. The type I error rate is inflated to for 8% in *Emax* model with an informative prior. This inflated type I error rate would need to be communicated to the study team who may consider this to be too high a development risk.

Additional simulations for the NDLM model were performed to examine the impact of informative prior on the half adaptive design (Scenario 3) and are displayed below with two prior choices a) the evolution variance has prior of Inverse Gamma (IG) distribution (IG(0.5,0.5)) and b) IG(2,4).

The additional simulations seem to show that the NDLM model fitting is sensitive to the choice of evolution variance (W) and the probably of success and type I error are impacted by the choice of priors such that with an informative prior, the type I error was reduced to 7% with little impact of the probability of success in other dose response curve. These considerations would also need to be weighed up by the study team. If the Type I error is important then the priors may be further investigated to reduce these to an acceptable level.

Table 6.17 Probability of success and failures at interim and final analysis with Bayesian *NDLM* model with informative prior in the half adaptive design (Scenario 3)

a) Evolution variance ~IG(0.5,0.5), initial dose ~N(-0.5, 1.2\*1.2)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| True Dose Response | Bayesian NDLM model | | | | | |
| Early success | Early failure | Final success | Final failure | Total Success | Mean subjects |
| Placebo like flat Curve | 0.00 | 0.01 | 0.07 | 0.92 | 0.07 | 63.9 |
| *Emax* Curve | 0.71 | 0.00 | 0.28 | 0.11 | 0.99 | 58.0 |
| Log Linear Curve | 0.69 | 0.00 | 0.28 | 0.03 | 0.97 | 57.2 |
| U Shaped Curve | 0.34 | 0.00 | 0.51 | 0.15 | 0.85 | 61.5 |

1. Evolution variance ~IG (2,4), initial dose ~N(-0.5, 1.2\*1.2)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| True Dose Response | Bayesian NDLM model | | | | | |
| Early success | Early failure | Final success | Final failure | Total Success | Mean subjects |
| Placebo like flat curve | 0.00 | 0.00 | 0.10 | 0.90 | 0.10 | 64.0 |
| *Emax* Curve | 0.74 | 0.00 | 0.24 | 0.02 | 0.98 | 57.5 |
| Log Linear Curve | 0.70 | 0.00 | 0.26 | 0.03 | 0.96 | 57.2 |
| U Shaped curve | 0.49 | 0.00 | 0.40 | 0.10 | 0.89 | 59.8 |

### Simulation results with Calibrated type I error rate

If control of the Type I is important then further work could be undertaken to bring this down to the nominal 5%. The type I error rate can be calibrated by adjusting the decision criteria for the placebo flat response to maintain type I error rate at ~5%. The final success criteria of probability of any doses with maximal effect or ED90 effect achieves a drug effect greater than the control was 95% (Pr(|Rdmax–Ctrl| > 0) > 0.95 or Pr(|RED90 –Ctrl| > 0) >95%) and then was adjusted to 0.94 for Emax model and 0.98 for NDLM during type I error rate calibration. The vague prior distribution of evolution variance for NDLM modelis inverse-gamma distribution (IG(0.001,0.001)). The results are given in Table 6.18.

Table 6.18 Probability of success and failures at interim and final analysis after calibrating type I error rate in the half adaptive design (Scenario 3)

A: Bayesian *Emax* model (Type I error rate was calibrated at 5%)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| True Dose Response | Bayesian *Emax* model | | | | | |
| Early success | Early failure | Final success | Final failure | Total Success | Mean subjects |
| Placebo like flat curve | 0.00 | 0.27 | 0.05 | 0.68 | 5% | 60.7 |
| *Emax* curve | 0.79 | 0.00 | 0.20 | 0.01 | 99% | 55.6 |
| Log Linear Curve | 0.66 | 0.00 | 0.31 | 0.03 | 97% | 57.7 |
| U Shaped curve | 0.06 | 0.11 | 0.13 | 0.70 | 19% | 61.9 |

B. Bayesian NDLM (Type I error rate was calibrated at 5%)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| True Dose Response | Bayesian NDLM model | | | | | |
| Early success | Early failure | Final success | Final failure | Total Success | Mean subjects |
| Placebo like flat curve | 0.00 | 0.21 | 0.05 | 0.74 | 5% | 62.8 |
| *Emax* curve | 0.70 | 0.00 | 0.18 | 0.11 | 88% | 57.6 |
| Log Linear Curve | 0.62 | 0.02 | 0.20 | 0.25 | 82% | 57.8 |
| U Shaped curve | 0.53 | 0.00 | 0.21 | 0.17 | 74% | 58.0 |

After type I error calibration, the total probability of success is 99% and 88% if data follow Emax curve; 97% vs. 82% in case of Log Linear curve, 19% vs. 74% in case of U-Shaped curve for Emax and NDLM models respectively. As expected, after calibration, the probability of success decreases for monotonic dose response using NDLM model and continue to perform better if the dose response follows a U-Shaped curve. The results continue to support the conclusion that Bayesian Emax model performs better in case of monotonic dose response and if there is a belief that the dose response could be non-monotonic based on prior knowledge as in our case study - where a compound in the same class seemed to have non-monotonic dose responses - then the NDLM is the superior model to assess the dose response.

In earlier comparisons of the *Emax* and NDLM models, the same decision rules were applied and to assess the type I errors. To facilitate for a fair comparison of power without the need for recalibrating type I error at each design, Receiver Operating Characteristic curves (ROC curves) for the fixed design (S1) and half-adaptive (S3) are presented in Figure 3 and 4 respectively. The ROC curves draw a plot of the true positive rate against the false positive rate for the different possible decision criteria. Since any increase in sensitivity is accompanied by a decrease in specificity, the ROC curves show the tradeoff between sensitivity and specificity. For each design, the true positive rates from Bayesian *Emax* and NDLM model at assumed U-shaped, *Emax* or Loglinear curves are plotted against the corresponding false positive rates from flat curve. The closer the curve follows the left border and the top border of the ROC space, it shows the better sensitivity given specificity.

Under half adaptive design, the ROC curve for the Bayesian *Emax* model is closer to the left and top borders than NDLM model when the assumed curves follow *Emax* or Loglinear curves. Thus. the *Emax* model performs better. When the type I error rate is at 5%, the true positive rate of Bayesian *Emax* model is approximately at 97% for both *Emax* curve and Loglinear curve and the true positive rate is 90% and 85% for both *Emax* curve and loglinear curve using NDLM model. For the U-shaped curve, the Bayesian NDLM model performed better than *Emax* model.

The results are in line with earlier conclusion that *Emax* model outperforms if dose response is monotonic and NDLM model is better when the dose response is U-shaped.

B

A Curve

C

Figure 6.9. ROC curves display the true positive rate (statistical power) and false positive rate for Bayesian *Emax* (red) and NDLM model (blue) under Fixed design (S1) with the dose response following A) U-Shaped, B) Emax or C) Loglinear curve.

A

B

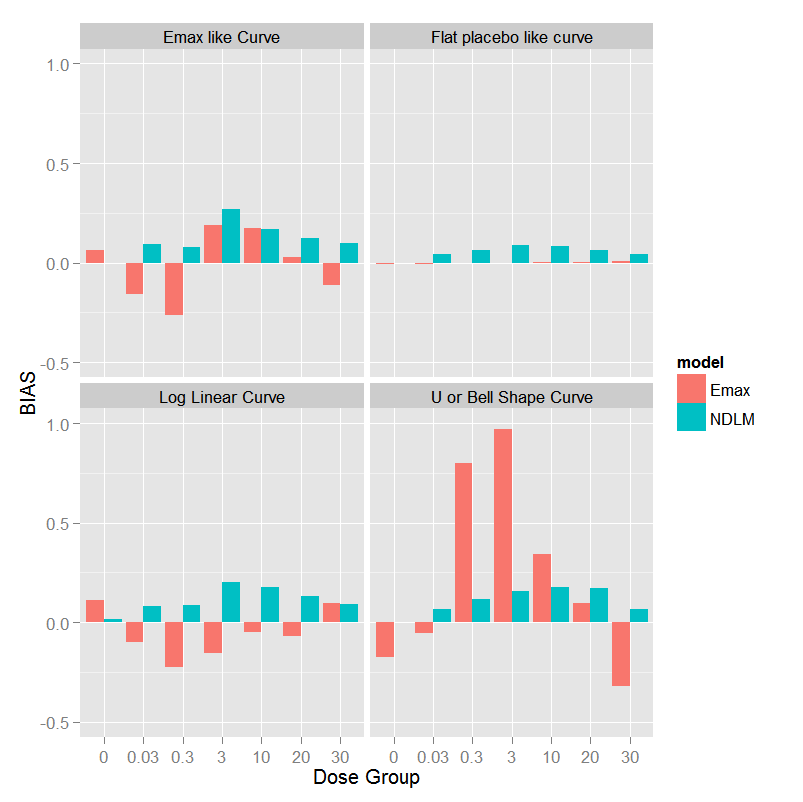
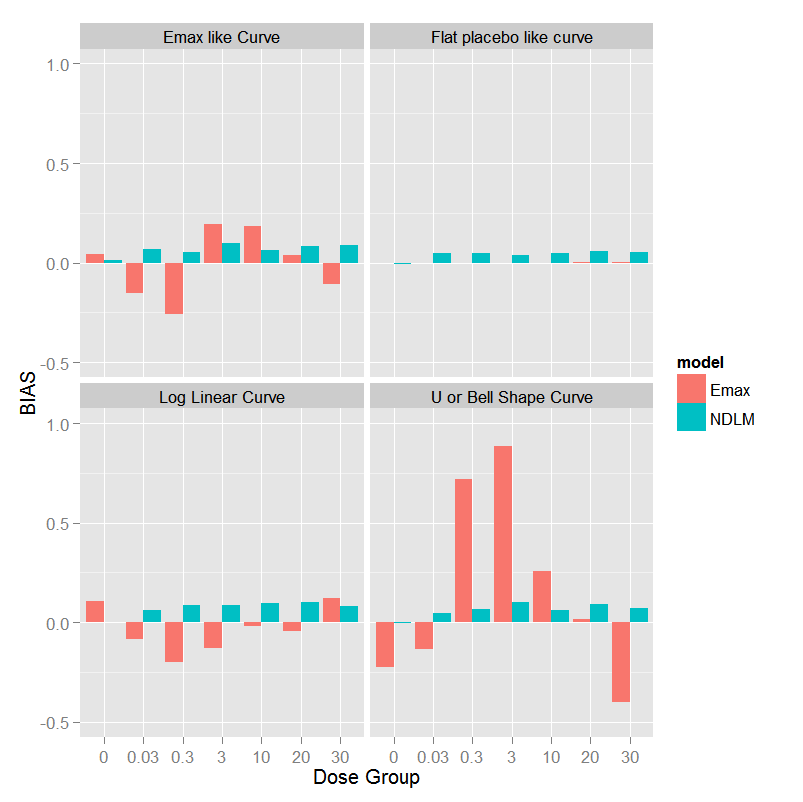
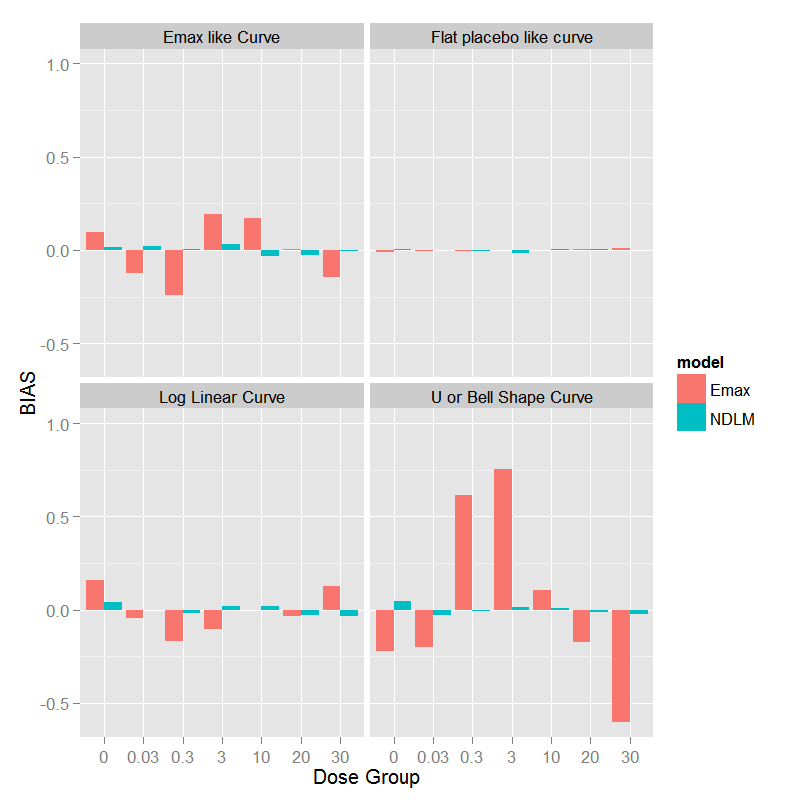
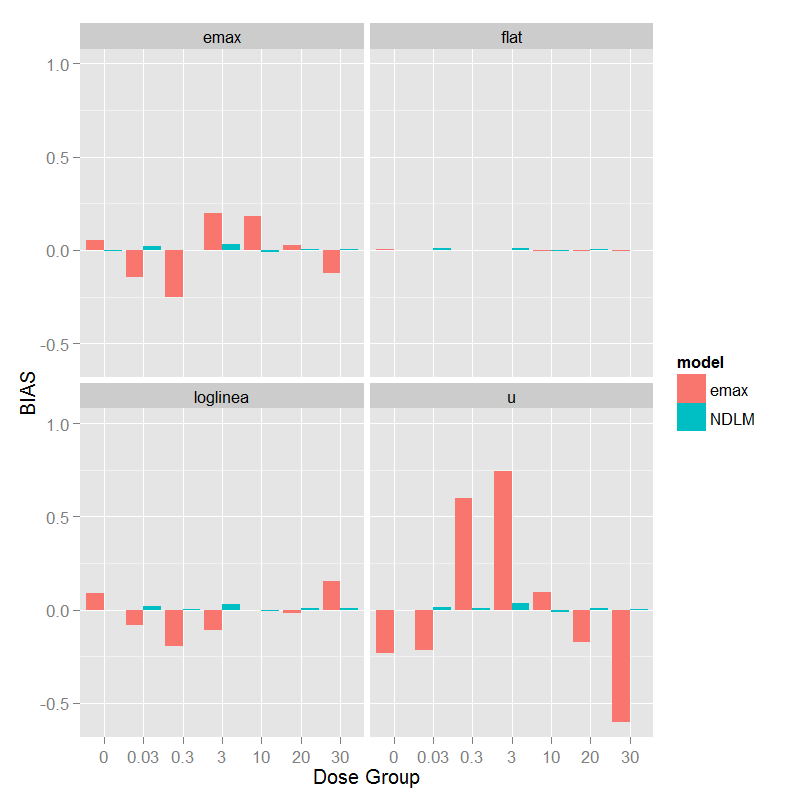
C

Figure 6.10 ROC curve display the true positive rate (statistical power) and false positive rate for Bayesian *Emax* (red) and NDLM model (blue) under Half adaptive design (S3) with the dose response following A) U-Shaped, B) *Emax* or C) Loglinear curves

### Assessment of Bias

The assessment of statistical bias through simulation at each dose level (placebo, 0.03, 0.3, 3, 10, 20, and 30 mg/kg) is calculated as the difference in the estimated mean response using *Emax* or NDLM models against the assumed true response profile (at each dose level). NDLM model is a semi-parametric so there are no “true” model parameter, so the parameter of interest if the true assume response profile with aim to understand the deviation from true response in each simulation scenario. The difference from the true dose response profile is estimated for each simulation. The mean difference - and bias - is taken as the mean difference for the dose response from the truth across all 5000 simulations (Casella, 2001). The Bayesian *Emax* model is compared to the NDLM model under four scenarios of fixed design (Figure 6.11A), adaptive design with fixed allocation (Figure 6.11B), half adaptive (Figure 6.11C) and fully adaptive design (Figure 6.11D). The pre-defined interim futility/success decisions and final decisions were based on the criteria described in Table 6.5.

Under the no adaptive allocation and assumption of true dose response as *Emax* like curve or log linear shape curve, there is less bias (absolute bias) of mean response at lower dose levels using the NDLM model in comparison to the Bayesian *Emax* model. The bias using *Emax* model is less if the true dose response data follow a placebo like response than NDLM model and the absolute values of all bias are less than 0.02. If the true dose response curve is a U shaped non-monotonic curve, the bias is much bigger at 0.3 mg/kg and 3 mg/kg if analyzing using the *Emax* model (0.6510 vs. -0.0062 at 0.3mg/kg in the NDLM model; 0.7523 vs. 0.0155 at 3mg/kg in the NDLM model), since the *Emax* model makes the assumption of monotonic changes and still fits the line between the lowest dose and highest dose, ignoring the U-shaped response.



D

C

A

B

Figure 6.11 The statistical bias for each planned dose group (placebo, 0.03, 3, 10, 20, and 30 mg/kg) under the assumption of the true dose response curve being flat curve, *Emax* like curve, log linear curve, and U shape curve based on the fixed Design with no adaptation. A: Fixed Design; B: No adaptation; C: Half Adaptation; and D: Fully Adaptation

Under the half adaptive allocation design and the assumption of true dose response as *Emax* like curve or log linear shape curve, similar to fixed design, there are less bias (absolute bias) of mean response at lower dose levels but more bias at 20 mg/kg using the NDLM model in comparison to the Bayesian *Emax* model. The individual bias from each dose level shows that *Emax* model tends to underestimate the dose response effect while NDLM tends to overestimate the effect in the mean response. The bias using *Emax* model is less if the true dose response data follow a placebo like response than NDLM model and the absolute values of all bias are less than 0.06. If the true dose response curve is a U-shaped non-monotonic curve, the bias is much bigger at 0.3 mf/kg and 3 mg/kg if analyzing using the *Emax* model (0.7182 vs. 0.0656 at 0.3mg/kg in the NDLM model; 0.8835 vs. 0.0992 at 3mg/kg in the NDLM model), since the *Emax* model makes the assumption of monotonic changes and still fits the line between the lowest dose and highest dose, ignoring the U-shaped response.

Under the fully adaptive allocation design and the assumption of the true dose response as an *Emax* like curve, the bias of the Bayesian *Emax* model and NDLM model is similar. The individual bias from each dose level shows that *Emax* model tends to underestimate the mean response effect at 0.03, 0.3 and 30 mg/kg while NDLM tends to overestimate the effect at 3 and 20 mg/kg in the mean response. The bias is also similar if the true dose response data follow a log linear curve and *Emax* model tends to underestimate the mean response while NDLM tends to overestimate the mean response. NDLM model also overestimate the mean response if the true response is placebo like curve. If the true dose response curve is a U-shaped non-monotonic curve, the bias is much bigger at 0.3 mg/kg and 3 mg/kg if analysing using the *Emax* model (0.8013 vs. 0.1170 at 0.3mg/kg in the NDLM model; 0.9678 vs. 0.1553 at 3mg/kg in the NDLM model), since the *Emax* model makes the assumption of monotonic changes and still fits the line between the lowest dose and highest dose, ignoring the U-shaped response.

### Summary of Model Comparison: *Emax* Model versus NDLM Model

Analysis of the NDLM model led to a significant increase in the statistical power of detecting the treatment difference, when the true dose response is non-monotonic, compared to the Bayesian *Emax* Model. The probability of success using NDLM model was similar regardless of which underlying true dose-response profile was assumed, but less sensitivity in the analysis of selecting the dose response of ED90 and an increase in the statistical bias, compared to the Bayesian *Emax* model. The Bayesian *Emax* model excelled with a higher probability of selecting ED90 and a smaller average sample size, when the true dose response followed *Emax* like curve, compared to NDLM model.

The Type I error is inflated in the Bayesian NDLM model in all scenarios though the type I errors were reduced with the informative prior. The high Type I error could potentially lead to a false investment decision and further work when a compound does not truly have an effect. Once the final study design is established the simulations will need to be investigated with varying decision criteria calibrated so that the Type I error is controlled. Though there were some variations, the bias is comparable if true dose response follows a placebo like curve, *Emax* like curve, or log linear shape curve under the no adaptive allocation, half adaptive and adaptive scenarios. The bias for *Emax* is significantly increased if the true dose response is assumed to follow a U-shaped non-monotonic curve

Analysis of the semi-parametric NDLM model led to a significant increase in the statistical power of detecting the treatment difference, when the true dose response is non-monotonic, compared to the Bayesian *Emax* Model. The probability of success using NDLM model was similar regardless of the true effects models, but less sensitivity in the analysis of selecting the dose response of ED90 and an increase in the statistical bias, compared to the Bayesian *Emax* model. Bayesian *Emax* model excelled with higher probability of selecting ED90 and smaller average sample size, when the true dose response followed *Emax* like curve, compared to NDLM model.

Under the same decision criteria, the Type I error rates are elevated to 12% for half-adaptive or fully adaptive scenario and to 18% for a non-adaptive scenario when analysing using the NDLM model, while the type I error is generally under control below 6% using *Emax* model. An inflated Type I error rate signals that the NDLM model is over-sensitive and is thus inflating the number of false positive trials. When controlling Type I error, it was shown from ROC curves that the statistical power is 8-10% lowers in NDLM model if the dose response follows *Emax* or Loglinear curves but much better in case of U-shaped curve. Analysis of the NDLM model led to a significant increase in the statistical power of detecting the treatment difference, when the true dose response is non-monotonic, compared to the Bayesian *Emax* Model. The probability of success using NDLM model was similar regardless of which underlying true dose-response profile was assumed, but less sensitivity in the analysis of selecting the dose response of ED90 and an increase in the statistical bias, compared to the Bayesian *Emax* model. The Bayesian *Emax* model was superior with a higher probability of selecting ED90 and a smaller average sample size, when the true dose response followed *Emax* like curve, compared to NDLM model.

Amongst all the design scenarios, a hybrid approach of half adaptive design with fixed allocation at 50% subjects before any adaptive allocation seems to have the most reasonable operating characteristics and will be considered to carry forward for GSK654321.

Though there were some variations, the bias is comparable if true dose response follows placebo like curve, *Emax* like curve, or log linear shape curve under the no adaptive allocation, half adaptive and adaptive scenarios. The bias is significantly increased if true dose response follows a U-shaped non-monotonic curve.

## Discussion

We have described and compared the statistical properties of the Bayesian *Emax* model and the Bayesian NDLM model for a Phase 2a PoC study for a new compound to treat patients with RA. We have then evaluated both models using simulation in the context of an adaptive Phase 2a PoC design for a new RA compound under a variety of assumed dose response curves: linear, *Emax* model, U-shaped curve, and flat. We have demonstrated that the NDLM model is more flexible and can handle a wide variety of dose-responses, including monotonic and non-monotonic relationships. The *Emax* model was superior, compared to NDLM model, with higher probability of selecting the ED90 and smaller average sample size, when the assumed underlying true dose response follows an *Emax* like curve.

Although the *Emax* model is one of the most widely applied models relating drug concentrations to effects (Pinheiro, 2006), it is not uncommon to observed non-monotonic U-shaped dose response in biological cell and organism (Calabresa, 2014), especially in Phase 2a clinical development of novel biological in the treatment of RA patients (Stohl, 2013 and Owyang, 2010).

Due the fact that the results for a PoC RA study of a drug in the same class followed a U-shaped dose response there was a wish to investigate if the analysis could be improved for a new compound in development. Of particular interest, in context with the development for GSK654321, the NDLM model was able to maintain the probability of success even in the case of a non-monotonic dose response.

In 2007, a Pharma working group on dose response published a White Paper to compare the Bayesian NDLM and other designs or dose response models including optimal design, MCP‑MOD, ANOVA, smoothing spline (Bronkamp, 2007) etc. In the context of the RA clinical trial in this Chapter, a Bayesian *Emax* model was used in the original PoC design and analysis of GSK123456 and a NDLM model was proposed for the analysis of a U-shaped dose response. It was shown (Bronkamp, 2007) that the NDLM model outperformed others models i.e. MCP-MOD model with regard to the probability of identifying the dose response and the correct dose, therefore the NDLM was chosen in the comparisons with *Emax* model. Our simulation results extend that into the context of with much smaller sample size (total N=64, n=8 each arm) and the results match with the results from previous publications.

The NDLM model is a good alternative to the *Emax* model for a RA Phase 2a PoC design. The main cause for concern with NDLM is the inflation of the Type I error. It is demonstrated from the ROC curves that the power of *Emax* model is superior than that of the NDLM model but the difference is quite negligible for *Emax* and Loglinear curves. To mitigate the risk of Type I error inflation, the decision criteria may need to be adjusted to control the Type I error by adopting different decision rules. Further work is would be required therefore for any individual study to optimise the design characteristics. It is also acknowledged that NDLM model did not have high specificity in finding ED90 compared with the *Emax* model when the data follow *Emax* model.

*Emax* model is one of the best model to estimate the dose response if the dose response followed monotonic curve, however, while biological exposure response relationships are often monotonic, down-turns of the clinical dose-response relationship at higher doses have been observed, one example in biologics development is the immunogenicity observed at high dose in the patients treating with biological. Therefore, we recommend to routinely considering an umbrella-shaped dose-response model unless umbrella profiles can be excluded with certainty at the trial design stage.

If there is a chance that the dose response shall follow non-monotonic increase or decrease NDLM is a better model, the thorough operating characteristics are needed to run to understand the model performance the benefits and the concerns of using the non-monotonic model should be communicated and highlighted to clinical team.

In comparison to Temple’s research on NDLM and *Emax* comparison within the context of Phase 2b trial with larger sample size (n~ 30 per arm, total 250-280), this work in this dissertation investigated the comparison of both NDLM and *Emax* with the context of a Phase 2a trial – with smaller sample sizes and a wide dose range. It was shown that *Emax* model was a better model, compared to NDLM model, with higher probability of selecting the ED90 and smaller average sample size, when the assumed underlying true dose response follows an *Emax* like curve. When the dose responses follow a non-monotonic dose response such as U-shaped curve, NDLM method has better allocation and power in correctly identifying a dose more often than Bayesian *Emax* model.

It was also shown in the research of Temple that Bayesian NDLM tended to underestimate the response at the early doses, therefore resulting in higher doses being selected, however, our simulation showed a similar or better model fitting in Bayesian NDLM model than *Emax* model within the context of Phase 2a setting. In addition, we found out that an adaptive design performed better with smaller average sample size but there was little difference in different allocation methods using NDLM model, which agree with the finding in Temple (2012).

It should be noted that the methods described in this dissertation were anchored in a single RA example with the simulations and results presented only applicable to this case study which motivated our work. This is of particular importance if different dose responses are anticipated or are of importance for an evaluation. Even for this case study there would be a need for further work once the study design has been finalised. Despite the apparent lack of generalisability of the results the methods proposed and our methods of evaluation could be generalised to other clinical trials to offer a solution to expedite drug development.

Therefore, it is plausible to detect non-monotonical dose response in the dose response analysis and this dissertation showed how to compare the statistical properties of the Bayesian Emax model and the NDLM model*.* The NDLM model is shown to be more flexible and can handle a wide variety of dose-responses, including monotonic and non-monotonic relationships although, within the parameters of the simulation.

The main cause for concern with NDLM was the inflation of the Type I error. To minimise this problem, the decision criteria or informative prior may need to be adjusted to control the Type I error if the same decision rules are used in the comparison. After controlling for the type I error rate at 5%, the statistical powers of *Emax* model are ~8% higher than that of NDLM models in *Emax* and Log-linear dose responses, which was further supported by ROC results. The NDLM model works better when dose response follows U-shaped curve. Further work is would be required therefore for any individual study to optimise the design characteristics. It is also acknowledged that NDLM model did not have high specificity in finding ED90 compared with the *Emax* model when the data follow *Emax* model.

It should be noted that the methods described in this paper were anchored in a single RA example with the simulations and results presented only applicable to this case study which motivated our work. This is of particular importance if different dose responses are anticipated or are of importance for an evaluation. Even for this case study there would be a need for further work once the study design has been finalised. In case that a U-shaped curve us expected or there is potential physiological/pharmacological rationale of down-turn response, Bayesian NDLM model is generally recommended and this conclusion can be generalized to other case studies. In addition, our methods of evaluation in finding the best design could be generalised to other clinical trials to offer a solution to expedite drug development.

Many published clinical trials use the data from interim analyses when the subjects complete the study, or at least reach the primary endpoint, for adaption or decision making purposes. RA disease is a chronic disease area, the onset duration of which is long. From patient enrolment to the time of the clinical outcome can take months. This posted a challenge to adaptive designs.

In the next chapter, a literature search is conducted and a meta-analysis is performed to explore the relationship between early and late time points in the same study. If a late response can be predicted using early response, adaption based on predicted late time point from an early time point could lead to a potential time and resource saving.

# Improving the RA proof of concept desigN: Relationship of early response and late response

In previous chapters, the Phase 2a adaptive design with dose response model was implemented in a PoC study. A revised dose response model based on a Bayesian NDLM model was proposed to fit the dose response and to give a good estimate of ED90, especially in a situation where the dose response increases or decreases non-monotonically.

The merits of an adaptive design were discussed in Chapters 3 and 4. An adaptive design with pre-specified stopping criteria allows researchers to stop a failing compound early if the drug does not show the anticipated efficacy or, conversely, a development can be expedited, potentially leading to a new drug application earlier if there is an overwhelming drug effect.

Although there are a number merits to adaptive designs, there were two issues identified with the implementation of adaptive design in the Phase 2a PoC trial.

The first issue was that the trial failed to predict the ED90 (estimate from Bayesian *Emax* model) when there was a non-monotonic dose response. In Chapter 6, we compared the *Emax* model with the alternative semi-parametric model: the NDLM model. It was shown that the semi-parametric NDLM model was able to identify the ED90 efficiently when data followed a U-shaped curve and maintained a similar probability of success in the adaptive design setting when the dose response decreases monotonically. It was concluded that the NDLM model is a flexible and alternative model to investigate the dose response of a new drug when the dose response is not known or fully explored.

The second issue was that the time required for the interim analysis and adaptation was too long in the context of the planned study design. RA is a chronic disease, the development of which usually takes years. Due to the nature of the disease, the primary endpoint is usually measured months after treatment. For example, in the Phase 2a trial, the primary clinical endpoint is the DAS28 change from baseline at Day 56 (Month 2) as a result, the part A cohort randomization design took more than 2 years to complete the first part of the study (Part A) due to the time taken to complete the 7 planned between cohort interim analysis. Slow enrolment and a longer time to complete the adaptations posed a challenge to clinical development.

It is estimated that the average annual revenue of biological medication for RA disease globally is $2.5 billion for the top 7 RA drugs in the market in 2012 (Palmer, 2013) and the sale of RA drugs is expected to increase greatly in the next 3-5 years (Palmer, 2013). The investment impact is huge if the decision-making of Phase 2a trials can be improved and the time can be shortened. There is thus an urgent need to expedite the Phase 2a decision-making.

One possible solution to the challenge is to predict the response at a late time point using an early time point that would facilitate the interim analysis being conducted earlier in the study and as a consequence make the study duration shorter. The time can be significantly reduced if the interim decision and adaption are based on the early time point. To predict the late response, it is necessary to investigate the relationship between early and late response. In this chapter, we will explore the relationship between early endpoint and late endpoint by analyzing the PoC study data retrospectively. We will also perform a systematic meta-analysis of published studies.

## The Aims of the Chapter

The aims of this chapter are to

1. Explore the relationship of DAS28 at the early time point and later time point using single trial data from the PoC trial of GSK123456;
2. Perform a systemic literature review;
3. Investigate the relationship of DAS28 between the early and later time point.
4. Extend the investigation to the relationship of DAS28 change from baseline between the early and later time point;
5. Review potential structural correlations in correlating DAS28 changes from baseline and DAS28 change at Day 14 as surrogate endpoint of Day 56.

## Relationship of DAS28 at Early Time Point and Late Time Point in the PoC Study

DAS28 is usually collected at multiple time points at screening, baseline, and across the span of post-dose periods in a typical PoC trial (Genovese, 2010; Burmester, 2012). The baseline is defined as the last available assessment prior to the start of the dose. Therefore, the change from baseline at post-dose will be calculated as the baseline assessment subtracted from post-dose assessments.

The Phase 2a PoC data have been discussed in Chapter 5 and 6. Individual data of DAS28 change from baseline at early time point (Day 14) are plotted against later time points (Day 56 vs. Day 14) in Figure 7.1. A strong linear correlation was found in DAS28 change from baseline between early time point at Day 14 and late time point at Day 56 (Figure 7.1). The adjusted correlation coefficient of the line relationship of DAS28 at Day 14 vs. Day 56 is 0.73. That means that the mean DAS28 score at the later time point is likely to have a higher response if the early time point showed a higher response. The findings imply that an adaptive clinical trial based on DAS28 response at Day 14 instead of Day 56 could potentially be undertaken. Using Day 14 would reduce the decision per cohort by 42 days. The findings may provide a path to reducing the time to decision making using early time point (DAS28 at Day 14) to predict late time point (Day 56 or Day 84), to inform the faster decision-making.

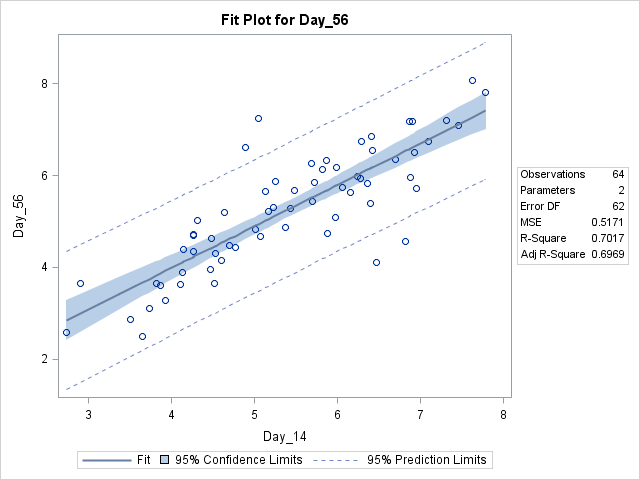


Figure 7.1. Scatter plot and linear model fitting of DAS28 score at Day 56 and Day 14. Both model fitting with 95% confidence limit and 95% prediction limit using linear model are presented in the plot. 64 subjects in Part A of the PoC trial combining both treatment and placebo are included in the graph and the r-square of the line relationship of DAS28 at Day 56 vs. Day 14 is 0.73

The individual data from Day 7, Day 14, Day 28, Day 56 and Day 84 are displayed in matrix plots to investigate the relationship of paired time points in DAS28 change response (all data combined in Figure 7.2, all subjects who received placebo only in Figure 7.3, and all subjects who received GSK123456 in Figure 7.4). The corresponding correlation coefficient, P‑values and number of subjects between DAS28 at Day 7, Day 14, Day 28, Day 56 and Day 84 are presented in Tables 7.1, 7.2, and 7.3.

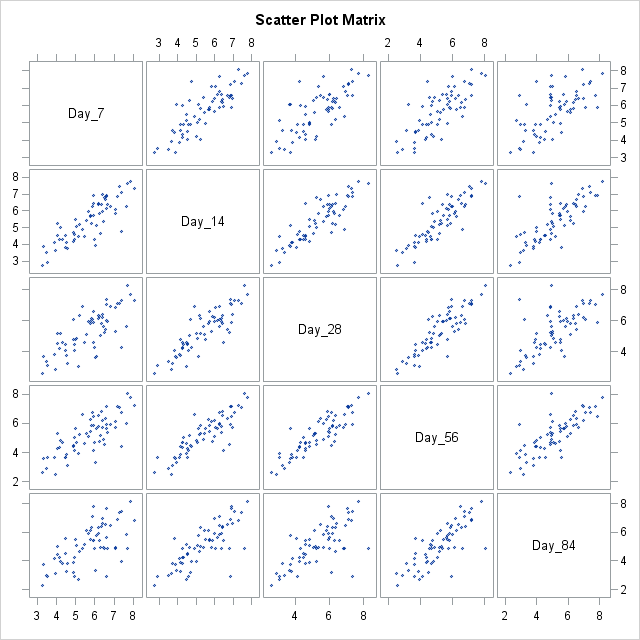


Figure 7.2 Individual correlation scatter matrix plot of DAS28 score at Day 7, Day 14, Day 28, Day 56 and Day 84 (from left to right), all subjects in both the treatment and placebo groups in Phase 2a Part A of GSK123456 are included in the plots.

Table 7.1 Correlation coefficient, P-value and number of subjects between DAS28 score at Day 7, Day 14, Day 28, Day 56 and Day 84 (from left to right). The analysis includes all subjects receiving both study drug and placebo in Phase 2a Part A of GSK123456.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Pearson Correlation Coefficient of DAS28 (Part A All Patients)** | | | | | |
|  | **Day 7** | **Day 14** | **Day 28** | **Day 56** | **Day 84** |
| **Day 7** | 1.00  63 |  |  |  |  |
| **Day 14** | 0.83 <.0001 63 | 1.00  64 |  |  |  |
| **Day 28** | 0.74 <.0001 62 | 0.86 <.0001 63 | 1.00  63 |  |  |
| **Day 56** | 0.75 <.0001 63 | 0.84 <.0001 64 | 0.88 <.0001 63 | 1.00  64 |  |
| **Day 84** | 0.64 <.0001 62 | 0.77 <.0001 63 | 0.68 <.0001 63 | 0.79 <.0001 63 | 1.00  63 |

A total of 64 patients from Part A of the GSK123456 PoC study were used to calculate the between day correlations and P-values in Table 7.1. The correlations between DAS28 scores at Day 14 and Days 28, 56 and 84 are 0.77, 0.73, and 0.63 respectively. There appears to be a strong association/correlation between DAS28 score on Day 14 and DAS28 scores on Day 28, 56 and 84 in the RA population combining treatments and placebo.

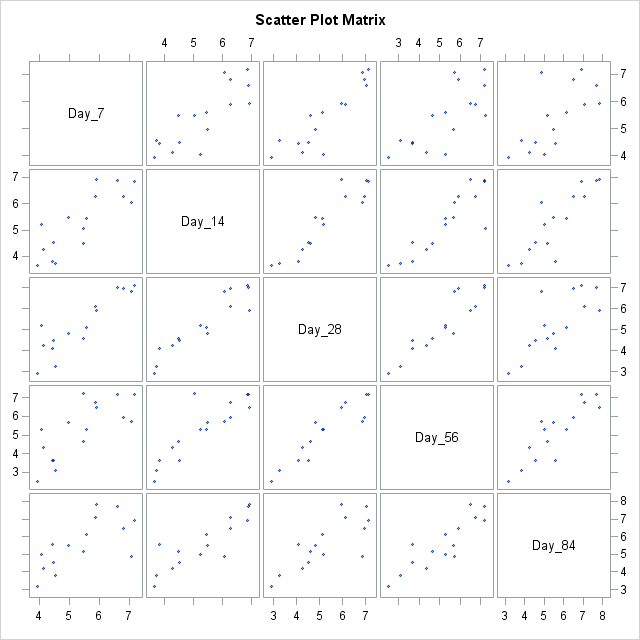


Figure 7.3 Individual correlation scatter plot of DAS28 score at Day 7, Day 14, Day 28, Day 56 and Day 84 (from left to right), placebo subjects in Phase 2a Part A of GSK123456 are included in the plots.

Table 7.2 Correlation coefficient, P-value and number of subjects between DAS28 score at Day 7, Day 14, Day 28, Day 56 and Day 84 (from left to right). The analysis includes placebo subjects in Phase 2a Part A of GSK123456.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Pearson Correlation Coefficient of DAS28 (Part A Placebo)** | | | | | |
|  | **Day 7** | **Day 14** | **Day 28** | **Day 56** | **Day 84** |
| **Day 7** | 1.00  16 |  |  |  |  |
| **Day 14** | 0.83 0.0061 16 | 1.00  16 |  |  |  |
| **Day 28** | 0.90 <.0001 15 | 0.93 <.0001 15 | 1.00  15 |  |  |
| **Day 56** | 0.75 0.0009 16 | 0.88 <.0001 16 | 0.92 <.0001 15 | 1.00  16 |  |
| **Day 84** | 0.69 0.0041 15 | 0.87 <.0001 15 | 0.78 0.0005 15 | 0.88 <.0001 15 | 1.00  15 |

A minimum of 15 placebo patients from Part A of the GSK123456 PoC study were used to calculate the between day correlations and P-values (Table 7.2). Larger correlation coefficients were observed between Day 14 vs. Day 28 (0.87) than between Day 14 vs. Day 56 (0.80) or Day 84 (0.66) which indicates correlations are not the same across days. Day 14 and Day 28 seem to be the best predictors of Day 56 (correlations are 0.80 and 0.84) and Day 14 is the best early time point to predict Day 84 (correlation is 0.66). Similar results have been observed in the RA patients receiving treatment (Table 7.3).

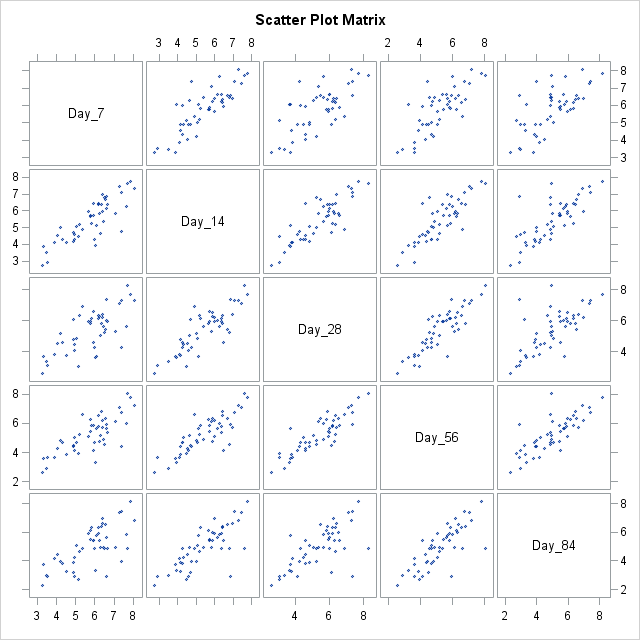


Figure 7.4 Individual correlation scatter plots of DAS28 score at Day 7, Day 14, Day 28, Day 56 and Day 84 (from left to right), all subjects receiving GSK1123456 in Phase 2a Part A of GSK123456 are included in the plots.

Table 7.3 Correlation coefficient, P-value and number of subjects between DAS28 score at Day 7, Day 14, Day 28, Day 56 and Day 84 (from left to right). The analysis includes subjects in all GSK123456 dose levels in Phase 2a Part A of GSK123456.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Pearson Correlation Coefficient of DAS28 (Part A Treatment combined GSK123456)** | | | | | |
|  | **Day 7** | **Day 14** | **Day 28** | **Day 56** | **Day 84** |
| **Day 7** | 1.00  47 |  |  |  |  |
| **Day 14** | 0.84 <.0001 47 | 1.00  48 |  |  |  |
| **Day 28** | 0.70 0.0007 47 | 0.84 <.0001 48 | 1.00  48 |  |  |
| **Day 56** | 0.77 <.0001 47 | 0.83 <.0001 48 | 0.87 <.0001 48 | 1.00  48 |  |
| **Day 84** | 0.68 0.0009 47 | 0.77 <.0001 48 | 0.68 0.0001 48 | 0.79 <.0001 48 | 1.00  48 |

The relationship of DAS28 and other RA biomarkers has also been explored to investigate whether DAS28 can be predicted using an early biomarkers other than DAS28. The RA inflammation biomarkers, including CRP and ESR, follow a similar trend (data not shown). Both biomarkers are components of DAS28 composite endpoints. The correlation of DAS28 data and other RA inflammation biomarker responses were examined. However, no strong correlation could be established.

### Summary of Late and Early Time Point Relationship

Linear correlations have shown that DAS28 at the early and late time points in the subjects who received placebo, GSK123456, and all data combined were well correlated. The closer the time between the early and late time point, the stronger the observed correlations. It is concluded from the RA trial that DAS28 at early time point (i.e. Day 14) is the best predictor of the late time point (i.e. Month 2 Day 56).

To further understand the relationship in DAS28 between early and late endpoints and generalize the relationship from a single trial to all RA clinical trials, in the next section of this chapter, a systematic literature reviews and a meta-analysis were conducted of published trials to investigate the relationship in DAS28 scores between early and late timepoints.

## Meta-analysis to investigate the Relationship of Early Time Point and Late Time Point

An analysis of a single PoC trial (Section 7.2) has shown that DAS28 response at late time point was correlated to early time point. However, single studies are often not reliable enough to generalize the results; therefore, a meta-analysis from multiple studies is performed in this section to explore whether the linear relationship can be generalized to all other studies in this section. The meta-analysis combines results from published studies to investigate the relationships and address the pre-defined hypotheses, questions or patterns across studies (Greenland 2008).

In this section, all literature that meets the criteria is collated for the meta-analysis to investigate whether the linear relationship of early and late time point from PoC data can be extended to studies in general in patients with RA disease.

### Identification of Trials

The first meta-analysis was conducted in 2012 and the literature between 2012 and 2015 was reviewed in 2015 and combined with first meta-analysis. Twenty eligible RCTs were combined which include DAS28 related terms as listed below.

MEDLINE and Cochrane Central Register of Controlled Trials databases were searched from 2001 to Feb 2015 to identify all eligible randomized, controlled trials (RCTs). Details of each trial included in the analysis are given in Appendix. The aim was to identify all published trials that were available from 2001 to Feb 2015.

Only the literature that has both early (Week2, Month 1) and late (Month 2, Month 3, and Month 6) DAS28 data (DAS28-CRP or DAS28 ESR) are used. The search terms for the systematic review included "DAS28", "rheumatoid arthritis", “disease activity”, “double blind controlled”, and "randomized clinical trial". Only trials that met the following criteria were included:

1) Results reported in publicly available sources including publication;

2) Had treatment durations of at least 2 months;

3) Placebo/comparator controlled trials, with double blinded randomization;

4) Clinical trials reported only in abstract form were not included;

5) All the eligible clinical trials must have at least a baseline and 8-week measurement of DAS28, so the early time point and late time point response can be calculated. In addition, the trials must have a study duration of at least 8 weeks of post-treatment. If only DAS28 and baseline DAS28 are captured in the literature, the DAS28 change will be calculated by subtracting DAS28 baseline from the DAS28 at pre-defined days;

6) Non-drug related trials, such as acupuncture or physical therapy trials were not included;

7) Full access to the complete literature was available;

8) If there are multiple reports for the same compound, a careful review of literature was performed to cover multiple phases of clinical trials and to minimize duplication as well domination by a single drug.

### List of Literature Included in the Meta-Analysis

The complete list of literature that meets the criteria described in Section 7.3.1 is shown in Appendix 11.3. All the eligible literature is included in the references of this dissertation. Since the correlation may be different in different study treatments, to facilitate analysis and explore the between-treatment variations in the meta-analysis, the study treatments are categorised into three categories: 1: Placebo/MTX; 2: Biological with or without MTX/DMARDs; 3: Other RA Treatments.

There are a total of 71 sets of eligible treatment data from 20 Phase 2 and Phase 3 clinical trials which had both early and late time points included in the meta-analysis. Not all the treatment studies had all the data.

## Relationship of DAS28 at Early and Late Time Point based on Meta-Analysis

The mean DAS28 change data from all the eligible literature are pooled and listed by the time point post-treatment. The baseline data such as baseline DAS28, the disease duration and health assessment questionnaire disability index (HAQ-DI) are recorded. The data from the following time point are selected for analysis since they are closely related to the time point in the PoC study: Day 14 (Week 2), Month 1, Month 2, and Month 3. Data across all the treatment are combined for the meta-analysis to explore the relation of DAS28 change from baseline presented in the literature.

The objective of the analysis is to explore the relationship of early results and late results with a focus on building a model for future data prediction in RA adaptive design. In this section, we start with a weighted linear regression with trial sample size as a weight and treating all the clinical trial data as the same regardless the study treatment. Correlations between the DAS28 at early time point and late time point are calculated separately by each treatment group. Finally, ANCOVA model with study treatment group as fixed effects and early time point as covariate is used to fit the meta-analysis data.

### Correlation and Simple Linear Regression

Figure 7.5 and Figure 7.6 are matrix plots of the DAS28 at early time point (x-axis) and at late time point (y-axis) from the meta-analysis. Similar positive correlations are found in DAS28 change between the early time point and late time. The mean of DAS28 change data at Day 14 (Week 2), Day 28 (Month1), Day 56 (Month 2), Day 84 (Month 3), and Day 168 (Month 6) from each study is presented with all treatments combined in Figure 7.5 and biological presented in Figure 7.6. The correlation coefficients for all treatments combined and biological are presented in Table 7.4 and Table 7.5 respectively. The correlation was calculated using the mean of each treatment without adjustment for sample size.

The correlations between Day 14 and Day 56 are 0.89 and 0.88 in all study treatments combined and biological respectively. It is also shown that, as time point gets closer, the R square is relatively higher which indicates that adjacent early time points are a better predictor of late time points. Day 14 appears to be a good predictor of Day 56 in the RA population.

### Summary of Late and Early Time Point Relationship

It is evident that there are significant correlations in mean DAS28 change between the mean of early time points (i.e. Day 14) and the mean of late time points (i.e. Day 56), which further indicate how predicting late clinical response using short-term data can be undertaken (two arms were removed due to baseline not available, two data points were removed due to outliers Cook D value).

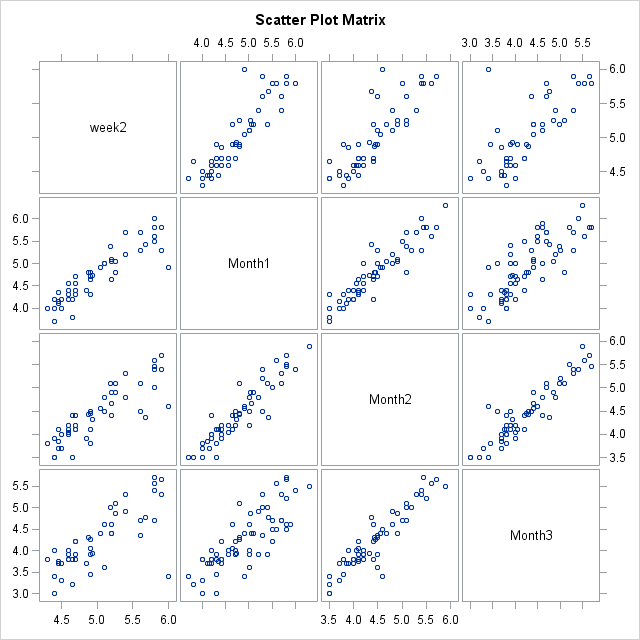


Figure 7.5 Matrix plot from the meta-analysis. All time points of DAS28 at Week 2, Month 1, Month 2, and Month 3 from the meta-analysis are displayed and all study treatments are combined. Each point is the mean from each treatment of clinical trials in the meta-analysis.

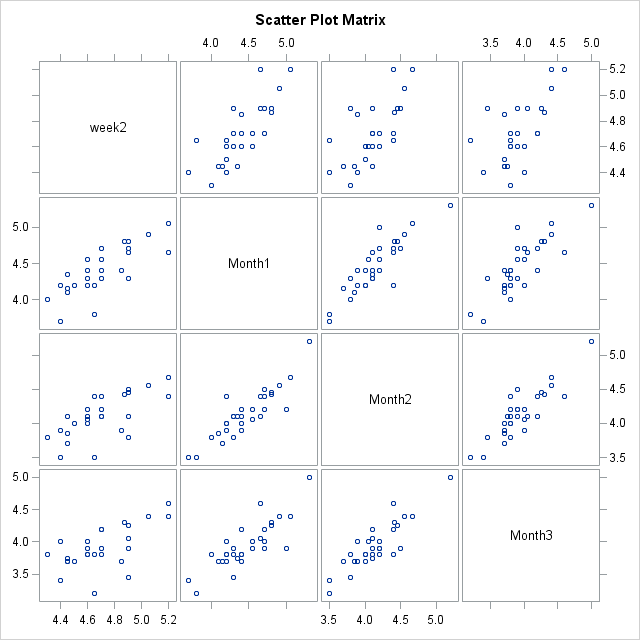


Figure 7.6 Scatter matrix plots of all time points of DAS28 at Week 2, Month 1, Month 2, and Month 3 from the meta-analysis with study treatment of biological. Each point is the mean from each clinical trial in the meta-analysis.

Table 7.4 Correlation coefficient, P-value and number of treatment arms between DAS28 at Week 2, Month 1, Month 2, and Month 3 from the meta-analysis combining all study treatments.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Week 2** | **Month 1** | **Month 2** | **Month 3** |
| **Week 2** | |  | | --- | | 1.00 | |  | | 46 | |  |  |  |
| **Month 1** | |  | | --- | | 0.93 | | <.0001 | | 46 | | |  | | --- | | 1.00 | |  | | 64 | |  |  |
| **Month 2** | |  | | --- | | 0.88 | | <.0001 | | 46 | | |  | | --- | | 0.93 | | <.0001 | | 53 | | |  | | --- | | 1.00 | |  | | 58 | |  |
| **Month 3** | |  | | --- | | 0.86 | | <.0001 | | 46 | | |  | | --- | | 0.83 | | <.0001 | | 64 | | |  | | --- | | 0.95 | | <.0001 | | 55 | | |  | | --- | | 1.00 | |  | | 66 | |

Table 7.5 Correlation coefficient, P-value and number of treatment arms between DAS28 at Week 2, Month 1, Month 2, and Month 3 from the meta-analysis with study treatment of biological.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Week 2** | **Month 1** | **Month 2** | **Month 3** |
| **Week 2** | |  | | --- | | 1.00 | |  | | 26 | |  |  |  |
| **Month 1** | |  | | --- | | 0.78 | | <.0001 | | 26 | | |  | | --- | | 1.00 | |  | | 30 | |  |  |
| **Month 2** | |  | | --- | | 0.68 | | <.0001 | | 26 | | |  | | --- | | 0.88 | | <.0001 | | 34 | | |  | | --- | | 1.00 | |  | | 30 | |  |
| **Month 3** | |  | | --- | | 0.62 | | 0.0008 | | 26 | | |  | | --- | | 0.83 | | <.0001 | | 34 | | |  | | --- | | 0.87 | | <.0001 | | 30 | | |  | | --- | | 1.00 | |  | | 30 | |

## Relationship of early response and late response in change from baseline

In previous sections, Pearson Correlation Coefficients are 0.83-0.84in individual Phase 2a trial and 0.68-0.88 for meta-analysis, which indicated that there are strong correlations in DAS28 raw scores between early response i.e. Day 14 and late response i.e. Day 56. One thing to highlight is that the primary endpoint of the Phase 2a study of GSK654321 is DAS28 change from baseline, therefore it is necessary to predicate the DAS28 change from baseline at Day 56 using early response at Day 14. In this section, the correlations between early and late response in change from baseline will be calculated and compared with those in raw scores. In addition, an important issue to consider is that with change from baseline a structural correlation is introduced, since baseline is involved in the calculation of change from baseline in both the responses at Day 14 and at Day 56. The structural correlation is then discussed in the next Section.

### DAS28 Change at Early Time Point and Late Time Point in PoC Study

To explore the relationship of endpoint to disease activity at multiple time points, individual data of DAS28 change from baseline at early time point (Day 14) are plotted against later time points (Day 56 vs. Day 14) in Figure 7.7. A strong linear correlation was found in DAS28 change from baseline between early time point at Day 14 and late time point at Day 56 (Figure 7.7). The adjusted correlation coefficient of the line relationship of DAS28 at Day 14 vs. Day 56 is 0.73. That means that the mean DAS28 at the later time point is likely to have a higher response if the early time point showed a higher response. The findings imply that an adaptive clinical trial based on DAS28 response at Day 14 instead of Day 56 could potentially be undertaken. Using Day 14 would reduce the decision per cohort by 42 days. The findings may provide a path to reducing the time to decision making using early time point (DAS28 at Day 14) to predict late time point (Day 56 or Day 84), to inform the faster decision-making.

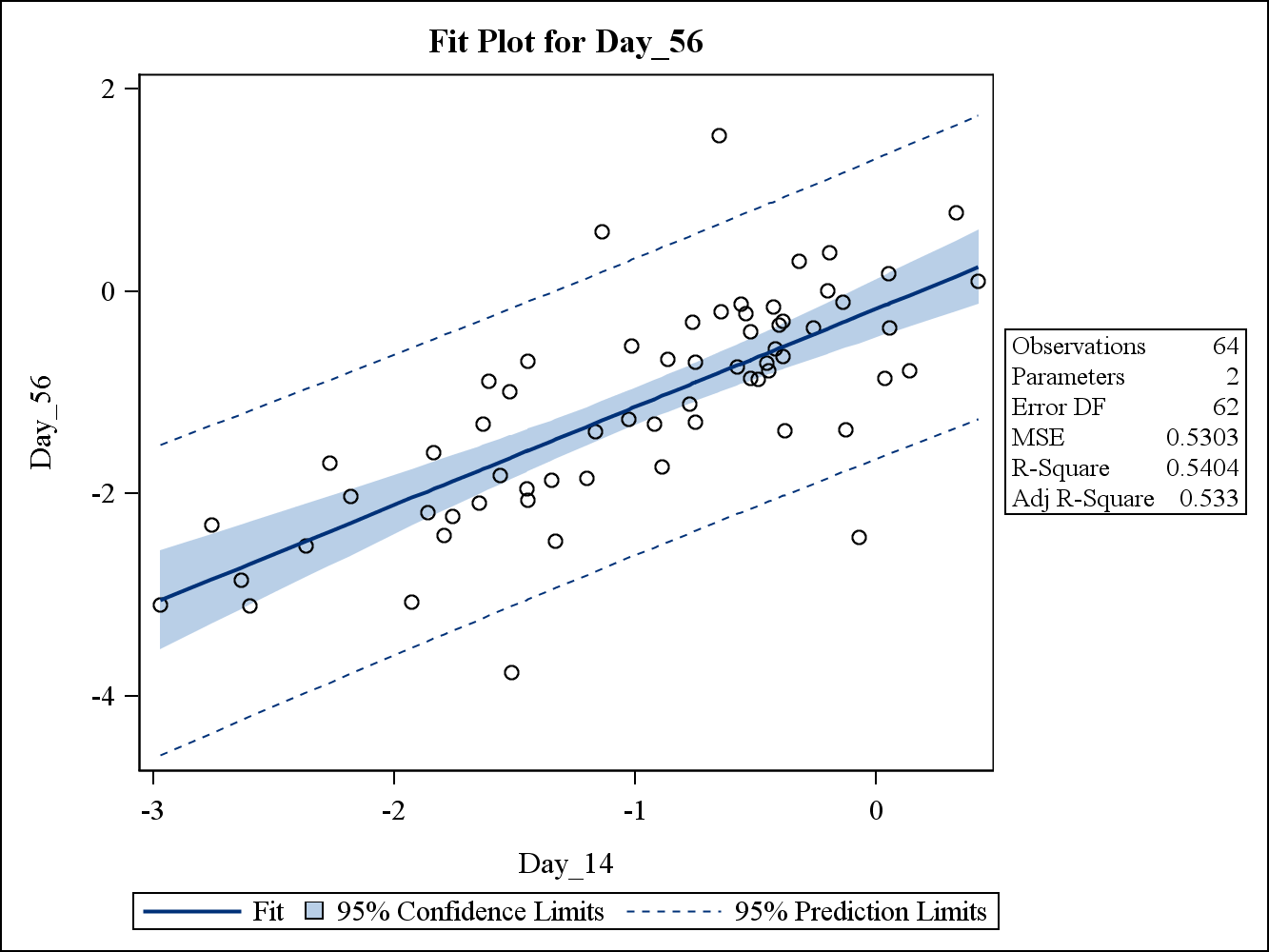


Figure 7.7. Scatter plot and linear model fitting of DAS28 change from baseline at Day 56 and Day 14. Both model fitting with 95% confidence limit and 95% prediction limit using linear model are presented in the plot. 64 subjects in Part A of the PoC trial combining both treatment and placebo are included in the graph and the adjusted correlation coefficient of the line relationship of DAS28 at Day 56 vs. Day 14 is 0.73

The individual data from Day 7, Day 14, Day 28, Day 56 and Day 84 are graphed in matrix plots to investigate the relationship of paired time points in DAS28 change response (all data combined in Figure 7.8). The corresponding correlation coefficient, P-values and number of subjects between DAS28 at Day 7, Day 14, Day 28, Day 56 and Day 84 are presented in Table 7.6.

### DAS28 Change at Early Time Point and Late Time Point in meta-analysis

Figure 7.8 and Figure 7.9 are matrix plots of the DAS28 at early time point (x-axis) and at late time point (y-axis) from the meta-analysis. Similar positive correlations are found in DAS28 change between the early time point and late time. The mean of DAS28 change data at Day 14 (Week 2), Day 28 (Month1), Day 56 (Month 2), Day 84 (Month 3), and Day 168 (Month 6) from each study is presented with all treatments combined in Figure 7.8 and biological presented in Figure 7.9. The correlation coefficients for all treatments combined and biological are presented in Table 7.6 and Table 7.7 respectively.

The correlations between Day 14 and Day 56 are 0.89 and 0.88 in all study treatments combined and biological respectively. It is also shown that, as time point gets closer, the R square is relatively higher which indicates that adjacent early time points are a better predictor of late time points. Day 14 appears to be a good predictor of Day 56 in the RA population.

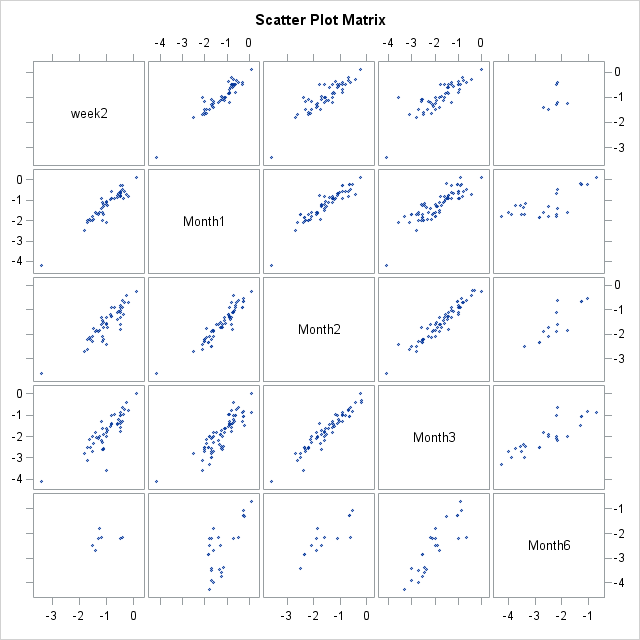


Figure 7.8 Matrix plot from the meta-analysis. All time points of DAS28 change from baseline at Week 2, Month 1, Month 2, Month 3, and Month 6 from the meta-analysis are displayed and all study treatments are combined. Each point is the mean from each treatment of clinical trials in the meta-analysis.

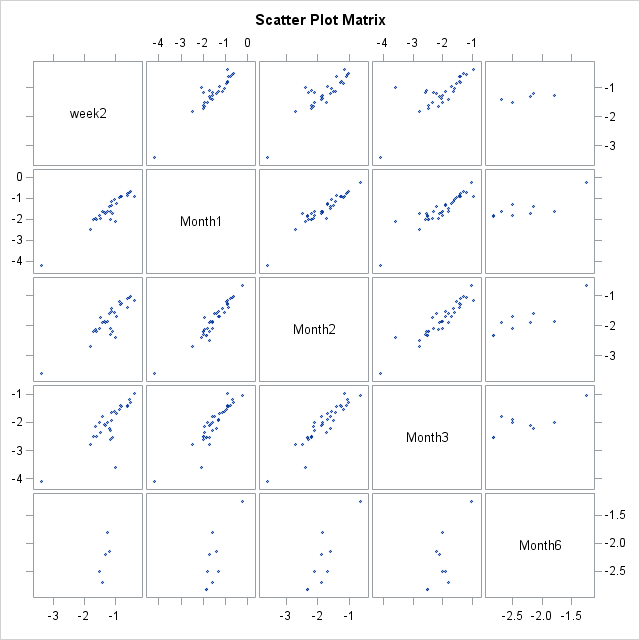


Figure 7.9 Scatter matrix plots of all time points of DAS28 change from baseline at Week 2, Month 1, Month 2, Month 3, and Month 6 from the meta-analysis with study treatment of biological. Each point is the mean from each clinical trial in the meta-analysis.

Table 7.6 Correlation coefficient, P-value and number of treatment arms between DAS28 change from baseline at Week 2, Month 1, Month 2, Month 3, and Month 6 from the meta-analysis combining all study treatments.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Pearson Correlation Coefficient of DAS28 from meta-analysis** | | | | | |
|  | **Week 2** | **Month 1** | **Month 2** | **Month 3** | **Month 6** |
| **Week 2** | |  | | --- | | 1.00 | |  | | 48 | |  |  |  |  |
| **Month 1** | |  | | --- | | 0.94 | | <.0001 | | 48 | | |  | | --- | | 1.00 | |  | | 68 | |  |  |  |
| **Month 2** | |  | | --- | | 0.89 | | <.0001 | | 48 | | |  | | --- | | 0.93 | | <.0001 | | 57 | | |  | | --- | | 1.00 | |  | | 62 | |  |  |
| **Month 3** | |  | | --- | | 0.85 | | <.0001 | | 48 | | |  | | --- | | 0.87 | | <.0001 | | 68 | | |  | | --- | | 0.96 | | <.0001 | | 59 | | |  | | --- | | 1.00 | |  | | 70 | | |  | | --- | |  | |  | |  | |
| **Month 6** | |  | | --- | | 0.32 | | 0.4771 | | 7 | | |  | | --- | | 0.72 | | <.0001 | | 25 | | |  | | --- | | 0.85 | | 0.0001 | | 14 | | |  | | --- | | 0.87 | | <.0001 | | 25 | | |  | | --- | | 1.00 | |  | | 25 | |

Table 7.7 Correlation coefficient, P-value and number of treatment arms between DAS28 change from baseline at Week 2, Month 1, Month 2, Month 3, and Month 6 from the meta-analysis with study treatment of biological.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Pearson Correlation Coefficient of DAS28 from meta-analysis** | | | | | |
|  | **Week 2** | **Month 1** | **Month 2** | **Month 3** | **Month 6** |
| **Week 2** | |  | | --- | | 1.00 | |  | | 28 | |  |  |  |  |
| **Month 1** | |  | | --- | | 0.94 | | <.0001 | | 28 | | |  | | --- | | 1.00 | |  | | 34 | |  |  |  |
| **Month 2** | |  | | --- | | 0.88 | | <.0001 | | 28 | | |  | | --- | | 0.95 | | <.0001 | | 34 | | |  | | --- | | 1.00 | |  | | 34 | |  |  |
| **Month 3** | |  | | --- | | 0.79 | | <.0001 | | 28 | | |  | | --- | | 0.91 | | <.0001 | | 34 | | |  | | --- | | 0.93 | | <.0001 | | 34 | | |  | | --- | | 1.00 | |  | | 34 | |  |
| **Month 6** | |  | | --- | | 0.75 | | 0.1462 | | 5 | | |  | | --- | | 0.80 | | 0.0089 | | 9 | | |  | | --- | | 0.87 | | 0.0022 | | 9 | | |  | | --- | | 0.77 | | 0.0158 | | 9 | | |  | | --- | | 1.00 | |  | | 9 | |

It is evident that there are significant correlations in mean DAS28 change between the mean of early time points (i.e. Day 14) and the mean of late time points (i.e. Day 56), which further indicate how predicting late clinical response using short-term data can be undertaken.

## Structural correlation of early response and late response at change from baseline

A number of research articles have investigated the relationship of both initial value and change from initial value (Oldham, 1962; Blomqvist, 1977; Blance, 2005). It has been shown that there is a negative bias in the change from initial values when compared to the correlation of the initial values (Chiolero, 2013).

There are several methods proposed to account for the structural correlation (Oldham 1962; Blomqvist, 1977; Blance, 2005). In 1962, Oldham suggested the correlation should be adjusted in the analysis of repeated measure from the same subject with the correlation being calculated using the correlation of (U2-U1) and (U2+U1)/2 where U1 is the early response and U2 is late response. An adjustment of the correlation coefficient method was proposed to account for the differences in the correlation between when the change and initial values are used (Blomqvist, 1977; Edland, 2000). A more complicated random effect model to adjust for the difference in association between baseline and change from baseline was also proposed (Blance, 2005).

In the context of this dissertation to understand the impact of the structural correlation using change from baseline for DAS28, an analytical approach is taken to mathematically derive the relationship of the correlation between early and late response in change from baseline and correlation in original value. A simulation analysis is then conducted to explore the impact on the correlation in change from baseline when the correlation between baseline and post-dose is 0.1, 0.2, 0.3, …, 0.7.

### Relationship of early response and late response in Change from baseline and original value

The aim of this section is to derive the correlation between early and late response in change from baseline as a function of correlation in original score, the function is

 7.1

with . Here we assume Yij is a vector with original measures of Yij1, Yij2, and Yij3 of correlated repeated measure of baseline, change from baseline at Day 14 and Day 56 respectively from the same patient of i treatment, j subjects, where i=1 if placebo and 2 if treatment.

**.**

Here, Y1, Y2, and Y3 denote the DAS28 response at baseline/pre-dose, Day 14 and Day 56; Cov is the covariance, σY1 the standard deviation of Y1, σY2 is the standard deviation of Y2, σY3is the standard deviation of Y3. µY2 and µY3 are the mean of Y2 and Y3 respectively.

Since the relationship is different for each treatment, we will focus on the correlation from a single treatment, so the formula of correlation coefficient of DAS28 between early response (Y2) and late response (Y3) is defined as (7.1) and can be rewritten as (7.2)

 7.2

**with**  **,**

**.**

For each treatment, the correlation of the DAS28 early response (Y2) and late response (Y3) is defined as

, , 7.3

where cov(Y2, Y3) = E[(Y2-µY2)(Y3-µY3)] = E[[(Y2-E[Y2])(Y3- E[Y3])] =E[Y2 Y3]-E[Y2] E[Y3]; E is the expectation

σY22= E[[(Y2-E[Y2]2) = E[Y22]-(E[Y2])2; and σY32= E[[(Y3-E[Y3]2]= E[Y32]-(E[Y3])2.

Similarly, the correlation coefficient of early response and baseline is given as

, 7.4

and the correlation coefficient of late response and baseline is written as

7.5

For a given individual trial, the correlation coefficient of change from baseline of early and late response can be written as

, 7.6

=

= ,

=

= ,

= ,

== .

Defining a, b, c, and d as

a=;

b=;

c=,and d=.

The correlation coefficient of change from baseline of early and late response can be given by

7.7

If the correlation of post-baseline response (Y2 or Y3) and baseline is zero, Eqn 7.7 becomes

Thus, the correlation of DAS28 change from baseline has a linear relationship to the correlation of raw DAS28 score with slope

and intercept of .

The correlation in change from baseline response is dependent of the correlations in raw score as well as the correlation of post-baseline and baseline and is equal to

,

when there are no correlations between the raw scores Y1, Y2, and Y3. Let ϒ2 be the ratio of σY2 and σY1, and ϒ3 to be the ratio of σY3 and σY1.ϒ2 = σY2 /σY1 and ϒ3 = σY3 /σY1,so the denominator can be written as

.

After replacing σY1, σY2, and σY3 with ϒ2 and ϒ3, the a, b, c, and d the above result is transformed to

Let a=; b=; c=; and d=.

In summary, there is a linear relationship between the correlation in change from baseline and the raw score. The slope and intercept of linear relationship is dependent on the variance in baseline and their correlation between baseline and post-dose.

Given the above result, the linear relationship of correlation of change from baseline and raw values are further illustrated in Figure 7.10 and Figure 7.11. It is shown in Figure 7.10 that the linear relationship varies with the correlation between baseline and post-dose and the standard deviation ratio of postdose/predose measurement.

* In the case of equal predose and postdose variance (ϒ2 and ϒ3=1), the correlation in change from baseline is similar to correlation in raw score when postdose/predose correlation is close to 0.5.
* Change from baseline is likely to overestimate the correlation in raw score when correlation is less than 0.5 between postdose/predose (ρY1,Y2 and ρY1,Y3 <0.5) and underestimate the correlation when are much higher correlation than 0.5.
* Change from baseline is likely to overestimate the correlation in raw score when the baseline variance is higher (i.e. ϒ2 and ϒ3 <1) and under-estimate the raw score correlation when the postdose variance is higher than that in pre-dose (i.e. ϒ2 and ϒ3 >1) (Figure 7.10 and Figure 7.11).

Therefore, structurally the correlation of change from baseline (time 1) to time 2 vs change from baseline (time 1) to time 3, is positive (> 0) even if the raw scores of early response (time 2) and late response (time 3) are independent. The analytical formula confirms the existence of the structural correlation. In the next section, we will investigate the impact of this structural correlation in the context of a Phase 2a trial in RA.

A: ϒ2= ϒ3 = 0.6

B: ϒ2= ϒ3 = 1.0

C: ϒ2= ϒ3 = 1.4

Figure 7.10 Plots of correlation of change from baseline (CFB) and raw response of Day14 and Day56 when the correlation of baseline and post-dose is 0, 0.1, 0.3, 0.5, 0.7. ϒ2 and ϒ3 are the ratios of standard deviation of Day 14/predose and Day 56/predose respectively and assumed to be equal. Top (A) ϒ2 = ϒ3 = 0.6; Middle (B) ϒ2 = ϒ3 = 1.0; Low (C) ϒ2 = ϒ3 = 1.4.

Figure 7.11 A plot of correlation of change from baseline (CFB) and raw response of Day14 and Day56 when the ratios of Standard Deviation (Std) of postdose and baseline is 0.6, 0.8, 1.0, 1.2, and 1.4. The correlation of baseline and postdose is fixed at 0.7.

In the PoC study of GSK123456, the correlation coefficient between pre-dose and Day 14 (ϒ12) or Day 56 (ϒ13) response is approximately 0.7 and standard deviations are 0.9, 1.2, and 1.3 respectively for pre-dose response, Day 14 and Day 56 response. The relationship of correlation in original value and change from baseline (CFB), which is calculated using (7.7), is presented in Table 7.8. This is a likely scenario for future PoC study of GSK654321.

The results in Table 7.8 shows that the correlation in CFB is always lower than correlation in original value and the ratios are 0.04, 0.35, 0.59, 0.76, 0.89, and 0.95 when correlation in original scores are 0.5, 0.6, 0.7, 0.8, 0.9 and 0.95 respectively. It is counter-intuitive that the correlation can get close to zero (correction = 0.02) in CFBs between Day 14 and Day 56 and results are confirmed further by simulations (simulation is performed similar to Section 7.6.2). This may be a result of the structural correlation discussed in this Section, however, the exact reason is not known.

Table 7.8 The correlation of change from baseline (CFB) and raw response of Day14 and Day56 This is the likely scenario in PoC study of GSK123456 when correlation coefficient between pre-dose and Day 14 (ϒ12) or Day 56 (ϒ13) response is 0.7 and standard deviations are 0.9, 1.2, and 1.3 respectively for pre-dose response, Day 14 and Day 56 response.

|  |  |  |
| --- | --- | --- |
| Correlation in original value | Correlation in change from baseline (CFB) | Ratio of Correlations  CFB/Original |
| 0.50 | 0.02 | 0.04 |
| 0.60 | 0.21 | 0.35 |
| 0.70 | 0.41 | 0.59 |
| 0.80 | 0.61 | 0.76 |
| 0.90 | 0.80 | 0.89 |
| 0.95 | 0.90 | 0.95 |

### Simulation to investigate Change from baseline and original value at meta-analysis

In previous sections, the relationship of the correlation in change from baseline compared to the correlation in original score is derived in the context of a single study. In the meta-analysis, there are no individual patient level outcome data available, so the mean of early response is correlated with mean of late response. Therefore, in this section, the relationship will be explored to understand the impact of change from baseline over the mean of early response and late response through simulations. The aim of this section is thus to use simulations to investigate the impact on correlation in mean change from baseline when the correlation between baseline and post-dose varies in a meta-analysis.

The original score of baseline, Day14 and Day56 responses are simulated using tri-variate model as illustrated below. Yj1, Yj2, and Yj3 denote the DAS28 response at baseline/predose (time 1), Day 14 (time 2) and Day 56 (time 3); Cov is the Covariance; σY1 is the standard deviation of Y1, σY2 is the standard deviation of Y2, σY3 is the standard deviation of Y3. µY2 and µY3 are the mean of Y2 and Y3 respectively.

The parameters in the simulation are described below in Table 7.9.

 7.9

**with**  **,**

**.**

Table 7.9. The assumed value of the tri-variate normal distribution in the simulation.

|  |  |
| --- | --- |
| **Parameters** | **Value assumed in the simulation** |
| µ1:Mean DAS28 of pre-dose response\* | 6.5 |
| µ2:Mean DAS28of Day 14 response\* | 5.5 |
| µ3:Mean DAS28of Day 56 response\* | 5.0 |
| σ1: Standard Deviation of pre-dose response\* | 0.9 |
| σ2: Standard Deviation of Day 14 response\* | 1.2 |
| σ3: Standard Deviation of Day 56 response\* | 1.3 |
| ρ12 = ρ13: Correlation coefficient between pre-dose and Day 14 (ϒ12) or Day 56 (ϒ13) response | 0, 0.1, 0.3, 0.4, 0.5, 0.6, 0.7 |
| ρ23: Correlation coefficient between Day 14 and Day 56 response | 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 |

\*: the assumed values were chosen to closely reflect the data from Phase 2a trial of GSK123456.

Simulation steps are illustrated below:

1. Simulate 64 patients with baseline (pre-dose), Day14 and Day56 response using tri-variate Normal distribution as illustrated as below in (7.9).
2. Yj1, Yj2, and Yj3 denote the DAS28 response at baseline/predose, Day 14 and Day 56 respectively for patient j. The change from baseline for Day 14 and Day 56 are calculated by subtracting baseline response from the post-treatment response.
3. Calculate the mean change from baseline for Day 14 and Day 56.
4. Repeat 10,000 times and calculate the correlation of change from baseline results at Day 14 and Day 56 over the mean changes.

The relationship of the correlation between change from baseline and raw scores for different scenarios in pre-dose and post-dose correlation is shown in Figure 7.12. It is shown that the correlation of early and late response in raw value is higher than that using change from baseline (CFB) results when pre-post-dose correlation is greater than 0.4. If the pre-post-dose correlation is less than 0.4, using the analysis in CFB is likely to be lower than the correlation in raw scores. It is estimated that the correlation coefficient is 0.60 and 0.75 for baseline/Day 14 and baseline/Day56 respectively. Therefore, the estimated correlation of Day 14 and Day 56 response using CFB method is not over-estimating the correlation compared to the raw scores. The simulation was carried out in R version 3.2.5 and R code is presented in Appendix 11.1.5.

Figure 7.12. The relationship of change from baseline (CFB) and raw response of Day14 and Day56 in the context of meta-analysis based on simulation when the correlation of baseline and post-dose is 0, 0.1, 0.3, 0.4, 0.5, 0.6 and 0.7.

### Summary of Structural Correlation

As presented in Figure 7.10 and 7.11, the correlation of CFB is linearly correlated with that of raw score between early response and late response and the slope is determined by the variance and correlation of pre-dose (baseline) and post-dose values. The analytical formula confirms the existence of the structural correlation. The use of change from baseline would result in structural correlation since the baseline is incorporated in both early response and late response, Senn (2006) suggested that analysis of covariance (ANCOVA) model with raw score as the dependent/outcome variable and baseline as covariate would perform better comparing to univariate model with change from baseline as the dependent/outcome variable.

Frison and Pocock (1992) investigated and compared the variance using ANCOVA, post-baseline and change from baseline methods in randomized clinical trials and showed that ANCOVA has a smaller variance and is superior but the difference between ANCOVA and change from baseline method is smaller as the correlation between baseline and post-baseline increases. The change approach is the method based on change from baseline. While comparing post-baseline method and change method, change method is better than post-baseline method when correlation between baseline and post-baseline is higher than 0.5. The variance of ANCOVA and change method are dependent on number of pre-treatment visits, extra pre-treatment visit(s), which improve efficiency for methods with baseline involved i.e. ANCOVA and change method (Frison, 1992) and make change method closer to ANCOVA method in statistical efficiency as measured by variance (Frison, 1992).

Our investigation into the potential structural correlation supported the results of correlation in change approach and showed that the relationship between association of change scores and its association between raw scores is linearly correlated and the slope is close to 1 when correlation between baseline and post-baseline data is around 0.5, assuming equal variance of baseline, and post-dose value. If there is no or little correlation between baseline and pose-baseline, the correlation of change scores is likely to overestimate the raw score correlation. The association of change scores will under-estimate the association in raw score when there is moderate or strong correlation between baseline and post-baseline data.

In summary, due to the potential structural correlation with change from baseline, it is generally recommended to use raw score in the calculation of correlation between early and late response. However, for endpoints such as DAS28 change from baseline have been commonly used as the preferred outcome in many RA clinical trials (Fleischmann 2015; Dhir 2014; Smolen 2013, 2014; Genovese 2014). Although the ANCOVA method using raw scores with baseline as covariate would be the preferred method of statistical analysis (because it results in the smallest treatment variance estimate); our results showed that the estimated correlation of the Day 14 and Day 56 responses results in a similar or under-estimating the correlation using raw score. With moderately strong correlation in change score, it is reasonable to use change from baseline as primary endpoint for the DRPM prediction in late response using early response in the context of Phase 2a RA trial.

## ****Correlation and**** causality in v****alidation of surrogate Endpoint****

In earlier sections, a correlation analysis was conducted for a single trial of GSK123456 PoC study and a meta-analysis for the DAS28 response at an early time point (i.e. Day 14) and a late time point (i.e. Day 56) with the aim show how the DAS28 late response can be predicted using a surrogate the DAS28 early response. The correlation coefficient was 0.84 for individual patient responses in the PoC trial, between change in DAS28 at day 14s and day 56, and 0.93 for overall means from 20 clinical trials in meta-analysis. Higher correlations are indicative of higher statistical association, but high correlation does not always suggest the causality of response or the validation the surrogate endpoint to final endpoint (Molenberghs, 2005; Buyse, 2000; Alonso, 2004).

There are examples in the literature of wrong predictions of the true endpoints using surrogate endpoints resulting due to the failing to account for the totality of treatment effect on final endpoints (Fleming and DeMets, 1996, Buyse 2000). Just because a surrogate is predictive of the clinical outcome it does not mean that the treatment effect observed with a surrogate would be predictive of the treatment effect on the clinical outcomes.

An example of a situation where a surrogate failed to predict the clinical outcome is the antiarrhythmic drug Encainide, which was approved by FDA with the surrogate endpoint of arrhythmias (Alonso 2007). It was then found that although the association was high between the arrhythmias suppression and final endpoint of cardiac mortality, that did not imply the treatment induced changes between changes in arrhythmias and cardiac events. The drug Encainide was withdrawn from market in 1991 (Alonso 2007).

The International Conference on Harmonisation (ICH, 1998) guidance describes the strength of surrogate endpoint which should be established with three conditions and they are

1. Biological plausibility of relationship;
2. Prognostic value of surrogate endpoint for final endpoint;
3. The treatment effect on the surrogate endpoint corresponds to the effect on the final endpoint in trials.

As a result, Buyse et al (2000) suggested that a surrogate endpoint (i.e. DAS28 change at Day 14) be “validated” before a surrogate endpoint can replace a final endpoint (i.e. DAS28 change at Day 56) in the evaluation of an experimental treatment. To achieve this, there should be a clear distinction when there are clinical effects and no effect using surrogate endpoint to predict the true final endpoint (Buyse 2000).

Biological causality is mainly discussed at the 1st condition and the 2nd and 3rd conditions focus on statistical evidence. In our example, both early response and late response of DAS28 are measured longitudinally from the same patients therefore, it is logical to believe biological plausibility between the two endpoints, which satisfies the first condition.

In this section, the statistical evidences of condition #2 and #3 of surrogacy are discussed. We will review the methods that were proposed to validate the surrogate endpoint in single trial and multiple trials followed by evaluation and discussion of early response of DAS28 i.e. Day 14 as a surrogate.

It is noted that in the proposed study design, DAS28 from both day 14 and day 56 are collected. DAS28 response predicted from Day 14 is used to expedite the decision making at interim analysis. Thus, the Day 14 response serves as a surrogate for Day 56 in early decisions. As data accrue, Day 56 response will be available when assessing the doses response on Day 14 for the current dose to help predict the Day 56 response. This approach is similar to the ASTIN trial (Grieve 2005) which was an adaptive design where a similar surrogate endpoint was used in the decision: the Scandinavian Stroke Scale.

### ****Validation of surrogate endpoints in Single-trial setting****

For the method of surrogate endpoint validation in individual trial, Prentice (1989) proposed validation criterion for surrogate endpoint to account for the relationship of treatment and final true response. If we use symbol S to represent surrogate endpoint, T to represent final true endpoint, and Z to represent treatment. Prentice’s validation criteria (1989) require that the treatment Z should have a significant effect on the surrogate and true endpoints and the surrogate endpoint should have a significant effect on the true endpoint T and the treatment effect on the true endpoint should be fully captured by surrogate endpoint. Symbolically, let *Tj* and *Sj* be random variables of true (DAS28 at Day 56) and surrogate endpoints (DAS28 at Day 14), respectively, for the *j*th subject and let *Zj* be an indicator variable for treatment, so Prentice’s method (1989) can be illustrated using below formulas (Buyse 1998, 2000),

Sj = µS + αZj + εSj 7.7

Tj = µT + βZj + εTj 7.8

Tj = µ + ϒSj + εj 7.9

where α and *β* are the slopes for the treatment Zon surrogate endpoint S and true endpoint T. ϒ is the slope of the true endpoint (T) on surrogate endpoint S as linear regression. *Sj, Tj* and Zj are surrogate response, true response, and treatment in jth patients. µS, µT, and µ are the intercepts for the linear regressions from (7.7), (7.8), and (7.9) respectively. ej is error term for effect of true effect on surrogate endpoint. Both eSj and εTj, are Normally distributed correlated errors which satisfy

, and 

Lastly, the true response is examined using linear regression model with both treatment and surrogate endpoint as shown below.

Tj = µT + βsZj + ϒzSj + εTj 7.10

where Sis the estimate of the effect of *Z* on *T* after adjusting surrogate endpoint S. s is the parameter estimate for treatment effect Z; ϒz is the slope estimate of effect of surrogate endpoint S on true endpoint T after adjusting treatment Z. To achieve surrogacy using Prentice’s criteria, all four elements of the Prentice criteria are needed, i.e. null hypothesis testing of zero slope effect of α, β andϒ are rejected and null hypothesis testing of βs is not rejected.

Though Prentice’s method is a direct and intuitive approach, all four criteria are difficult to achieve (Buyse 2005). In addition, there are inherent issues. The main concern for Prentice’s criterion is that the method is useful to reject a bad surrogate when rejecting βs as zero but it is less suitable to use to validate a good surrogate (Buyse 2005). In addition, non-significant of βs may be due to lack of statistical power resulting from a small sample size. More importantly, the fourth Prentice criteria test (Ho: βs =0), even if not rejected, cannot explain 100% of treatment captured in true response T. Therefore, a proportion of explained (PE) method was further developed (Freedman 1992).

The proportion of treatment effect explained (PE)is defined as proportion of effect of treatment Z on T that is explained by surrogate endpoint S (Freedman, 1992) and the PE is close to 1 in case of a good surrogated endpoint. Per definition, Freedman’s proportion explained (PE) is calculated using the formulae below:

,

where β and βs are the slope estimates of formulae and regression models 7.8 and 7.10. The PE is a ratio of parameters so the confidence interval of PE can be calculated using delta method (Buyse, 2005). PE extended Prentice’s method and simplified the validation criteria, but the PE method (Buyse, 2005) also has its flaws, it is expected that PE is 1 when all the treatment effect is captured by S (i.e. βs =0), however, mathematically the βs may be larger or smaller than β, so that PE is not constrained to be between 0 and 1. Another criticism of the PE method by Frangakis and Rubin (2002) was that PE does not consider a causal effect since the treatment effect on true response T is after conditioning on the post-randomization S.

Due to these limitations, two new quantities, the relative effect and adjusted association, were proposed to assess surrogacy by Buyse and Molenberghs (1998). For continuous Normally distributed endpoints S and T, the relative effect (RE) is defined as the ratio of effect of Z on T and effect Z on S and the formula is

, 7.11.a

where α and β are the slope estimate from (7.7) and (7.8). The adjusted association shows the correlation coefficient of S and T after adjusting the treatment effect. It is derived as individual-level association between surrogate and true endpoints, after accounting for the effect of treatment, the formula is

ρZ = r(S, T|Z ) = Corr(εSj, εTj) = 7.11.b

where σSS, σTT, and σST are the parameters in the covariance matrix of εSj and εTj. The confidence interval of Zcan be calculatedusing bootstrap methods (Buyse 2005).

In this section, three types of methods were discussed to validate surrogate endpoint in a single-trial setting. Both Prentice and PE methods have fundamental issues as described above so the method of two quantities, RE and adjusted correlation. Each calculates the effect of the true and the surrogate endpoints and the correlation after adjusting for the treatment effect and are recommended in practice (Alonso and Bigirumurame 2016) to evaluate the validation. All the methods being proposed are in a single-trial setting so there is no trial replication assessed in multiple-trial or meta-analysis setting. In the next subsection, validation methods will be discussed multiple-trial setting.

### ****Validation of surrogate endpoints in multiple-trial setting****

In the multiple-trial setting, the surrogate endpoint and final true endpoint are collected in multiple trials and both individual patient data and trial data are assumed to be available for the analysis. Both fixed effect models and random effect models with treatment as an independent variable are suggested in the literature (Buyse and Molenberghs 1998) in the validating the surrogate endpoint, the method is called meta-analytic approach (Buyse and Molenberghs 1998).

The two-stage fixed effect model is illustrated as below. At the first stage, the surrogate endpoint and true final endpoint are assumed to be jointly Normally distributed with treatment effect model presented in both surrogate endpoint and true endpoint in a linear model.

Sij = µSi +αiZij + εSij 7.12

Tij = µTi + βi Zij + εTij 7.13

where µSi and µTi are the intercepts for the linear regressions respectively. Here, αi and *β*i are the corresponding treatment effect with εj being the error term for effect of true effect on surrogate endpoint. εSij and εTij, are correlated errors

, where 

At second stage, it assumes the µSi, µTi, αi and *β*i are a mixture of fixed and random effects, so

,

where the first component is a vector of fixed effects and the second component is vector of random effects and the random effects follow a zero mean normal distribution with dispersion matrix of D with dSS, dTT, daa, and dbb as standard deviation for the random effects and corresponding matrix as below.

D =

In comparison to the two-stage approach, the second method is to have of random or mixed effect model so both fixed effect and random effect variables are fitted in the model. Assume *Tij* and *Sij* be random variables of true (DAS28 at Day 56) and surrogate endpoints (DAS28 at Day 14), respectively, for the *i*th trial and *j*th subject and let *Z ij* be an indicator variable for treatment. The mixed-effect model (Buyse 2000) is

Sij = µSi + mSi + (α+ ai) Zij + εSij 7.13

Tij = µTi + mTi + (β+ bi) Zij + εTij 7.14

where µSi and µTi are the intercepts for the linear regressions, α and *β* are the fixed slope of the treatment *Z* on surrogate endpoint S and true endpoint T; ai and *b*i are the corresponding random treatment effect; mSi and mTi are the corresponding random trial effect at ith trial; εj is error term for effect of true effect on surrogate endpoint. εSij and εTij, are correlated error terms and random effects follow a zero mean Normal distribution with dispersion matrix of D similarly as discussed above.

It was suggested that the surrogacy should be established through both methods of trial-level surrogacy and individual patient -level surrogacy (Buyse, 2000). The trial-level surrogacy can be assessed using trial level coefficient of determination between treatment effect on true response T and the treatment effect on S. The individual level surrogacy is accessed using individual level coefficient of determination after adjusting for both trial and treatment effect.

The full form of trial level coefficient of determination formula can be found in the literature (Buyse, 1998 and 2000). The reduced form of trial-level, which assume random slope is independent of intercept, can be expressed as

7.15

The individual level surrogacy is calculated using the association between surrogate and final endpoint adjusting for the treatment effect.

7.16

Buyse et al (2000) extended these concepts to multiple randomized trials and suggested the surrogacy if the coefficients of determination of both trial level and individual level are close to one.

Previous discussions have focused on the methods when surrogate S and true T endpoints are Gaussian random variables; other methods have been developed to evaluate surrogate endpoints in binary, time-to-event or ordinal endpoints. Although the meta-analytic approach can be extended to non-continuous endpoints, there were challenges in numerical issues with the hierarchical model (Alonso, 2016), so other methods are proposed include as likelihood reduction factor method (Alonso, 2004), Information Theoretic Approach (Buyse, 2006; Alonso, 2016), and correction based on cross-validation (Abrahantes, 2008). Since the interest of this dissertation are continuous endpoints, therefore, only meta-analytic methods are discussed.

### Validation of DAS28 at Day 14 as Surrogate Endpoint

In the single trial setting of phase 2a PoC trial, both the surrogate endpoint Day 14 and final true endpoint Day 56 are fitted in a linear model with treatment as independent variable. The relative effect and adjusted association are then calculated using (7.11.a) and (7.11.b). The results with estimate of RE and confidence interval are presented in Table 7.10 for a comparison of placebo with all treatment doses combined. The calculations were performed in R (version 3.2.5).

Table 7.10 The relative effect and adjusted association in phase 2a PoC trial of GSK123456 between early response Day 14 and late response Day 56.

1. The calculation is based on DAS28 change from baseline between Day 14 and Day 56

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Relative effect | | Adjusted association (ρZ) | |
|  | Estimate (Std Err) | 95% Confidence Limit | Estimate (Std Err) | 95% Confidence Limit |
| Placebo vs all doses | 1.84 (1.43) | (-0.97, 4.64) | 0.73 (0.07) | (0.58, 0.84) |
| Placebo vs 3 mg/kg | 1.66 (0.88) | (-0.06, 3.38) | 0.70 (0.09) | (0.54, 0.89) |
| Placebo vs 10 mg/kg | 1.67 (1.42) | (-1.12, 4.46) | 0.73 (0.12) | (0.41, 0.93) |

1. The calculation is based on DAS28 raw score between Day 14 and Day 56

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Relative effect | | Adjusted association (ρZ) | |
|  | Estimate (Std Err) | 95% Confidence Limit | Estimate (Std Err) | 95% Confidence Limit |
| Placebo vs all doses\* | 1.54 (1.21) | (-0.83, 3.91) | 0.83 (0.07) | (0.75, 0.90) |
| Placebo vs 3 mg/kg\* | 1.42 (0.80) | (-0.14, 2.98) | 0.86 (0.10) | (0.71, 0.94) |
| Placebo vs 10 mg/kg\* | 1.67 (1.54) | (-1.34, 4.69) | 0.82 (0.11) | (0.64, 0.91) |

\*The calculation is based on ANCOVA model with baseline as covariate.

When placebo is compared to the average of all doses, the relative effect is 1.84 with standard error of 1.43. The adjusted association is 0.73. The relative effects are lower at 1.66 or 1.67 when placebo is compared to 3 mg/kg or 10 mg/kg respectively. The adjusted association is similar at between 0.7 and 0.73. The estimate of relative effect and adjusted association are quite consistence for all 3 scenarios although there was much lower variance when placebo is compared to 3 mg/kg. The relative effect and adjusted association with DAS28 raw value are presented in Table 7.10A. When placebo is comparing to average of all doses, the relative effect is lower at 1.54 with smaller standard error of 1.21. The adjusted association is 0.83, a higher than those using change from baseline values. The PoC study is a relatively smaller study with total sample size of 64 (16 patients receiving placebo and 48 treated with 6 difference doses). With the small sample size, it is still shown a strong adjusted association after adjusting for treatment effect. The results are similar in the validation of surrogacy using raw scores and change from baseline values and all the relative effects are greater than 1. The primary endpoint of the follow-on study is the change from pre-dose DAS28 – not the actual DAS28 at a time point. We demonstrated that, despite the issue of structural correlation with change from baseline, there is still strong correlation in change score after adjusting for treatment effect.

Overall, in a single-trial setting, the relative effect of DAS28 at Day56 and Day 14 is between 1.66 and 1.84, which mean a treatment effect of Day 56 response is at 1.66-1.84 times the effect of Day 14 response. The point estimate indicates that magnitude of treatment in true final endpoint is 1.66-1.88 times the magnitude of treatment in surrogate endpoint. The lower confidence interval of relative effect contains 0 which indicates that the relative effect is not significant with the small cohort of sample size.

These results are limited to single-trial data from Phase 2a trial of GSK123456, the meta-analysis data, which provides means from each of 20 trials and is external to GSK, provides more assurance of relationship in DAS28 change between Day 14 and Day 56. The meta-analytics model as described in Section 7.6.2 is fitted to the meta-analysis data. The variance and covariance matrix D of the parameters is then obtained from the two-stage fixed effect model. For meta-analysis data, the R-square of trial-level surrogacy is 0.83 with 95% confidence interval of (0.57-1.00), so R (correlation)–value is 0.91. The calculation was based on in a reduced effect model. The confidence interval is obtained via delta method. Meta-analysis provides a way to inform the actual analysis using data external to pharmaceutical company, i.e. GlaxoSmithKline, that the relationship between day 14 and 56 seems valid.

An R-value close to 1 would mean the trial-level treatment effect in final true endpoint is a multiply of the surrogate endpoint response, so an R-value of 0.91 suggests a strong surrogacy; the precision of the R-value estimate is still small given the lower bound of 95% confidence interval is 0.57. All the individual data are required in the validation of surrogate endpoint at individual-value. However, there was no subject level data, only mean or summary measure data available in meta-analysis being conducted in early response and late response of DAS28, the individual surrogacy is therefore not calculated.

### Discussion of DAS28 at Day 14 as Surrogate Endpoint

As discussed in earlier sections, though there are strong correlations in DAS28 between early and late response, the DAS28 at Day 14 being used as surrogate endpoint for Day 56 response is needed to be validated. ICH guidance (1998) discussed three conditions to access the causality and validate the surrogate endpoint. Both early response and late response of DAS28 are measured longitudinally from the same patients, therefore, it is logical to believe biological plausibility between the two endpoints. The second condition discusses the predictive value of surrogate endpoint to final endpoint and the third condition covers the relative treatment effect between surrogate and final true endpoint. It was shown that relative treatment effect (RE) is 1.66-1.84 with adjusted association of 0.70-0.73. The adjusted association of 0.70-0.73 is indicated a moderately strong correlation after adjusting for treatment effects. Therefore, for a given patient, the final true response of DAS28 at Day 56 can be predicted using his or her response at Day 14 with moderate level of accuracy.

At multiple-trial level, a strong R-value of 0.91 is indicative of a strong association at trial-level, however, no individual level association is available due to lack of individual data from the meta-analysis. Piedbois and Buyse (2004, 2008, and 2010) suggested a full validation should be conducted using individual patient data in meta-analysis. It would be optimal for GSK654321 for future research should be conducted to further validate the surrogate endpoint using individual level data with a meta-analysis.

### Model Parameters from Covariate Model with DAS28 Early Response as Covariate

Since the relationship may be impacted by the study treatment received, the statistical analysis of meta-analysis data is repeated to investigate the effect of the treatment group using analysis of covariance (ANCOVA) model in the prediction of late time point DAS28. ANCOVA model assumes a linear relationship between the response of DAS28 change at late time point and covariate DAS28 change at early time point. The formula of ANCOVA model is

,

where αi is the model corrected effect on y given group i, i = 1 is the placebo/MTX, 2 is biological/MTX, 3 is other RA drugs, Xij is DAS28 change response at early time point, β is the slope of late response as a function of early response and µ is intercept (also known as grand mean).

The meta-analysis data are modelled using the ANCOVA and results are presented in Table 7.11 and Table 7.12. Model parameters including parameter estimates and their standard errors are presented. The highest influential values are identified visually and by Cook’s distance method and were removed from the analysis. The interaction term of treatment and covariate is included to evaluate if there are interactions between treatment and early response. It is shown in the analysis that the P-value of interaction for predicting Month 2 (model 2) from Week 2 is 0.3724. The interaction term is excluded from the final model since the P-values of interim terms are not significant at 5%. Additional model fitting and diagnostics are shown in the Appendix.

Table 7.11 Model parameters including the intercept, the slope of covariates and their standard errors between DAS28 change from baseline at Day 14 and late time point from meta-analysis. PBO refers to placebo/MTX, Bio refers to biological drugs and the intercept of other RA drugs is zero.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | **Late Time Point from Early Time Point** | Parameter | Estimate | Standard Error |
| 1 | **Month 1 (Day 28) from Week 2 (Day 14)** | µ | 0.120 | 0.120 |
|  |  | β | 1.214 | 0.093 |
|  |  | α1 (PBO) | -0.250 | 0.117 |
|  |  | α2 (Bio) | -0.190 | 0.084 |
|  |  | MSE | 0.278 |  |
| 2 | **Month 2 (Day 56) from Week 2 (Day 14)** | µ | -0.388 | 0.148 |
|  |  | Β | 1.100 | 0.115 |
|  |  | α1 (PBO) | -0.055 | 0.144 |
|  |  | α2 (Bio) | -0.105 | 0.1020 |
|  |  | MSE | 0.422 |  |
| 3 | **Month 3 (Day 84) from Week 2 (Day 14)** | µ | -0.473 | 0.212 |
|  |  | β | 1.283 | 0.165 |
|  |  | α1 (PBO) | -0.011 | 0.207 |
|  |  | α2 (Bio) | -0.033 | 0.148 |
|  |  | MSE | 0.869 |  |

Table 7.12 Model parameters including the intercept, the slope of covariates and their standard errors between DAS28 change from baseline at Day 28 and late time point from meta-analysis. PBO refers to placebo/MTX, Bio refers to biological drugs and the intercept of other RA drugs is zero.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | **Late Time Point from Early Time Point** | Parameter | Estimate | Standard Error |
| 4 | **Month 2 (Day 56) from Month 1 (Day 28)** | µ | -0.482 | 0.105 |
|  |  | β | 0.963 | 0.064 |
|  |  | α1 (PBO) | 0.168 | 0.106 |
|  |  | α2 (Bio) | 0.079 | 0.082 |
|  |  | MSE | 0.317 |  |
| 5 | **Month 3 (Day 84) from Month 1 (Day 28)** | µ | -0.825 | 0.132 |
|  |  | β | 1.046 | 0.087 |
|  |  | α1 (PBO) | 0.373 | 0.140 |
|  |  | α2 (Bio) | 0.320 | 0.109 |
|  |  | MSE | 0.694 |  |
|  |  |  |  |  |

### Other Models

Baseline DAS28 data are included in the ANCOVA model to explore if the DAS28 effect at Month 2 (Day 56) can be explained by the additional covariate in addition to DAS28 at early time point. The P-value is 0.33 and baseline effect is not significant, so baseline effect is removed from the analysis model. Other statistical model with baseline covariate such as mean disease duration at baseline or mean HAQ-DI are attempted to adjust the ANCOVA model based on the two prognostic factors, but neither effect was significant.

### Model Comparisons

ANCOVA analysis is considered a superior and more sensible method since the analysis uses all the data compared to the separated analysis method by treatment (Section 7.4.3). The model parameters from ANCOVA could be used to construct a predictive model to predict DAS28 change from baseline data at Day 56 using Day 14.

## Discussion

The primary endpoint of the Phase 2a study of GSK654321 is DAS28 change from baseline, therefore it is necessary to predicate the DAS28 change from baseline at Day 56 using early response at Day 14, which may result in potential structural correlation. It was shown in this chapter that there were positive correlations in clinical efficacy endpoints of DAS28 raw score as well as change from baseline between early time point and late time point from single PoC study and a meta-analysis. Our investigation into the potential structure correlation showed that the relationship between association of change scores and association between raw scores is linearly correlated and the association of change scores will under-estimate the association in raw score when there is moderate or strong correlation between baseline and post-baseline data, which is a likely case in our study in Phase 2a trials. With moderately strong correlation in change score, it is reasonable to use change from baseline as primary endpoint for the DRPM prediction in late response using early response in the context of Phase 2a RA trial. With limited data, it is challenging to fully validate the DAS28 Day 14 as surrogate endpoint of Day 56. Therefore, future research is still needed to validate the surrogate endpoints.

In summary therefore, the ANCOVA model based prediction of late time points using early time endpoint has the potential to be used to construct a predictive model to predict DAS28 change from baseline data at Day 56 using Day 14. In the next chapter, the model based predictive model will be used to predict the late response based on earlier responses and to facilitate faster and early decision-making in the context of Phase 2a PoC.

# Improving the RA proof of concept desigN USING Early time points to speed UP decisions

In chapter 7, a retrospective analysis and systematic literature review was conducted to investigate the relationship between DAS28 at early time points and late time points. It was highlighted that there was a significant linear relationship in DAS28 change from baseline between late time point - specifically that the outcome assessed at Day 14 was a predictor of the late DAS28 response. In the Phase 2a PoC RA study, DAS28 responses at Day 14 and Day 56 are measured with Day 56 being the primary endpoint. To complete the 8-cohort dose response study for GSK123456 it took more than 2 years. If prediction of the Day 56 response from Day 14 could be done to expedite the decision-making, there is the potential that the time to complete the study could be reduced.

In this chapter, we will apply the predictive model by re-running the PoC for GSK123456 imagining the interim study decisions were being made at Day 14. Later, the delayed response predictive models are evaluated using simulation in the context of a Phase 2a PoC clinical trial.

## The Aims of the Chapter

The aim of this chapter is to develop a delayed response model to predict the DAS28 outcome at later time point using the DAS28 outcome data collected at earlier time points to facilitate faster and earlier decision making.

## Retrospectively Fitting Between Cohort Data

One of concerns in the implementation of adaptive designs in RA is that the time to complete adaptive design trials is too long due to the chronic nature of the disease. In the case of long-term endpoints/outcomes, it is optimal for an analysis decision to be based on a shorter-term endpoint, which could serve as a surrogate for the long-term endpoint. This section will illustrate the use of several statistical models to predict later time point using early time point in the Phase 2a PoC study.

In the context of this chapter, the early time point is DAS28 score collected on Day 14 and late time point is DAS28 score on Day 56.

### Retrospective Model Fitting in PoC Study

A schematic example of dose escalation and dose finding trial to use Day 14 to predict Day 56 in the context of the PoC study is displayed below in Figure 8.1. The predictive model uses all available DAS28 data at the time of analysis, including Day 14 data from the current cohort and Day 56 data from past cohorts. For example, for the Cohort 3 analysis, when the DAS28 at Day 14 is available for the last subject in Cohort 3 (6 assigned treatments and 2 Placebo) there are also DAS28 at Day 56 available for previous cohorts (here cohort 1 and cohort 2). Thus the Day 56 from past cohorts can be used also to predicted Day 56 data in Cohort 3 with the day 14 data in Cohort 3.

Cohort 3

Cohort 2

Cohort 1

Figure 8.1 Schematic plot of DAS28 change from baseline for between cohort interim analyses. The Day 14 data of DAS28 at Cohort 3 is projected to Day 56 and combined with early complete cohort data for the dose response model analysis.

To further understand the relationship of Day 14 and Day 56 and explore whether there is a viable option to achieve the potential time saving in the projection of later time point based on early read-outs, DAS28 data of GSK123456 from early chapters are analysed retrospectively by cohort in the same sequential order as in the original clinical trial. The dose levels as stated in the RA Phase 2a study (PoC) of GSK123456 were undertaken to compare the efficacy (Choy 2013). As described in previous chapters the study design was a dose staggering dose escalation design and the range of doses were 0.03-30 mg/kg based on patients’ body weight. Initially, six cohorts of eight patients each were enrolled (Cohorts 1 through 6). After the first interim analysis, an additional two cohorts of patients were enrolled (Cohort 7 and Cohort 8).

The next section describes how delayed response prediction models (DRPM) can be used to analysis the DAS28 outcome data in GSK654321 using early time points (day 14) to predict later data (day 56).

## Delayed Response Predictive Model

DRPM have been used in clinical trials (Hardwick, 2006, Koopmeiners, 2014, Fu, 2010). Weir et al. (Weir, 2007) assumed that early responses were a strong predictor of late responses and the last observation carried forward (LOCF) method could be used here to impute the late responses. In addition, Fu and Manner (2010) proposed an integrated two-component prediction (ITP) longitudinal model for this scenario. Their approach uses all the available early and late data at the interim instead of only imputing the missing final responses. Though DRPM methods have been developed to link early and late responses, however, none of the models response has been specifically applied to DAS28 - the typical primary endpoints in Phase 2a RA clinical trials.

In this section, a predictive model utilizing the linear relationship between early and late response is proposed and evaluated in the context of the PoC study design for follow-on compound GSK654321. Specifically, the topics to be discussed are:

1. the type of data in the decision-making at interim analysis;
2. the mathematical details of the proposed delayed response predictive model (DRPM) and consequently;
3. comparison of proposed model vs last observation carry forward (LOCF) method with regard to the estimated treatment effect, probability of stopping for futility, and bias.

### Illustrated Delayed Response Predictive Models

Figure 8.2 presents a diagram illustrating individual patient data collection and projection of Month 2 (Day 56) data based on Week 2 (Day 14) response. The primary endpoint of the PoC study is DAS28 change from baseline at Month 2 Day 56. The patients’ data at interim analysis may come from multiple sources:

1. Patients who have complete DAS28 data at Day 56,
2. Patients who have complete DAS28 data up to Day 14 but don’t have Day 56 data, and

3) Patients who have dropped out for various reasons.

The proposed DRPM method combines the predicted DAS28 response at Day 56 from sources (#2) and those patients who have complete DAS28 data at Day 56 from sources (#1).

The diagram illustrates individual patient data collection and projection of Month 2 (Day 56) data based on Week 2 (Day 14) response. All subjects in the same cohort have DAS28 assessment at baseline, Day 14 and Day 56. Depending on the rate of enrolment, the subjects are enrolled sequentially. The vertical red line denotes interim analysis (IA) being conducted when early time point at Day 14 from all subjects are collected. At the interim analysis, the first three subjects have Day 56 assessment completed and the other two subjects only have day 14 assessment completed. The proposed DRPM method combines those patients who have complete DAS28 data at Day 56 and the predicted DAS28 response at Day 56 from those patients who have assessment of Day 14 available but do not have Day 56 data at interim for futility or efficacy decisions. Therefore, there is no additional follow-up required for the last patient and a time saving of 6-week is reduced for this interim analysis and for future planned interim analysis or dose escalation analysis.

The potential time-savings for the proposed design are presented in Table 8.4. A maximum of 294 days could be saved in the 8-cohort study design of phase 2a PoC design with 7 between cohort dose escalation analysis. The DRPM uses all the available Day 56 response based on accumulating data and predicting Day 56 response from Day 14, in a cohort design of GSK123456 or GSK654321, there is a chance that the trial would end up with a success for those trials that stopped based DRPM model, therefore decision criteria are needed to be carefully selected to minimize those chance of “false negative”. There is still a chance of “false positive” if a success stopping criteria is implemented, which is not covered in this dissertation.

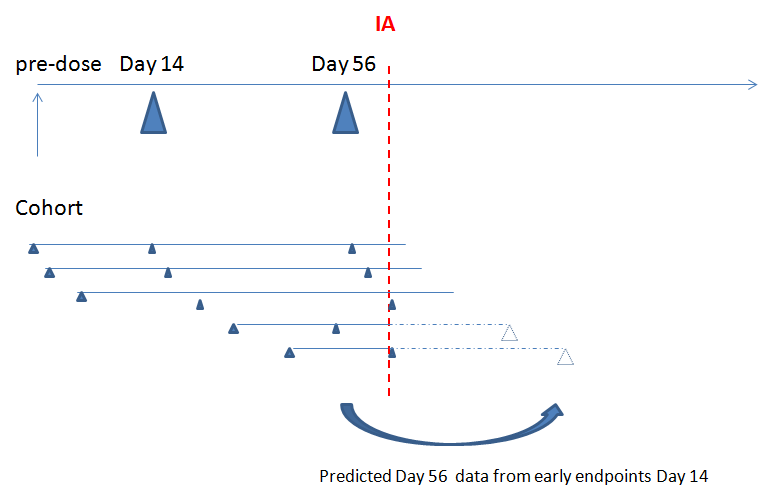


Figure 8.2. Diagram illustrating individual patient data collection and projection of Month 2 (Day 56) data based on Week 2 (Day 14) response. The vertical red line denotes the interim analysis (IA) being conducted when early time point at Day 14 from all subjects are collected. The solid triangle is the available observed data and the dotted triangle is the late day 56 data to be collected. The predicted data will be pooled with all the available data in the previous cohort of the time point for the interim analysis.

### Delayed Response Predictive Model: Mathematical Details

An analytic formula provides a way to apply the delayed response method without re-running the statistical analysis. The proposed model assumes that all patients are independent of each other.

DAS28 change from baselineis denoted as **Δ**Yijt for the jth subject in the ith treatment arm at time t, where t is 1 at early time point (i.e. Day 14) or 2 at late time point (i.e. Day 56), i is 1 in placebo arm or 2 in treatment arm. **Δ**Yijt follows a normal distribution with mean µit and variance σ2it, **Δ**Yijt ~ N(µit, σ2it).

For example, for the 3rd subject on the placebo arm, DAS28 change from baseline at early time point 1is denoted as **Δ**Y131, which follows *N*(µ11, σ211) distribution, while DAS28 change from baseline at late time point 2is denoted as **Δ**Y132, which follows N(µ12, σ212) distribution. Without loss of generality, assume there are ni subjects in each arm, and **Δ**Yijt data are complete for j=1,…, m, and **Δ**Yij2 are missing for j=m+1, …, n.

The traditional method will only use subjects with complete data to perform data analysis, i.e., only the first m subjects are used for each arm. This is also called “complete-case” analysis (Gelman 2007). This would result in power reduction and bias in the estimate (Gelman 2007). The late response data become **Δ**Yi12, **Δ**Yi22, …, **Δ**Yim2

#### Last Observation Carry Forward (LOCF) method

The LOCF method will carry over the last observation which is the response at early time point **Δ**Yij1 to impute the response at late time point **Δ**Yij2. The LOCF late response data are **Δ**Yi12, **Δ**Yi22, …, **Δ**Yim2, **Δ**Yi(m+1)1, …, **Δ**Yi(n)1 with mean

and variance

#### Delayed Response Predictive Model 1 (DRPM1)

The DRPM 1 approach uses the mean early response µi1 to predict the mean late response µi2 and impute **Δ**Yij2 with mean µi2. It has been shown in Chapter 7 from a meta-analysis of 20 studies that the mean late response has a linear relationship with the mean early response in the same study. For the proposed DRPM1, assume that µi2 = αi+βi µi1,where αi and βi are based on the meta-analysis results. So **Δ**Yij2 = αi+βi µi1, here , DRPM1 method late response data are **Δ**Yi12, **Δ**Yi22, …, **Δ**Yim2, **Δ**Yi(m+1)2, …, **Δ**Yi(n)2 with mean

and variance

.

#### Delayed Response Predictive Model 2 (DRPM2)

The DRPM 2 approach adds additional variance by using imputing **Δ**Yij2 by a random variable with mean µi2. So **Δ**Yij2 ~ *N*(αi+βi µi1, βi2 σ2i1), here , the distribution of late response data are normal distribution with

mean

and variance

.

#### Delayed Response Predictive Model 3 (DRPM3)

The DRPM 3 approach predicts future databy adding additional subject variability σ2 to the variance of DRPM2 model. So **Δ**Yij2 ~ *N*(αi+βi µi1, βi2σ2i1+σ2), here σ2 is the variability from the linear relationship of mean early response and mean late response and is obtained from the meta-analysis (Table 7.5). The DRPM method late response data are **Δ**Yi12, **Δ**Yi22, …, **Δ**Yim2, **Δ**Yi(m+1)2, …, **Δ**Yi(n)2 with mean

and variance is

.

As illustrated in the formula, the DRPM3 model has same mean as DRPM1 and DRPM2 but t additional variability was added due to the prediction therefore DRPM3 has wider variance compared to DRPM1 and DRPM2.

### Summary of Approaches

Three DRPM approaches and the LOCF approach were presented in Section 8.3.2. LOCF is a simple method to replace the missing Day 56 response with Day 14 data. Depending on the rate of missing data and model assumptions, this is considered a conservative approach since it may lead to underestimating the true treatment effect. DRPM approaches use the linear relationship from the meta-analysis and replace the missing data with the predicted Day 56 response based on Day 14 data. For each method, the treatment difference can be calculated as the difference of expected treatment and placebo response and can be expressed as the mean of μ22 – μ12 with the variance of as sum of individual variances assuming independence. Normal approximation is then used to derive statistical inference based on the mean and variance.

In summary, three DRPM approaches and the LOCF approach were discussed in this section. Predicted Day 56 response is estimated based on the ANCOVA model parameters (the slope, intercept, and variability) since the predicted Day 56 response is believed to be a function of Day 14 response from the meta-analysis. The performance of the three DRPM approaches and the LOCF approach will be evaluated in next section using simulation.

## Illustration of Delayed Response Predictive Models

DRPM1 is used here in the analysis for expository purposes. The predicted DAS28 at Day 56, the observed DAS28 at Day 56 and the difference (between predicted and observed), and the percentage difference are presented in Table 8.1. The ANCOVA model parameters from meta-analysis of systematic review in Table 7.5 are used in the prediction of mean DAS28 at Day 56 from Day 14 assuming a linear relationship between DAS28 at Day 56 and DAS28 at Day 14 (intercepts of -0.055 and -0.105 the slope of 1.1 and 1.1 for placebo and treatment respectively).

The most difference and percentage change occur in Cohort 1 when the sample size is low and the percentage change is below or equal to 10% after Cohort 1, with the exception of Cohort 3. The time saving of each cohort analysis is 42 days and the total time saving of 7 between-cohort analyses is 294 days compared to the original design at the cost of 10% difference in the true estimate of dose effect at Day 56.

Table 8.1. The predicted DAS28 at Day 56, observed DAS28 at Day 56, the difference, and the percentage difference based on the retrospectively analysis of the PoC study. The analysis by cohort is performed in the same sequential order as in the clinical trial.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cohort** | **Treatment (mg/kg)** | **n** | **DAS28 Change From Baseline** | | | |
| **Predicted mean change in DAS28 scores at Day 56 56**  **from Day 14** | **Observed mean**  **change in DAS28 at Day 56** | **Diff (predicted – observed)** | **Percent Change (%)** |
| 1\* | 0.03 | 3 | -1.88 | -1.76 | -0.12 | 7 |
| 0.3 | 2 | -1.12 | -1.37 | 0.25 | -18 |
| placebo | 2 | -1.47 | -1.29 | -0.18 | 14 |
| 2 | 0.3 | 8 | -1.20 | -1.11 | -0.09 | 8 |
| placebo | 4 | -1.50 | -1.65 | 0.15 | -9 |
| 3 | 3 | 6 | -1.31 | -1.15 | -0.16 | 14 |
| placebo | 6 | -1.18 | -1.21 | 0.03 | -2 |
| 4 | 3 | 12 | -1.41 | -1.56 | 0.15 | -10 |
| placebo | 8 | -1.10 | -1.20 | 0.10 | -8 |
| 5 | 10 | 6 | -1.65 | -1.73 | 0.08 | -5 |
| placebo | 10 | -1.28 | -1.22 | -0.06 | 5 |
| 6 | 30 | 6 | -0.93 | -0.90 | -0.03 | 3 |
| placebo | 12 | -1.06 | -1.04 | -0.02 | 2 |
| 7\* | 10 | 8 | -1.34 | -1.27 | -0.07 | 6 |
| placebo | 13 | -1.00 | -0.98 | -0.02 | 2 |
| 8 | 20 | 6 | -0.88 | -0.83 | -0.05 | 6 |
| placebo | 15 | -1.06 | -1.00 | -0.06 | 6 |

\*1 subject in dose 0.06mg/kg in Cohort 1 was excluded from the analysis. One placebo subject in Cohort 7 was removed due to missing data. Percentage change is calculated as the ratio of difference and observed mean change in DAS28.

The time course plots of DAS28 results at the end of the study are shown in Figures 8.3 (0.3, 3, 10, 20 and 30 mg/kg) and 8.3 (placebo and other doses 0.3, 3, 10, 20 and 30 mg/kg). The treatment is not displayed in the graph if the sample size is less than 4.

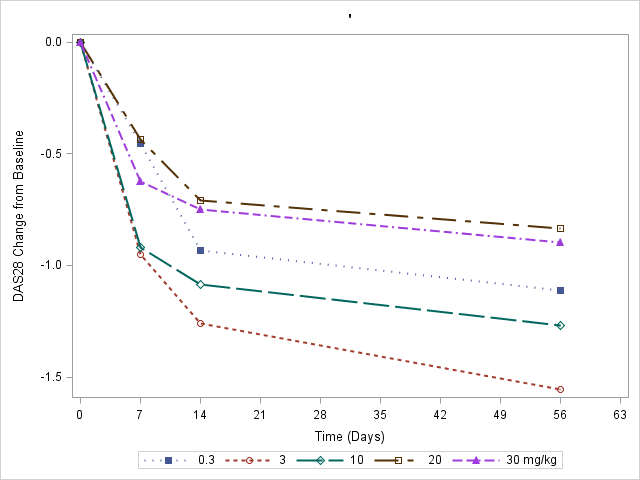


Figure 8.3 Time profile of DAS28 change from baseline for dose 0.3mg/kg, 3mg/kg, 10mg/kg, 20mg/kg and 30mg/kg. The means of each dose group are displayed at Day 0, Day 7, Day 14, and Day 56 time points. For demonstration purpose, the error bar was not shown in the figure.

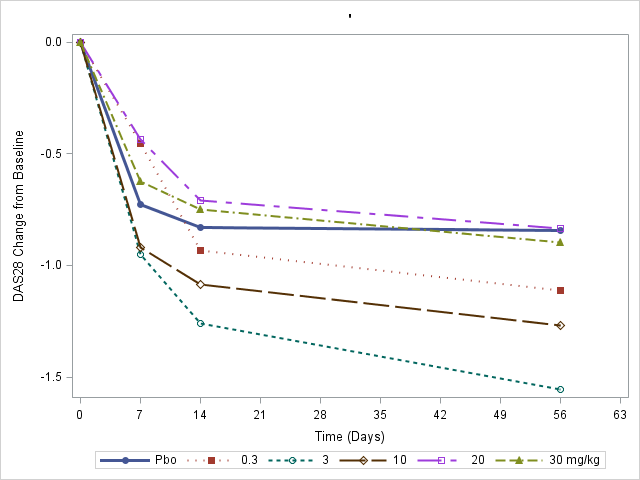


Figure 8.4 Time profile of DAS28 change from baseline for placebo, dose 0.3mg/kg, 3mg/kg, 10mg/kg, 20mg/kg and 30mg/kg. The means of each dose group are displayed at Day 0, Day 7, Day 14, and Day 56 time points. For demonstration purpose, the error bar was not shown in the figure.

It is evident that the slope of Day 14 and Day 56 in corresponding doses is relatively parallel in Figure 8.3 which is consistent with the linear relationship from the literature. The slope of Day 14 and Day 56 in placebo is different from that in study treatment (Figure 8.4).

### Summary of Retrospective Model Fitting in PoC Study

It was shown in the retrospective cohort analysis of PoC study using DRPM1 that the percentage changes of predicted Day 56 and observed Day 56 were within 10% in most cohort predictions. The average DAS28 change from baseline is -1.11 across all treatment groups and the percentage change of 10% is between 1.00 and 1.22, after cohort 4, the maximal percentage change is between 2% and 6%, therefore the deviation of percentage change 10% may have an impact in early cohorts but is unlikely to impact the interim decision of later cohorts.

If the DRPM1 method were successfully implemented in the PoC study using Day 56 as interim decisions, great time saving is anticipated. DRPM is a potential solution to achieve the time saving if projected Day 56 based on early read-outs is used in the clinical trial.

In next section, the different predictive models will be assessed using simulation in the context of a Phase 2a PoC design for follow-on compound GSK654321 to predict DAS28 response at Day 56 based on Day 14.

### Delayed Response Predictive Model: Simulation

#### Data Simulation

The longitudinal responses (Day 14 and Day 56) of the targeted 64 patients in 8 cohorts in the context of the Phase 2a PoC study of GSK654321 are simulated. In each cohort, out of a total of 8 patients, 2 patients receive placebo and each of the other 6 patients receive 0.03, 0.3, 3, 10, 20 or 30 mg/kg of follow-on compound GSK654321. It is assumed in the simulation that the data are reviewed by the internal safety review committee between each cohort after dosing since most of the adverse events occur within 48 hours of injection. DAS28 of each patient will be measured at baseline, Day 14 and at Day 56. A bivariate Normal distribution is used to simulate the Day 14 and Day 56 response and the true DAS28 response at Day 56 follows a Normal distribution with the mean as a linear function of Day 14 response (Table 8.2). The bivariate Normal distribution is illustrated in (8.1).

 8.1

**with**  **,**

**.**

Where µi1 and µi2 is mean of DAS28 change from baseline at Day 14 and Day 56 respectively. It is assumed that DAS28 change from baseline at Day 56 is linearly correlated with that at Day 14 with slope as βi and intercept as αi, i = 1 Placebo 2 treatment of GSK654321. ρ is the correlation coefficient between DAS28 change at Day 14 and DAS28 change at Day 56 response. σ1 and σ2are the standard deviation of DAS28 change from baseline at Day 14 and Day 56 respectively.

The simulations are anchored in line with scenarios for GSK654321 given the PoC study design. Safety reviews and interim analysis are planned between every cohort. For demonstration purposes, there are two planned interim analyses in this simulation. The first will occur after 3 cohorts of patients who complete the assessment of DAS28 at Day 14, and the second interim analysis takes place after 6 cohorts of subjects complete the assessment of DAS28 at Day 14. At the time of the interim analysis, DAS28 results at Day 56 from past cohorts are available so they can be pooled with predicted Day 56 data from Day 14 in the analysis. All the parameter assumptions for the data simulation are shown in Table 8.2.

Table 8.2 Assumed values of bivariate normal distribution in the simulation (A) and study design in the proposed assessment of DRPM model at Interim and Final analysis (B)

A: The assumed values of bivariate normal distribution

|  |  |
| --- | --- |
| **Parameters** | **Value assumed in the simulation** |
| µ1:mean DAS28 of Day 14 response | true effect mean at Day 14 |
| µ2:mean DAS28 of Day 56 response | Linearly correlated with that at Day 14 with slope as βi and intercept as αi. |
| βi: Slope of linear relationship of Day 14 and Day 56 | 0.5, 0.75, 1.0, 1.1, 1.2 |
| αi: Intercept of linear relationship of Day 14 and Day 56 | -0.443 for placebo, -0.493 for biological treatment |
| σ1: Standard Deviation of Day 14 response\* | 0.9 |
| σ2: Standard Deviation of Day 56 response\* | 0.6, 0.9, 1.2 |
| ρ: Correlation coefficient between Day 14 and Day 56 response | 0.1, 0.3, 0.5, 0.7, 0.9 |

B: Study design in the proposed assessment of DRPM model at Interim and Final analysis

|  |  |
| --- | --- |
| Interim and Final analysis | Study Design in the Proposed Assessment of DRPM model |
| 1st interim analysis | After the 3rd cohort completes assessment of DAS28 at Day 14 and the past cohorts have Day 56 assessment available. Total 24 patients including 6 patients in placebo, 3 patients each in GSK654321 0.03, 0.3, 3, 10, 20, and 30 mg/kg respectively. |
| 2nd interim analysis | After the 6th cohort completes assessment of DAS28 at Day 14 and the past cohorts have Day 56 assessment available. Total 48 patients including 12 patients in placebo, 6 patients each in GSK654321 0.03, 0.3, 3, 10, 20, and 30 mg/kg respectively. |
| final interim analysis | All 64 patients in 8 cohorts complete Day56 assessment and analysis are based on Day56 response including 16 patients in placebo, 8 patients each in GSK654321 0.03, 0.3, 3, 10, 20, and 30 mg/kg respectively. |

\* The slope (βi) and intercept (αi) are from ANCOVA model parameter Table 7.5. In base case scenario, the data are simulated assuming parameters are ρ=0.7 =0.9, =0.9,α1=-0.055, α2=-0.105, β1= β2=1.1.

As well as the based case scenario, there are three additional sensitivity analyses undertaken to evaluate the model performance as described below while keeping other simulation parameters as fixed:

1. Evaluate the DRPM1, 2, 3 and LOCF model performance in comparison to true model when the variance is lower than expected (or higher than expected ( while fixing other parameters as constant. The true model analyses the data of DAS28 at Day56 and there is no imputation in the true model.
2. Sensitivity analysis of DRPM1, 2, 3 and LOCF models when correlation coefficient is lower than expected (ρ=0.1, 0.3, and 0.5) or higher (ρ=0.9). In base scenario, the correlation coefficient is 0.70.
3. In addition, we expect that there is relationship of early and late response and the slope β is 1.1 in base scenario, in case the relationship of late and early response are lower than expected (β=0.5, 0.75 and 1.0) or higher than expected (β=1.2), the performance of DRPM models and LOCF are compared while keeping the test of parameters as fixed.

DRPM models 1, 2, and 3 and the LOCF approach are applied in the prediction of Day 56 and the design scenarios and decisions are discussed in the next section.

#### Design Scenarios and Decisions

There are two simulation scenarios to evaluate the predictive DRPM and LOCF approaches.

*Profile 1:* Under alternative hypothesis assuming the DAS28 change from baseline at Day 14 response of 6 doses levels (0.03, 0.3, 3, 10, 20 and 30 mg/kg) and placebo follow an *Emax* like curve as dashed red line in Figure 8.5. DAS28=-0.5-1.4\*Dose/(2.0+Dose)+ε

*Profile 2:* Under null hypothesis assuming the DAS28 change from baseline at Day 14 response of 6 doses levels (0.03, 0.3, 3, 10, 20 and 30 mg/kg) and placebo follow a flat curve as blue line in Figure 8.5. Flat curve: DAS28=-0.5+ε

Figure 8.5 Dose-response profiles used in the simulation. The model profiles include placebo like flat curve which is denoted as a blue line and is fixed at -0.5 for all dose levels, and dose proportional *Emax* like model shown as a red dashed line.

To facilitate the evaluation of design performance, the decision criteria are based on conditional power (Spiegelhalter, 2004) as follows:

1. Futility stopping criteria is that conditional power based on final sample size given the treatment effect as observed from interim analysis is less than 20%.
2. There is no stopping for success.

Criteria of 20% of conditional power is chosen because they give reasonable adaptive design operating characteristics such as type I error rate and power in proposed clinical trial design with the planned sample size. Conditional power is defined as the probability of a statistically significant result at the end of the trial given the data observed so far and the data yet to be observed assuming the same effect as was observed (Spiegelhalter, 2004). Spiegelhalter’s conditional power formula (formula 3.3; Spiegelhalter, 2004) is modified and used as decision rules for futility stopping at interim 1 and interim 2 with unequal sample size (Dmitrienko, 2007).

To further evaluate the design performance, the following results are reported at 1st and 2nd interim analysis:

1. The model based average treatment effect and standard deviation from DRPM 1, 2 and 3, LOCF and true. The true method uses only the late Day 56 response in the analysis without any imputation.
2. The percentage of trials that stopped for futility (conditional power <20%) and the total percentage of trials that stopped for futility
3. Bias of average treatment effect.

The average effects are reported for each model in this simulation. The average effect is defined as the DAS28 difference of mean of all 6 doses and mean placebo response.

The average effect method is analogous to the hierarchical testing procedure (also called gatekeeping procedure) (Dmitrienko, 2003; 2006) in which the statistical tests are performed in a sequential way to control the type I error rate in the dose response analysis. The average effect is tested first at 5% significant level, if the P-value of the average effect is less 5%, the patient’s DAS28 data at selected doses are tested at 5% level. For demonstration purposes, only the first hypothesis test of average effect is reported as decision making at interim analysis in this simulation. At final analysis, the trial will be significant if and only if both P-values of average effect and any of dose effects are significant. No other adaptations to aspects of the study design are adopted in the simulation.

All evaluations are based on 10,000 simulations, which provide sufficient precision (with a standard error of less than 0.001) for the sample estimate. All simulations are run in SAS 9.3. SAS codes to support the simulation are included in Appendix 11.1.4 and will be available for download.

#### Simulation Results

The model based treatment effect and standard deviation from DRPM 1, 2 and 3, LOCF and true data under the assumption of the *Emax* model are presented in Table 8.2. Within the parameters of the simulation, the model based treatment effects are -0.932, -0.932, -0.934 and -0.889 for DRPM 1, 2, 3 and LOCF approaches respectively at the first interim analysis. The true mean treatment effect is -0.934 so the bias is 0.002, 0.002, 0 and 0.045 for DRPM 1, 2, 3 and LOCF approaches respectively. The effect estimates from the DRPM 1, 2, and 3 approaches are unbiased which is in line with the analytical approach. The LOCF method replaces the missing Day 56 response with Day 14 and therefore under-estimates the treatment effect by 5% in the simulation. All the treatment estimates of DRPM 1, 2, and 3 models were unbiased and the mean treatment effect of LOCF under-estimated the true effect under both the null hypothesis and alternative hypothesis.

Within the parameters of the simulation, among the three DRPM approaches, the standard deviation of DRPM2 is 0.895 which is similar to the true effect (0.886) under the current assumption, the DRPM1 approach under-estimates the variance by replacing the missing Day 56 values with mean predicted Day 56 while the DRPM3 approach over-estimates the variance. Percentages of stopping for futility are 11.5%, 16.8%, 20.6%, 13.6% and 15.2% for DRPM 1, 2 and 3, LOCF approach, and true data respectively. Therefore, of the 4 approaches, DRPM2 was the best predictive model in predicting the treatment estimate and standard deviation within the parameters of the simulation. DRPM2 approach also had reasonable operating characteristics in the context of the proposed Phase 2a PoC design for GSK654321.

The 2nd interim analysis is performed for the trials not meeting the futility stopping criteria (Table 8.2). The biases of average treatment effects are 0.017, 0.009, 0.005 and 0.032 for DRPM 1, 2, 3 and LOCF approaches respectively, within the parameters of the simulation. DRPM2 and DRPM3 further show the least biases compared to the true data. In the final analysis, the final success and total futility for DRPM2 and DRPM3 are final success: 74.6% and 69.3% and total futility: 24.5% and 29.9% respectively. DRPM2 approach had reasonable probability of futility stopping, the statistical power and type I error comparing to true model in the context of the proposed Phase 2a PoC design for GSK654321.

The average sample sizes are 58, 56, 54, 58, and 57 for DRPM 1, 2, 3, LOCF, true model respectively and is comparable for all the models being evaluated (Table 8.2).

#### Type I error rate

The treatment effect and standard deviation from DRPM 1, 2 and 3, LOCF and true data under null hypothesis (flat dose response curve) are presented in Table 8.3. Within the parameters of the simulation, the biases of the treatment effects are 0.002, 0.002, 0, and 0.009 for DRPM 1, 2, 3 and LOCF approaches respectively at the first interim analysis. The true effect estimates from the DRPM 1, 2, and 3 approaches are unbiased which is in line with the analytical approach in Section 8.3.2. The LOCF method under-estimates the treatment effect by 2% in the simulation. Similar to the alternative hypothesis, of the three DRPM approaches, the standard deviation of DRPM2 is 0.895 which is closest to the true effect (0.886). The DRPM1 approach under-estimates and the DRPM3 approach over-estimate the variance. Percentages of stopping for futility are 83.8%, 83.9%, 84.7%, 85.5% and 84.9% for DRPM 1, 2 and 3, LOCF approach, and true data respectively.

The 2nd interim analysis is performed for the trials not meeting the futility stopping criteria (Table 8.3). Within the parameters of the simulation, DRPM2 and DRPM3 show the least biases compared to the true data. In the final analysis, the final success for DRPM2 and DRPM3 are 1.1% which is similar to the true data (1.5%). The predicted DRPM model didn’t inflate the type I error under the simulation assumptions. The average sample size is comparable under all models being evaluated (Table 8.3).

Table 8.2 Estimated treatment effect (treatment-placebo) and the interim futility probability of predicted model (the prediction of DAS28 at Day 56 based on Day 14), LOCF method and true data from Day 56 (Month 2) assuming data follow an *Emax* curve under alternative hypothesis. This is the base case and the data are simulated assuming parameters are ρ=0.7 =0.9, =0.9,α1=-0.055, α2=-0.105, β1= β2=1.1 for placebo and treatment respectively.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.932 (0.806) | -0.932 (0.895) | -0.934 (0.967) | -0.889 (0.816) | -0.934 (0.886) |
| % Futility based on Cond. Power (<20%) | 11.5% | 16.8% | 20.6% | 13.6% | 15.2% |
| Bias | 0.002 | 0.002 | 0 | 0.045 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -0.975 (0.852) | -0.983 (0.897) | -0.987 (0.935) | -0.960 (0.855) | -0.992 (0.887) |
| % Futility based on Cond. Power (<20%) | 5.8% | 7.7% | 9.3% | 6.0% | 5.7% |
| Bias | 0.017 | 0.009 | 0.005 | 0.032 | - |
| Final Analysis | | | | | |
| **Final Success (P-value <0.05)** | 81.5% | 74.6% | 69.37% | 79.5% | 78.4% |
| **Final Failure (P-value >=0.05)** | 1.1% | 2.4% | 0.8% | 0.9% | 0.7% |
| **Total Futility at Interim Analysis** | 17.3% | 24.5% | 29.9% | 19.6% | 20.9% |
| **Average Sample Size** | 58 | 56 | 54 | 58 | 57 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

Table 8.3 Estimated treatment effect (treatment-placebo) and the interim futility probability of predicted model (the prediction of DAS28 at Day 56 based on Day 14), LOCF method and true data from Day 56 (Month 2) assuming data follow a *flat* curve under null hypothesis. This is the base case scenario assuming flat curve and the data are simulated assuming parameters are ρ=0.7 =0.9, =0.9,α1=-0.055, α2=-0.105, β1= β2=1.1 for placebo and treatment respectively.

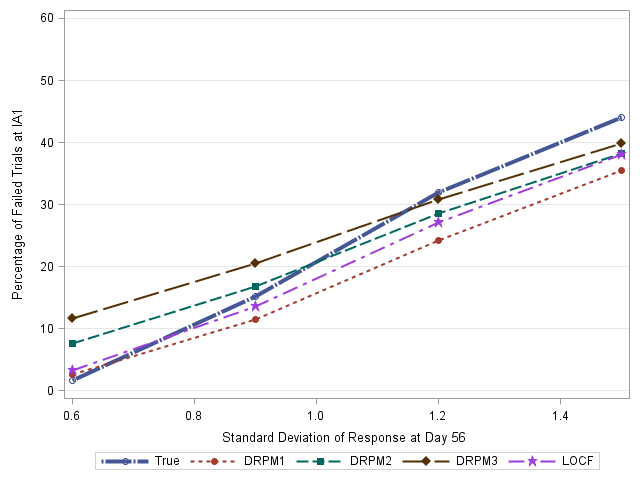
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Null Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.052 (0.809) | -0.052 (0.895) | -0.054 (0.967) | -0.035 (0.811) | -0.054 (0.886) |
| % Futility based on Cond. Power (<20%) | 83.8% | 83.9% | 84.7% | 85.5% | 84.9% |
| Bias | 0.002 | 0.002 | 0 | 0.009 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -0.332 (0.850) | -0.305 (0.894) | -0.301 (0.936) | -0.334 (0.850) | -0.368 (0.881) |
| % Futility based on Cond. Power (<20%) | 13.5% | 13.7% | 13.0% | 12.1% | 12.4% |
| Bias | 0.016 | 0.049 | 0.059 | 0.027 | - |
| Final Analysis | | | | | |
| **Final Success (P-value <0.05)** | 1.4% | 1.1% | 1.1% | 1.3% | 1.5% |
| **Final Failure (P-value >=0.05)** | 1.3% | 1.2% | 1.2% | 1.2% | 1.2% |
| **Total Futility at Interim Analysis** | 97.3% | 97.6% | 97.7% | 97.6% | 97.3% |
| **Average Sample Size** | 28 | 28 | 28 | 28 | 28 |

The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analysis. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

#### Sensitivity analysis

The simulations undertaken so far are only for anticipated slopes, correlations and variances for the GSK654321 based on retrospective data. Additional sensitivity analysis are now conducted to investigate the DROM1, 2, 3 and LOCF model performance comparing to true model, in the case of assumption violation when the variance of late response lower or higher than expected (Figure 8.6), the correlation coefficient of late and early response decreases (Figure 8.7), and slope of late and early response is lower than expected (Figure 8.8).

Here the simulations are performed by only varying one parameter of interest and fixing other parameters. Additional supplemental data Tables (Table 11.2-11.4) are presented in Appendix 11.4 which supports Figures 8.6 to 8.8.



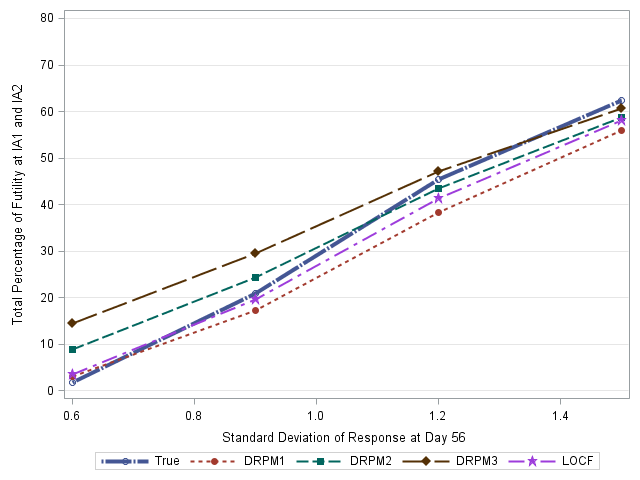
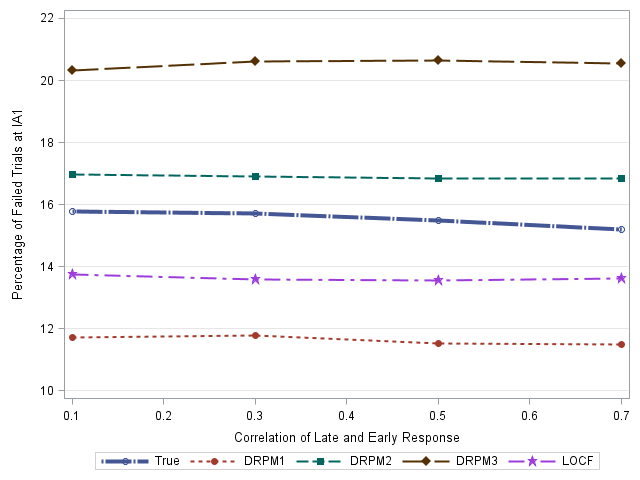


Figure 8.6 Sensitivity of futility at IA1 (top) and IA1/IA2 (bottom) for DRPM1, 2, 3, LOCF, and true model when the variability of late response is lower (=0.6) or higher than expected (=1.2 or 1.5). Other parameters (ρ=0.7 =0.9, α1=-0.055, α2=-0.105 and β1=β2=1.1) are fixed in the simulation.

From the simulations it seems that the percentage of trials stopping for futility as variability in late response increases. In addition, DRPM3 seems to perform better and is closer to the true model in for futility stopping at IA1 and IA1/IA2 when the variability is higher than expected (Figure 8.6). DRPM2 performs better when the variability is lower than expected. The bias is relatively small for DRPM2 and DRPM3 models while there is higher bias using LOCF method. The average sample size is comparable between DRPM2/DRPM3 and LOCF methods if the variability is lower or higher than expected (Appendix Table 11.2).

It is shown from the simulations that the impact on models in case of lower correlation coefficient between Day 14 and Day 56 response is relatively small. When the correlation is lower than expected between Day 14 and Day 56 response the impact on the models is not great. For ρ=0.1 the percentage of stopping for futility at the first interim analysis are 11.7%, 17%, 20.3%, and 13.8% respectively, which are similar to in base scenario (Figure 8.7 and Appendix Table 11.2). Other results are similar to base scenario. Within the parameters of the simulation of the 4 approaches, DRPM2 and DRPM3 model are the best predictive models in terms of: the futility at the interim analysis; the bias of the treatment estimate and the standard deviation. Similar results were observed when the correlation coefficient is less than expected (ρ =0.3 and 0.5) or higher than expected (ρ =0.9). The full results are given in Appendix Table 11.3.



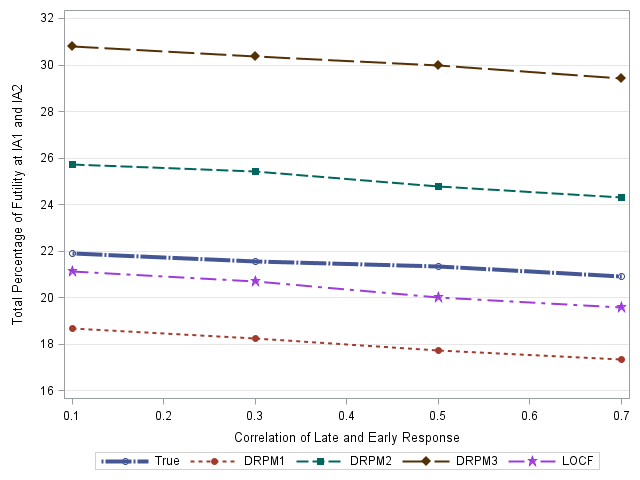


Figure 8.7 Sensitivity of percentage of futility at IA1 (top) and IA1/IA2 (bottom) for DRPM1, 2, 3, LOCF, and true model when correlation of late and early response are lower than expected (ρ=0.1, 0.3, 0.5 and 0.7). Other parameters are fixed in the simulation =0.9, =0.9,α1= -0.055, α2=-0.105 and β1=β2=1.1 for placebo and treatment respectively.

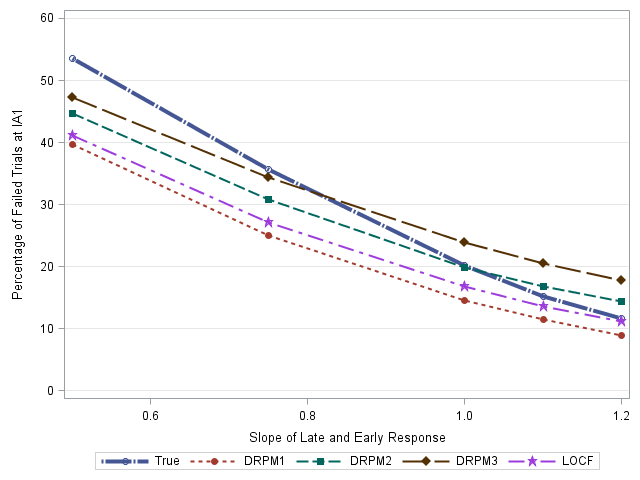
The fact that the approaches worked well even for low correlations was unexpected. An explanation could be that though there is little correlation between Day 14 and Day 56 there is still strong correlation between the dose and outcome. The simulations so far have only been done varying a single parameter. These will now be extended by varying both the dose and the correlation.

To further understand the examine the predictive model performance, further simulations were undertaken to explore the case when the slope of true late response and early response is much lower than expected (β1= β2= 0.5, 0.75, 1.0) as well higher than expected when β is 1.2 but the correlation coefficient is fixed 0.7.

For these extended simulations the percentage of futility stopping increases as the slope increases at the first interim analysis and combination of IA1 and IA2 (Figure 8.8). When the slope of linear relationship of DAS28 between Day 56 and Day 14 is much lower than expected (β1= β2=0.5), all the proposed models over-estimate the true effect at 0.158, 0.158, 0.160 and 0.115 for DRPM1, DRPM2, DRPM3 and LOCF model respectively (Appendix Table 11.4). The LOCF method seems to be better model in this scenario. At the second interim analysis the bias of DRPM1, 2, 3 reduces to 0.027, 0.027, and 0.024. When slope is lower than expect (=0.5, 0.75 or 1.0), all the proposed models over-estimate the true effect and have comparable percentage of futility stopping and final success to true model. The average sample size savings are also similar for all models.

From the simulations for the case when the Day 56 response is greater than expected (β1= β2=1.2), all the proposed model under-estimated the true effect at 0.0028, 0.0029, 0.0026 and 0.0072 for DRPM1,2,3 and LOCF model respectively at the first interim analysis (Appendix Table 11.4). The bias observed is higher for the LOCF model. It also shows that the percentage of stopping was higher in DRPM2 and DRPM3 model than true effect model. In both scenarios, more bias is observed. The DRPM models perform the best when the relationship of late and early is similar to the systematic literature. Both DRPM2 and DRPM3 models perform better when DAS28 response is greater than expected and LOCF model is better when the slope of linear relationship is 1. The full results of this simulation are given in Appendix Table 11.4.

To further evaluate the impact of the predicative model where the model may give a poor prediction, sensitivity analysis is conducted when slope of linear relationship of DAS28 between Day 56 and Day 14 is much lower or higher than expected (β1= β2=0.5, 0.75, 1.0, and 1.2) and correlation coefficient is fixed at 0.1 (ρ=0.1). The results of the sensitivity analysis (Appendix Table 11.5 and Figure 11.7) are similar to the other sensitivity results (Figure 8.8) that the DRPM3 performs better when the slope is lower than expected (β1= β2=0.5 or 0.75) in term of percentage of futility stopping at the first and second interim analysis with higher bias observed. The bias is -0.158, -0.157, -0.159, and -0.114 for DRPM1, 2, 3, and LOCF models respectively. DRPM2 and LOCF models slightly perform better when the slope is greater than expected (β=1.2). Type I error rates were not inflated for this evaluations within the parameters of the simulation (Appendix Table 11.6).



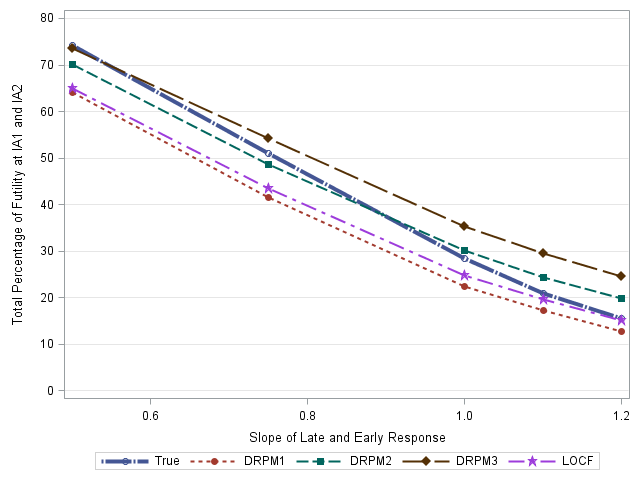


Figure 8.8 Sensitivity of percentage of futility IA1 (top)and IA1/IA2 (bottom) for DRPM1, 2, 3, LOCF, and true model if the slopes of late and early response are lower or higher than expected (β1= β2=0.5, 0.75, 1.0 and 1.2). Other parameters are fixed in the simulation ρ=0.7 =0.9, =0.9,α1=-0.055, α2=-0.105 for placebo and treatment respectively.

In summary, the percentage of futility stopping at the first and second interim analysis and average sample size are generally comparable in the sensitivity analysis being evaluated and the bias is larger (over-estimation or under-estimation) when Day14 outcome is not a good predictor of Day 56 outcome.

### Summary of Simulations

All the treatment estimates of DRPM 1, 2, and 3 models were unbiased and the mean treatment effect of LOCF under-estimated the true effect under both the null hypothesis and alternative hypothesis for the base case scenario. Of the 4 approaches, DRPM2 was the best predictive model in predicting the treatment estimate and standard deviation within the parameters of the simulation. DRPM2 approach also had reasonable operating characteristics in the context of the proposed Phase 2a PoC design for GSK654321.

In case of the assumptions for the variance are violated, DRPM3 performed better when the variability is higher than expected (Figure 8.6). DRPM2 performs better when the variability is lower than expected (Figure 8.6). Sensitivity analysis were conducted and it was shown the DRPM2 and DRPM3 model performed well when there are deviation in the assumptions in change in late response and correlation coefficients (Figure 8.7). However, if the slope of late and early response is lower or higher than expected, all evaluated models over-estimated or under-estimated the true effect. All models have comparable percentage of futility stopping and final success to true model.

Overall, the DRPM1, DRPM2, and DRPM3 approaches all gave unbiased estimates of treatment effects for Day 56 for the base case model and sensitivity analysis. The DRPM2 outperformed other approaches under the current simulation assumptions. Among the estimates of the true effect, the LOCF method under-estimated the true effect by 5% in base scenario.

Within the parameters of the simulations it seems that the DRPM models can be applied to future PoC dose finding design of GSK654321 by using predicted Day 56 data in the decision making. The method can be further extended to be used in conjunction with the adaptive dose allocation and Bayesian NDLM approach, as illustrated in Chapter 6.

## Time-saving with the Predictive model

In conventional RA designs, the interim analyses based on Day 56 data might run for years. In this chapter, it was shown using simulation that a DRPM approach can be implemented in the proposed PoC design of GSK654321. The time-savings resulting from the predictive model to predict DAS28 at late time points from early time points are presented in Table 8.4. Depending on the study design, the average time-savings range from 98 days to 392 days.

There is potential for a significant gain from additional years of sales recognized if the drug is successfully launched into the global RA market earlier. This highlights the importance of using an adaptive design in RA clinical trials and the DRPM model in the prediction of late response using early data to expedite the decision-making process.

Table 8.4 The time-saving for proposed study design using DRPM methods: 1) under a PoC Part A design with 8 cohorts in the dose finding/escalation phase or 2) under the PoC study design with two interim analyses.

|  |  |  |
| --- | --- | --- |
| Design | Prediction  Late endpoint from early time point | Average Time Saving (Days) |
| PoC dose escalation finding design with 8 cohorts | Month 1 (Day 56) from Week 2 (Day 14) | 98 |
| Month 2 (Day 56) from Week 2 (Day 14) | 294 |
| Month 2 (Day 56) from Month 1 (Day 28) | 196 |
| Month 3 (Day 84) from Month 1 (Day 28) | 392 |
| PoC design with two interim analyses | Month 1 (Day 56) from Week 2 (Day 14) | 28 |
| Month 2 (Day 56) from Week 2 (Day 14) | 84 |
| Month 3 (Day 84) from Week 2 (Day 14) | 140 |
| Month 2 (Day 56) from Month 1 (Day 28) | 56 |
| Month 3 (Day 84) from Month 1 (Day 28) | 112 |

## Summary

In this chapter, the DRPM approaches were developed to predict the late responses of DAS28 based on early data and it was shown that DRPM approach provides a simple and alternative way of data imputation to predict DAS28 at late response using early response by associating the response with early and late time point and the DRPM models could be applied to the Phase 2a PoC design in the clinical development of GSK654321.

It was demonstrated how DRPM1, DRPM2, and DRPM3 approaches all gave unbiased estimates of treatment effects within the parameters of the simulation undertaken for the anticipated responses for the study. The DRPM2 outperformed other approaches under the current simulation assumptions. Among the estimates of the true effect, the LOCF method under-estimated the true effect by 5%.

The simulations also showed that a DRPM approaches can be implemented in the proposed PoC design for the new compound GSK654321. The trial duration could potentially be significantly reduced if the interim decision and adaptation are based on earlier time points. DRPM2 and DRPM3 approach also had reasonable operating characteristics in the context of the proposed Phase 2a PoC design for GSK654321.

Analogous research has been conducted to predict the clinical response of ACR20 and ACR50 using earlier time points (Nixon, 2009) in RA Phase 3 clinical trials. The predictive method of this research (Nixon, 2009) was based a binary endpoint and shares the similar approach to this research regarding using short term endpoint to predict the long term endpoint.

It is acknowledged that a comparison based on day 56 on all data will have an advantage on power, with no additional bias introduced and Type I error rate. However, the significant time savings to the overall compound development can be achieved if a decision (about whether or not to proceed with further development of the new compound) can be made using earlier time point, which will expedite the drug development.

During the simulation, the proposed prediction methods were evaluated using an average effect with a hierarchy model, though it is expected that potential time savings could also be achieved through the application of methods to other appropriate designs. The basis of the model parameters estimates assumptions were from a systematic meta-analysis. These estimates could potentially be subject to publication bias. When there were deviations in the assumptions used in the simulations questions were raised with respect to issues such as bias for the different modelling approaches.

It is also acknowledged that enrolment rate may have an impact on design performance, i.e. there are more predicted late DAS28 response and less actual response at Day 56 at fast enrolment rate. When the enrolment rate is slower, there are more actual day 56 responses, which will have more accurate estimate of treatment effect. The enrolment rate was a major factor in Part B of the original Phase 2a study for GSK123456 as to was too quick for this part of the study to be truly adaptive. Using an early time point could also mean decisions are made before enrolment is completed.

In addition, the proposed approach was tested using simulation using average effect methods as first step of hierarchical comparisons followed by pairwise comparison at each dose level. The benefit and disadvantages of such approaches will need to be discussed carefully with clinical team. With the proposed methods described, and comparing the design of GSK123456 to GSK654321, there could be a potential time-saving of up to 294 days for an 8-cohort design if Day 14 (Week 2) data are used.

The methods described were anchored in am RA example with the simulations and results presented only applicable to the case study which motivated our work. The methods proposed, however, and our methods of evaluation, could be generalised to other clinical trials to offer a solution to expedite drug development.

In Chapters 9, the research questions at the beginning of dissertation will be reviewed and a Phase 2a PoC study design of GSK654321 will be recommended.

# Discussions and Proposed Phase 2a Design for GSK654321

An ongoing and serious challenge facing the pharmaceutical industry is the high failure rate in the late phases of drug development (Arrowsmith, 2011; Hay, 2014). It has been reported that approximately 80% of Phase 2 and 50% of Phase 3 clinical trials fail (Arrowsmith, 2011). Lack of efficacy resulting from poor decision making in Phase 2 trials is the main cause of Phase 3 failures (Arrowsmith, 2011; Hay, 2014). Therefore, more robust Phase 2 clinical trial designs, and Phase 2a trials in particular in context the of this PhD, are key to improving the Phase 3 success rate and increasing research and development (R&D) productivity.

Effective treatments for RA disease are relatively new, within the last 20 years. After the development of large molecular biological drugs such as anti-TNF, a few more effective medicines have been discovered and marketed. Most clinical trial designs are standard fixed designs with fixed sample sizes and no interim analyses. The implementation of flexible designs into RA Phase 2a clinical trials is challenging.

1. The drug response or drug effect may not follow a dose proportional increase or decrease in RA clinical trials. The dose response of GSK123456 follows a U-shaped curve as has been demonstrated earlier in this dissertation.
2. It takes a long time to obtain clinical endpoints (i.e. Day 56 for DAS28 and 6-24 months for ACR20) which make adaptation difficult based on interim data. In published Phase 2 and Phase 3 studies in the literature (Geusens, 2008) for an investigative medicine, more than 90% of studies have endpoints that are more than 3 months duration. The nature of chronic diseases and the duration of clinical endpoints, make it difficult to use adaptive designs, since it takes longer to make decisions at interim analysis based on the long term endpoint.

This dissertation attempted to find solutions to the above challenges. The two-part “learn and confirm” design was implemented in the Phase 2a PoC RA study (Chapters 4). The flexible Bayesian NDLM model improved the model fitting in the PoC design under the assumption of a U-shaped curve (Chapter 6) and the meta-analysis based DRPM approaches predicted Day 56 data using the early time point, and could be applied to the future PoC study of follow-on compound GSK654321 and achieve significant time-savings.

## The Aims of the Chapter

The aims of this chapter are to:

1. Review the aims of the research;
2. Recommend a better design for the future Phase 2a PoC trial of GSK654321;
3. Discuss the strengths and limitations of the proposed adaptive design;
4. Discuss extensions to the work.

## Review of Research Aims

At the beginning of the dissertation, there were five research questions.

1. **Can we apply a novel design to Phase 2a PoC trials in the clinical development of RA drugs?**

The Bayesian dose response adaptive design was successfully applied to a Phase 2a PoC clinical trial in patients with active disease RA in this dissertation.

The Bayesian dose response adaptive design has two parts. Part A of the study was a learning phase with single dose escalation and undertaken to identify efficacious and well tolerated dosing regimens in RA patients.

The first of the study (Part A) was a double blind, randomized, placebo controlled, multicentre study to investigate the efficacy and safety of single dose in RA patients. Only the starting dose in cohort 1 was pre-defined and subsequent doses were selected using a Bayesian dose response model. Between cohort analyses were undertaken for safety review and dose escalation.

The second part (Part B) was a confirming phase with repeated dose in a group sequential design, with the dose selected from the results of Part A. The primary objective of the study was to assess the effect of GSK123456 on DAS28, in subjects with active disease, compared with subjects receiving a placebo. A Bayesian dose-finding approach was selected as being the most efficient method for identifying the ED90 in the learning part and group sequential design was selected to control the type I error rate in the confirming part.

Although it is a failed study, the original design of the PoC study for GSK123456 had many merits. Firstly, the dose finding of the design combined two distinct functions in the trial, the dose escalation and dose finding. The dose escalation carefully escalates to higher doses to watch for any safety concerns. The dose finding selects an efficacious and optimal dose for the Part B of the trial.

To the best of my knowledge the adaptive design described and reported is this thesis is one of the first examples of a study that applied a Bayesian dose response adaptive design to a Phase 2a clinical trial in RA patients. This experience in adaptive design helps researchers compare the traditional fixed sample-size study design with an adaptive design and embrace more adaptive design.

Lastly, the adaptive study design provided flexibility to make an informed decision, on whether or not to proceed with further development of the drug/compound, earlier in the drug development pathway. The flexibility reflected in the decision at each dose escalation. The independent safety review committee met at each dose escalation meeting between cohorts to examine the data and all the decisions were made on accumulated data. The iSRC team used all the available information in the trial to review the safety data and determine the most appropriate dose level for the next cohort.

Although there are many merits, this study design can potentially be improved in two areas.

1) The *Emax* model that was used to analyze the data assumes a monotonic dose-response relationship. When the monotonic dose-response assumption is not satisfied, it is difficult to detect the dose response relationship and estimate ED90. Therefore, a more flexible dose response model is needed that can deal with monotonic and non-monotonic dose response relationship.

2) It took more than two years to go through cohort randomization design in Part A of the Phase 2a PoC study. The main reason was due to the long duration to obtain the late response at Day 56 at each interim analysis.

1. **What are the limitations of the traditional fixed sample size design in Phase 2a clinical trials in RA?**

In traditional fixed sample size design in Phase 2a clinical trials in RA, key elements, such as the dose levels, patients’ allocation to each dose level, and sample size for each dose level, remain unchanged throughout the duration of the trial. Making changes to the study design and/or stopping the ongoing study based on accumulated data is not possible. The statistical analysis of the outcome data is only conducted at the end of the study. The dose escalation, dose finding, and the confirmation of a selected dose, cannot be handled in a single study.

It was estimated that a minimum sample size of 270 is needed for a traditional Phase 2a design comparing placebo with one of 8 proposed study doses (0.03, 0.1, 0.3, 1, 3, 10, 20 and 30 mg/kg), to achieve 90% statistical power to detect a treatment difference of 0.95 in DAS28 change from baseline with pairwise comparisons without multiplicity adjustment in Part A alone. Such sample size cannot be afforded by study team in the early development of RA drug.

Alternative fixed sample size cohort design in Phase 2a RA study allocates each of 8 proposed study doses (0.03, 0.1, 0.3, 1, 3, 10, 20 and 30 mg/kg) evenly in 8 cohorts with 6 patients per cohort. Two of 16 placebo patients were allocated to each cohort. The total sample size is 64. There are certain levels of risks that relate to this design, if the intervention is supra-efficacious or non-efficacious, a trial would not be able to know until all subjects have been recruited at the completion of study, which meant that time, resources and money are wasted; if the drug is not working, it is not ethical to give a non-efficacious treatment to patients with active disease; or if there are safety concerns in the conduct of clinical trials, it potentially poses a safety risk to patients.

1. **How does the proposed Bayesian dose finding design compare with the traditional fixed sample size cohort design, in terms of statistical power and operating characteristics, in Phase 2a clinical trial design in the treatment of RA disease?**

In this dissertation, a fixed sample size cohort design was compared to a Bayesian dose finding design in the context of Phase 2a PoC design.

In the Phase 2a PoC study of GSK123456, Bayesian flexible designs such as Bayesian dose finding designs combined both dose escalation and dose findings in a single design and were able to respond to data by taking larger jumps in dose where there is no evidence of patient benefit. It was shown that the dose finding design was able to allocate more patients to doses around the ED90 (Chapters 6) in the simulation if the true response was *Emax* like effect. If a drug has no efficacy, the Bayesian designs would escalate quickly to the highest dose (Appendix 11.5). Fixed sample size cohort design would allocate each of 8 proposed study doses evenly in 8 cohorts.

Additional simulations were used to compare the Bayesian dose response adaptive design (i.e. fully adaptive) and fixed sample size cohort design (Chapter 6). It was shown that, analysed by NDLM model, the average sample size is 56 and 64 for Bayesian adaptive design and fixed sample size design respectively, if the true dose response is an *Emax* like curve. Thus, a sample size saving of 8 would be saved if Bayesian adaptive design is used. The overall probability of success, the posterior probability of treatment difference different from zero greater than 95%, was 92% (including early success and final success) in Bayesian adaptive design. Bayesian adaptive design allocated more patients to the doses around ED90 doses than fixed sample size cohort design, so fewer patients were exposed to non-efficacious drug.

In summary, the fixed sample size cohort design is inefficient for Phase 2a dose escalating trials since many patients may be enrolled into ineffective study arm/dose. Bayesian dose finding adaptive design allocates more patients to the doses around ED90, so fewer patients were exposed to non-efficacious drug doses, and the design achieved a reasonable statistical power with fewer patients.

1. **What are the statistical properties of the Bayesian dose finding design, with the normal dynamic linear model (NDLM), as compared to the Bayesian *Emax* model?**

Simulations were used to compare the *Emax* model and NDLM model. These simulations showed that NDLM was able to maintain the probability of success even in the case of non-monotonic dose response, while the *Emax* model could not. For a fully adaptive design, the Bayesian NDLM design has a higher probability of success (80% vs. 6%) and less probability of futility (18% vs. 92%) compared to the Bayesian *Emax* model in the case of non-monotonic dose response (Chapter 6). When using the Bayesian NDLM model, adaptation of dose allocation improved the model performance and the probability of success. Similar results have been shown in fixed and half-adaptive scenarios. An adaptive design - especially an half-adaptive design - is a more efficient design than fixed design due to a greater chance of the dose being selected being the ED90 with less average sample size being use in the clinical trial. In most cases the Bayesian *Emax* model works effectively and efficiently, with low bias and good probability of success in case of monotonic dose response. However, if there is a belief that the dose response could be non-monotonic based on prior knowledge as in our case study - where a compound in the same class seemed to have non-monotonic dose responses - then the NDLM is the superior model to assess the dose response. Within the parameters of the simulation the NDLM model was shown to be flexible with the ability to handle a wide variety of dose-responses, including monotonic and non-monotonic relationships.

Statistical biases for each planned dose groups under the assumption of the true dose response curve were compared and there was less bias in model fitting using NDLM method compared with the *Emax* model in most cases. This further supports NDLM methods in the context of Phase 2a clinical development, when the dose response relationship is unknown.

1. **How can decision-making be improved by applying a predictive model using early time point outcome to predict late time-point outcome?**

A retrospective analysis and systematic literature review were conducted to investigate the relationship of the clinical efficacy endpoint DAS28 at early time point versus late time point using a meta-analysis. It was shown in both the literature review and retrospective analysis (Chapter 7) that there was a significant linear relationship in DAS28 change from baseline between late time point i.e. Day 56 post-randomization and early time point i.e. Day 14 post-randomization. In an attempt to establish the surrogacy of early response of Das28 at Day 14 as surrogate endpoint of DAS28 at Day 56, a moderately strong correlation was shown after adjusting for treatment effects. At a multiple-trial level, a strong R-value is indicative of a strong association at trial-level, however, no individual level association is available due to lack of individual data from the meta-analysis. Future research needs to be conducted to further validate the surrogate endpoint when individual level data are available from meta-analysis.

Therefore, if a later time point of clinical efficacy endpoint DAS28 can be predicted using an earlier time point, interim decisions can be based on the early endpoints to expedite the decision-making in an adaptive design setting of Phase 2a development. The trial duration can be significantly reduced if the interim decision and adaptation are based on the outcomes collected at the early time point. The DRPM approaches using predicted Day 56 response were compared with the LOCF approach (Chapter 8) and it was shown that all the treatment estimates of DRPM 1, 2, and 3 models were unbiased under both the null hypothesis and the alternative hypothesis. However, for the LOCF approach the treatment estimates were found to be biased and the mean treatment effect of LOCF under-estimated the true effect under both the null hypothesis and the alternative hypothesis. The DRPM2 and DRPM3 approach also had reasonable operating characteristics in term of percentage of futility stopping and final success in the context of the proposed Phase 2a PoC design for GSK654321.

DRPM approaches are proposed in this dissertation and offered a way to impute the late response based on earlier responses. With the proposed method, there would be a predicted average time-saving of upto 294 days for an 8-cohort design with 7 between cohort analyses, if Day 14 (Week 2) data are used. Significant revenue saving (Palmer, 2013) could potentially be recognized if the drug is successfully launched into the global RA market earlier.

## A Better Design for the PoC Study of GSK654321

The study compound GSK123456 failed to deliver the expected benefit, and a Phase 2a clinical design is under consideration for the follow-on compound GSK654321, which has the same binding site as GSK123456 and a dose response profile that is similar to GSK123456. There are at least two methods to improve the dose response adaptive design, one is through a different dose response model – Bayesian NDLM model, the other is to make decisions based on the prediction of long term endpoint (Day 56) from the short term endpoint (Day 14). The design proposed by this dissertation utilizes a more flexible dose response model to accommodate non-monotonic dose response and predicted DAS28 at Day 56 from Day 14 for fast decision-making in an adaptive design.

Two dose response models, Bayesian *Emax* model and Bayesian NDLM model, were compared in the dose response setting of a RA Phase 2a PoC clinical trial. Simulations have shown that the NDLM model has better power and is a more flexible dose response method in monotonic and non-monotonic dose responses. The proposed DRPM 2 models based on predicted Day 56 response estimate the mean treatment effect and have good operating characteristic in the Phase 2a PoC setting of RA design.

To the best of my knowledge, this is the first time such methods have been used and reported in Phase 2a trials in RA.

## Comparisons of Phase 2a PoC Design in RA

Similar two-part clinical designs such as seamless Phase 2/3 design (Stallard, 2011) and two-part PoC and dosing finding design (Smolen, 2014) were conducted in clinical research, however, none of them integrated both the dose escalation and dose finding in the same design with Bayesian dose response model to identify the doses in RA patients with active disease in Part A and applied the group sequential design with selected Part A dosed in Part B. To the best of my knowledge, the two-part Phase 2a PoC design is a novel design and is the first design that was applied to RA clinical research.

## Comparisons of Bayesian NDLM model

The application of Bayesian NDLM model in medical research have been discussed in the listerature (Berry, 2002; Krams, 2003). The Bayesian NDLM model have been successfully applied to a Phase 2 dose selection or dose finding design in acute stroke patients - ASTIN study (Grieve, 2005). Similar to the study design in my research, the ASTIN design allowed early stopping for futility if there is no drug effect or stopping for success if a dose is efficacious. Nine hundred sixty-six acute stroke patients were allocated to one of 15 doses adaptively close to ED95 based on the accumulated data using Bayesian NDLM model, which improve learning of dose response model (Krams, 2003). No RA studies have used Bayesian NDLM design in Phase 2a studies.

Temple (2012) compared the NDLM and *Emax* within the context of Phase 2b trial with larger sample size (total sample sizen=250-280, n=~30 per arm) and a narrower dose ranges (0-8 mg) compared to the trials described and reported in my dissertation. This work showed that both Bayesian DNLM model and *Emax* model detected a dose response well but Bayesian NDLM tend to have the highest power in the probability of detecting a clinical response than *Emax* model. The work in this PhD investigated the comparison of both NDLM and *Emax* with the context of a Phase 2a trial – with a far smaller sample sizes (total sample size = 64, n=6-8 per arm) and a wide dose range (0 to 30 mg/kg).

To the best of my knowledge, this is the first work to apply a Bayesian NDLM method and compare it to an *Emax* model in a Phase 2a PoC study with a wide dose range and small sample size. It is also the first to apply such methods in RA patients.

## Comparisons of Delayed Response Models

Delayed response models have been considered in clinical trials since the early nineties of the last century (Harwick, 2001; Harwick, 2006; Koopmeiners, 2014; Fu, 2010). Similar to my research, Weir et al. (2007) also assumed that early responses were a strong predictor of late responses and the LOCF method could be used here to impute the late responses. The idea was that both the observed and imputed final responses are used for the interim analysis. Along the line of imputation, the NDLM (West and Harrison, 1989) was used to associate early responses with final responses (Berry, 2002). In addition, Fu and Manner (2010) proposed an integrated two-component prediction (ITP) longitudinal model for this scenario. Their approach uses all the available early and late data at the interim instead of only imputing the missing final responses. Other multiple imputation methods have been proposed to impute late responses (Royston 2004).

Analogous research has been conducted to predict the clinical response of ACR20 and ACR50 using earlier time points (Nixon, 2008) in RA Phase 3 clinical trials. Nixon et al. (2008) used a predictive method to estimate the 6-month ACR20 and ACR50 using 1-month or 3-month ACR20 and ACR50 data.

All the above methods were developed to link early and late responses; however, none of the above was applied to DAS28 responses in RA Phase 2a clinical trials.

The proposed DRPM models especially DRPM 3 in this dissertation provide a way to predict Day 56 response based on Day 14 data and offer a solution to expedite the decision-making in the interim analysis. To the best of my knowledge, this is the first dissertation to apply DRPM model to DAS28 responses and to Phase 2a PoC studies in RA clinical area.

## Generalization of Delayed Response Models

The delayed response models and its application in cohort design were developed for GSK654321 in RA patients and significant time saving and money saving can be expected if short term endpoint is used in the interim decision makings. The results cannot be generalised beyond the context of the work in the dissertation. It is thus acknowledged that there will be less time and money saving for other designs such as parallel designs

However, it is important to highlight, that the process of using the published literature and retrospective in house data to understand the relationship of early and late response as well as the use of simulations to investigate the properties of the study design can be generalized to other studies or other disease areas. Applying these approaches may bring different solutions for a particular study even if the same methods can be used.

## Limitations of the Research

### Limitations of Proposed Phase 2a Design

The two-part Bayesian dose finding adaptive design was proposed to combine “learning” and “confirming” in RA patients. It is acknowledged that the adaptive design may not be appropriate for every study, and is especially challenging in the Part B if the trial recruitment is fast. Here, the study enrolment may come to an end before the interim analysis can be performed.

### Limitation of NDLM model and Predictive Model

It has been shown that early data was able to reduce the time to the decision at interim analysis. However, the predictive model described in this dissertation has several limitations.

1. Publication bias, the literature used in the meta-analysis came from the public domain and there is a possibility for some publication bias to have occurred. Treatments at an early phase of development may not go on to achieve drug regulatory approval and may not be published;
2. Drop-outs were not considered in the NDLM model and simulation of predictive model. In the PoC study of GSK123456, there was no drop-out, therefore, it is assumed that there was no drop-out in the simulation for GSK654321; however, the sample size for each cohort is small, drop-outs may affect NDLM model and the predictive model performance. That may also impact on model generalisability.

## Recommendations for the Phase 2a PoC Study of GSK654321

For the future Phase 2a trial of GSK654321, the study design will be similar to that of GSK123456. There will be two improvements:

1. The proposed delayed response model will be used to predict the response at late time point based on the early time point, given the linear relationship from literature. The simulations have shown that the model can be applied to the adaptive design of a RA clinical trial which has been shown from the simulation that a significant time-saving can be achieved;
2. A more flexible dose response model will be used to handle the likely non-monotonic dose response without sacrificing the statistical power and model performance. The results of this dissertation provide an example of novel design, combining NDLM model and short term endpoints so that the interim decisions can be made more flexible.

## Extension and Future Work:

### Extending the NDLM Model and Prediction Model to Other Disease Areas

In this dissertation, the Bayesian NDLM model out-performed the Bayesian *Emax* model, especially when the dose response relationship was non-monotonic. Although we applied the models to RA data the NDLM model could be extended to other disease areas to investigate the dose response relationship with the approaches for investigation described in this PhD used to assess their suitability. In addition, the DRPM model can be extended to other disease areas if a similar relationship of early response and late response can be confirmed.

# Contribution to clinical research

In RA clinical research, there is a need for better designs to make informed decisions i.e. to stop a failing drug earlier or move more quickly to the next phase of clinical development if a drug demonstrates efficacy. Compared to traditional fixed design, adaptive designs offer a better alternative for the Phase 2a PoC trial by allocating patients more efficiently and modifying the study prospectively.

In this dissertation, the two-part adaptive design was successfully applied in a Phase 2a PoC study. Part A was the learning phase to learn about the dose response, while Part B was the confirming phase where a group sequential design was applied. Given the high expectation of a non-monotonic dose response it was shown how a Bayesian NDLM model better describes the dose response. This was the first study to compare NDLM to *Emax* in the context of a Phase 2a study design.

Lastly, to establish the surrogacy of the early response of Day 14 for late response of Day 56 a single-trial and multiple trial level surrogacy investigation was undertaken. A meta-analysis of published literature was used to show how decisions could be made faster if Day 14 and not Day 56 data is used to make adaptive design decisions. The methodological described in the PhD to determine if using a short-term outcome can be used as a predictor or proxy or surrogate for a longer term outcome can be generalised and applied to any clinical development where both short-term and longer term outcomes are collected.

To make a better informed design, the PoC trial of the follow-on compound GSK654321 will have a two-part adaptive design, with Bayesian NDLM analysis for dose response with decisions made using Day 14 data. If the predictive model can be applied to the GSK654321, a significant time-saving can be expected in the clinical development. This saving in time could potentially equate to significant revenue saving for pharmaceutical companies.

**Publication and Manuscript related to the Dissertation**

**Publication**

Choy E.H., Bendit M., McAleer D., **Liu F.,** Feeney M., Brett S., Zamuner S., Campanile A., Toso J.. Safety, tolerability, pharmacokinetics and pharmacodynamics of an anti- oncostatin M monoclonal antibody in rheumatoid arthritis: results from Phase 2a randomized, placebo-controlled trials. *Arthritis Res Ther.* 15(5):R132. (2013)

Liu, F., Julious, S., Walters, S.. Design Considerations and Analysis Planning of a Phase 2a Proof of Concept Study in Rheumatoid Arthritis in the presence of possible non-monotonicity. *BMC Medical Research Methodology* Accepted (2017).

**Conference Talk**

Liu, F., Austin, D., Julious, S. (2012) A Delayed Response Model to Assess the Proof of Concept Clinical Design in Treatment to Rheumatoid Arthritis, Joint Statistical Meeting San Diego

Liu, F. (2016) Efficient Design using Early Timepoint, Invited Talk at Society of Clinical Trial Annual Meeting Montreal Canada

**Poster**

Liu, F., Austin, D., Julious, S. (2014) Bayesian Model in Adaptive Dose Range Study in Proof of Concept Study of Rheumatoid Arthritis. University of Sheffield ScHARR.

**Acknowledgement**

I would like to gratefully and sincerely thank my supervisors, Prof. Steven Julious and Prof. Stephen Walters for their guidance, understanding, and patience during my 6-year studies. I couldn’t have done this without their supervision and encouragement. I would also like express my thanks to the two examiners.

My thanks go to GlaxoSmithKline who sponsored my PhD study and my line managers at GlaxoSmithKline, Dr. Daren Austin, Dr. Tim Montague, Dr. Graeme Archer and Mr. Phil Overend. Specifically, Dr. Darren Austin serves as my dissertation advisor at GlaxoSmithKline and helped to design the PoC study of GSK123456; Dr. Tim Montague allowed me to use flexible work hours for PhD work.

I would also like to thank all of the members of GSK123456 PoC study team and all investigators, specifically Dr. John Toso, Mrs. Andrea Campanile and Dr. Mark Baker. Dr. Toso was the study physician and Mrs Campanile was the clinical operation lead; without their leadership and support, dose escalation and interim analysis couldn’t be done. Dr. Baker was the clinical pharmacologist and we went to the University of Sheffield together on my first day trip to Sheffield.

I want to take this opportunity to express gratitude to all faculty members at the Department of Medical Statistics, especially to Prof. Mike Campbell for his input and valuable discussions at the time of PhD conversion. My thanks also go to all the students at ScHARR.

Finally, and most importantly, I would like to thank my wife Yun for her support, patience and love. I would also like to thank our two young kids, Alexander and Emma for being so sweet and considerate, especially when Dad is away for study. I also thank my parents, parents-in-law, and my family for their encouragement.

# Appendix

## Programming codes

This appendix contains the programming codes that have been used in this dissertation, including the WinBUGS code for the Bayesian *Emax* model (Chapter 4) and the NDLM model (Chapter 6), and the SAS code for DRPM simulations. The WinBUGS software (Lunn 2000) is used to produce the Bayesian model and NDLM model. WinBUGS is free software available to download at <http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/>. SAS is commercial software (SAS, 2002).

### WinBUGS Codes for Bayesian *Emax* Model

The following WinBUGS codes are developed to perform the Bayesian analysis for the dose escalation using Bayesian *Emax* model. The priors on model parameters are weak priors. SAS was used to call WinBUGS during the Bayesian analysis.

The dose response model below is a re-parameterization of a 3-parameter *Emax* model with baseline DAS28 as covariate, in *β*3 being the EC90 which is the dose level that produces 50% of maximal change in DAS28.

WinBUGS Code

model

{

for( i in 1 : N ) {

z[i] ~ dnorm(mu[i],tau)

mu[i] <- beta0 + beta1 \* y[i] + beta2 \* 9 \* x[i] /(9 \* x[i] + exp(1.151293\*(beta3 -7)))

}

tau ~ dgamma(0.001,0.001)

sigma <- pow(tau, -0.5)

beta0 ~ dnorm(0.0,1.0E-6)

beta1 ~ dnorm(0.0,1.0E-6)

beta2 ~ dnorm(0.3,1.0E-2)

beta3 ~ dnorm(8.0,1.0E-6)

}

### WinBUGS Codes for Bayesian NDLM Model

WinBUGS codes are developed for the Bayesian NDLM mode in Chapter 6. The SAS software was used to call WinBUGS during the Bayesian analysis of the clinical trial.

WinBUGS Code

model{

for (j in 1:nsub){

y[j]~dnorm(mu1[j],v)

mu1[j] <- theta[dose[j]]+ beta\*base[j]

theta.pred[j]~dnorm(mu1[j],v)

# standardise covariates

Base[j] <- BASE[j] - mean(BASE[])

}

for (j in 2:J){

theta[j]~dnorm(mu[j-1],w)

mu[j-1]<-theta[j-1]+delta[j-1]

delta[j]~dnorm(delta[j-1],s)

effect[j] <- -theta[j]+theta[1]

# Probability of efficacy at each dose which treatment diff # > 0 0.8 0.95 1.2

efficacy1[j] <- step(effect[j]-0.95)

efficacy2[j] <- step(effect[j]-0.8)

efficacy3[j] <- step(effect[j]-1.2)

futility[j] <- 1-step(effect[j])

}

theta[1]~dnorm(mu\_theta,sigma\_theta)

delta[1]~dnorm(mu\_delta,sigma\_delta)

mu\_theta~dnorm(a,b)

mu\_delta~dnorm(c,d)

beta ~dnorm(0,1.0E-3)

sigma\_theta~dgamma(0.001,0.001)

sigma\_delta~dgamma(0.001,0.001)

v~dgamma(0.001,0.001)

w~dgamma(0.001,0.001)

s~dgamma(0.001,0.001)

a~dnorm(0,0.001)

c~dnorm(0,0.001)

b~dgamma(0.001,0.001)

d~dgamma(0.001,0.001)

}

### SAS Code for Permutation Test

The SAS codes are presented below for permutation test being conducted in Chapter 6

/\*\*Macro For DataSet Permutation \*\*/

**%macro** randomize(

indata=\_last\_,

outdata=Randoutput,

treat=y,

numreps=**1000**,

seed=**1234**);

proc sql noprint;

select count(\*) into :numrecs from &INDATA;

quit;

data temp1;

retain seed &SEED;

drop seed;

set &INDATA;

do replicate=**1** to &NUMREPS;

call ranuni(seed,rand\_dep);

output;

end;

run;

proc sort data=temp1;

by replicate rand\_dep;

run;

data &OUTDATA;

array deplist{&NUMRECS} \_temporary\_;

set &INDATA(in=in\_orig) temp1(drop=rand\_dep);

if in\_orig then do;

replicate=**0**;

deplist{\_n\_}=&TREAT;

end;

else &TREAT=deplist{**1**+mod(\_n\_,&NUMRECS)};

run;

**%mend** randomize;

%***randomize***(indata=original\_56, outdata=original\_56\_2, treat=dose, numreps=5**000**, seed=**1234**)

**data** original\_56\_2;

set original\_56\_2;

logdose=log(dose+**1**);

logdosesq=logdose\*logdose;

**run**;

\* This is observed data analysis;

/\*

proc reg data=original\_56\_2 (where =(replicate=0)) ;

by replicate;

\*class dose;

model das28\_c= logdose ;

ods output ParameterEstimates=results\_reg\_obs;

run;

\*/

**proc** **reg** data=original\_56 ;

\*by replicate;

\*class dose;

model das28\_c= logdose logdosesq;

\*ods output ParameterEstimates=results\_reg\_obs;

**run**;

**proc** **reg** data=original\_56\_2 ;

by replicate;

\*class dose;

model das28\_c= logdose logdosesq;

ods output ParameterEstimates=results\_reg;

**run**;

**data** results\_reg1;

set results\_reg;

where variable="logdosesq";

if tvalue >**1.84** then flag= **1**;

else flag=**0**;

**run**;

**proc** **means** data=results\_reg1;

var flag;

**run**;

### SAS Codes for DRPM Simulation

The SAS codes are presented below for comparison and simulations of DRPM 1, 2, 3, and LOCF models in predicting DAS28 Day 56 response based on Day 14 data (Chapter 8). Below is a sample code and full SAS code will be available for download.

\*\*\*Data Simulation of dose placebo;

**data** datapbo(keep=SimID cohort r1 r2 i y y2true group);

call streaminit(&seed);

do SimID = **1** to &NumSim; /\* 1. create many simulations \*/

do cohort = **1** to &nchort;

do i = **1** to &N1; /\* sample of size &N \*/

r1=rand("NORMAL");

r2=rand("NORMAL");

y = &diffpbo + &SD\*r1/sqrt(**2**);

y2true = &int0true+ &beta0true\*&diffpbo + +&rho\*&SDy\*r1+sqrt(&SDy\*\***2**-&SDy\*\***2**\*&rho\*\***2**)\*r2;

group = **1**;

output;

end;

end;

end;

**run**;

\*\*\*Data Simulation of dose 0.03 mg/kg;

**data** datatrt1(keep=SimID cohort r1 r2 i y y2true group);

call streaminit(&seed+**1**);

do SimID = **1** to &NumSim; /\* 1. create many simulations \*/

do cohort = **1** to &nchort;

do i = **1** to &N2; /\* sample of size &N \*/

r1=rand("NORMAL");

r2=rand("NORMAL");

y = &difftrt1 + &SD\*r1/sqrt(**2**);

y2true = &int1true+ &beta1true\*&difftrt1 + +&rho\*&SDy\*r1+sqrt(&SDy\*\***2**-&SDy\*\***2**\*&rho\*\***2**)\*r2;

group = **2**;

output;

end;

end;

end;

**run**;

\*\*Repeated for doe 0.3, 3, 10, 20, and 30 mg/kg;

\*\*\*The predictive modle is to predicate mean to mean;

\*\*\*Calculate the mean of early time point response;

**data** dataall;

set datapbo datatrt1 datatrt2 datatrt3 datatrt4 datatrt5 datatrt6;

**run**;

**proc** **sort** data= dataall; by simID cohort group;

\*First interim analysis at Cohort 3;

**data** dataIA1;

set dataall;

where cohort <=**3**;

if cohort =**3** then flag\_miss=**1**;

else flag\_miss=**0**;

**run**;

**proc** **sort** data= dataIA1; by simID group;

**Proc** **means** data= dataIA1 noprint;

by simID group ;

var y;

output out=datamean mean=mean stddev=sd;

**run**;

**data** dataallall;

merge dataIA1 datamean(drop=\_type\_ \_freq\_);

by simID group;

**run**;

**Proc** **means** data= dataallall(where=(flag\_miss=**1**)) noprint;

by simID group;

var y;

output out=datamean\_miss mean=mean\_miss stddev=sd\_miss;

**run**;

**data** dataallall1;

merge dataallall datamean\_miss(drop=\_type\_ \_freq\_);

by simID group;

**run**;

\*\*\*\*Assume sigma square is known and is 0.422 from meta-analysis;

**data** dataallall2;

set dataallall1;

call streaminit(&seed+**7**);

if group =**1** then do;

\*\*\* This is the TRUE late response with a common slope beta1\*\*\*;

\*\*\* This is the predicted late response DRPM1\*\*\*;

call streaminit(&seed+**7**);

y2pred1=&int0+&beta0\*mean\_miss ;

\*\*\* This is the predicted late response DRPM2\*\*\*;

y2pred2=&int0+&beta0\*mean\_miss + &beta0\*sd\*rand("NORMAL") ;

\*\*\* This is the predicted late response DRPM3\*\*\*;

y2pred3=&int0+&beta0\*mean\_miss + sqrt((&beta0\*sd)\*\***2**+**0.422**)\*rand("NORMAL") ;

end;

else do;

\*\*\* This is the TRUE late response with a common slope beta1\*\*\*;

y2pred1=&int1+&beta1\*mean\_miss;

call streaminit(&seed+**8**);

\*\*\* This is the predicted late response DRPM2\*\*\*;

y2pred2=&int1+&beta1\*mean\_miss + &beta1\*sd\*rand("NORMAL") ;

\*\*\* This is the predicted late response DRPM3\*\*\*;

y2pred3=&int1+&beta1\*mean\_miss+ sqrt((&beta1\*sd)\*\***2**+**0.422**)\*rand("NORMAL");

end;

if flag\_miss= **1** then do;

y2drpm1 = y2pred1;

y2drpm2 = y2pred2;

y2drpm3 = y2pred3;

y2late=**.**;

y2locf=y;

end;

else do;

y2drpm1 = y2true;

y2drpm2 = y2true;

y2drpm3 = y2true;

y2late=y2true;

y2locf = y2true;

end;

\*\*\*\* y2 is the data combining the observed data and predicated data based on the model;

\*\*\*\* y2late is the data with only the observed data;

\*\*\*\* y2locf is the data combining observed data and LOCF;

**run**;

### R Codes for Structural Correlation at Meta-Analysis Simulation

The R codes are presented below for the simulations of structural correlation of DAS28 score between Day 56 and Day 14 data in the meta- analysis (Chapter 7).

library(mvtnorm)

#Create covariance matrix

for(c\_b\_14 in c(0,0.1,0.3,0.5,0.7)) {

c\_b\_14=0.7

c\_b\_56 = c\_b\_14

mean<-c(6.5,5.5,5.0)

for(c\_14\_56 in c(0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9)) {

corrm <- matrix(c(1,c\_b\_14,c\_b\_56,c\_b\_14,1,c\_14\_56,c\_b\_56,c\_14\_56,1),3,3)

std=c(0.9,1.2,1.3)

#Simulating Multivariate Normal distribution.

diag\_std=std %\*% t(std)

covm=diag\_std\*corrm

base<-f;

for(ii in 1:10000) {

d=mvrnorm(n = 64, mean, covm)

a<-d[,1]

b<-d[,2]

c<-d[,3]

# Change from baseline

change\_b<-b-a

change\_c<-c-a

mean\_b<-mean(change\_b)

mean\_c<-mean(change\_c)

d<-cbind(mean\_b,mean\_c,mean(b),mean(c))

e<-rbind(base,d)

base<-e

}

g<-cor(e)

#Combining Results

h<-cbind(c\_b\_14,c\_b\_56,c\_14\_56,g[2,1])

j<-rbind(base1,h)

base1=j

}

OutFile <- paste(OutDir, "correlation\_pre\_post\_new",".csv", sep="")

write.table(j, OutFile, append=TRUE, sep="," , col.names=NA, row.names=TRUE)

}

## List of literature in the Meta-analysis

The literature in the meta-analysis and the same size are displayed in Table 11.1. The first author of literature, years of publication, sample size and study description are tabulated in the meta-analysis.

Table 11.1 List of literature used in the meta-analysis.

|  |  |  |
| --- | --- | --- |
| **Literature** | **Sample Size (Placebo+MTX/ Test Drug(s))** | **Study Description** |
| Fleischmann 2015 | 41/41/40/41 | Jak-3 Inhibitor |
| Dhir 2014 | 47/53 | Two starting doses of MTX |
| Schiff 2014 | 10/27 | Certolizumab Phase IV, study |
| Smolen 2014 | N=219 in 7 groups | Sirukumab, two-part design |
| Genovese 2014 | N=219 in 9 groups | Two parts with 9 arms. Phase 2b olokizumab ipatients with rheumatoid arthritis with an inadequate response |
| Smolen 2013 | 162/163 | Adalimumab/Tocilizumab |
| Burmester 2013 | 75/39/41/39/39 | Mavrilimumab |
| Genovese 2013 | 50/54/49/43/41 | Secukinumab a Phase II, dose-finding, |
| Mease 2012 | 33/32/33/29 | Phase II, double-blind, BMS945429 |
| Yazici 2012 | 207/412 | Tocilizumab, Rose study |
| Nishimoto 2010 | 37/53 | Subgroup analysis of the SATORI study |
| Stefano 2010 | 60/60 | Combination therapies |
| Yamanaka 2011 | 229 | Tocilizumab |
| Cohen 2009 | 53/52/51/48 | Pamapimod, a p38 MAP kinase inhibitor |
| Vander Cruyssen 2008 | 511 | Infliximab: validation of the DAS28 score |
| Genovese 2008 | 370/751 | TOWARD study; IL-6; Tocilizumab |
| Fiehn 2007 | 21/19 | Leflunomide or methotrexate |
| Vanags 2006 | 8/8/7 | Chaperonin |
| Rintelen 2006 | 6/9 | Leflunomide/Chloroquin |
| Quinn 2005 | 10/10 | Nfliximab synovitis and damage |

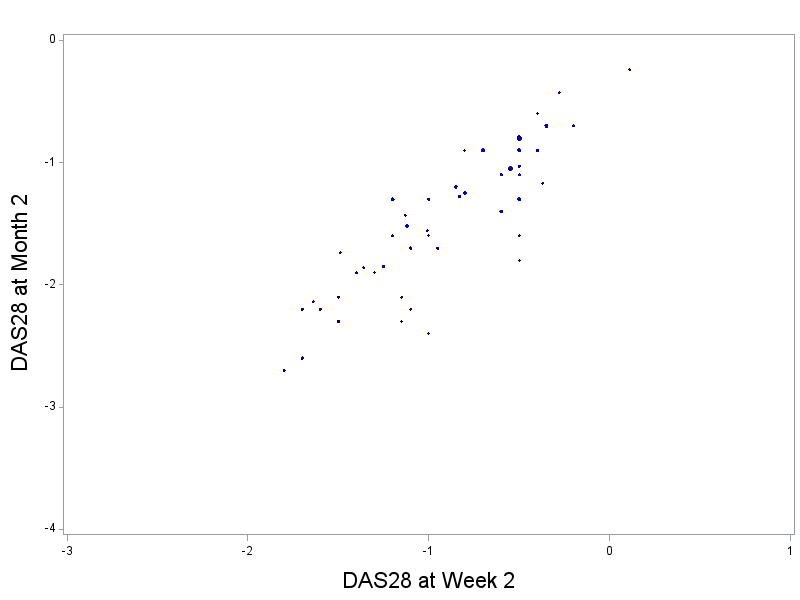
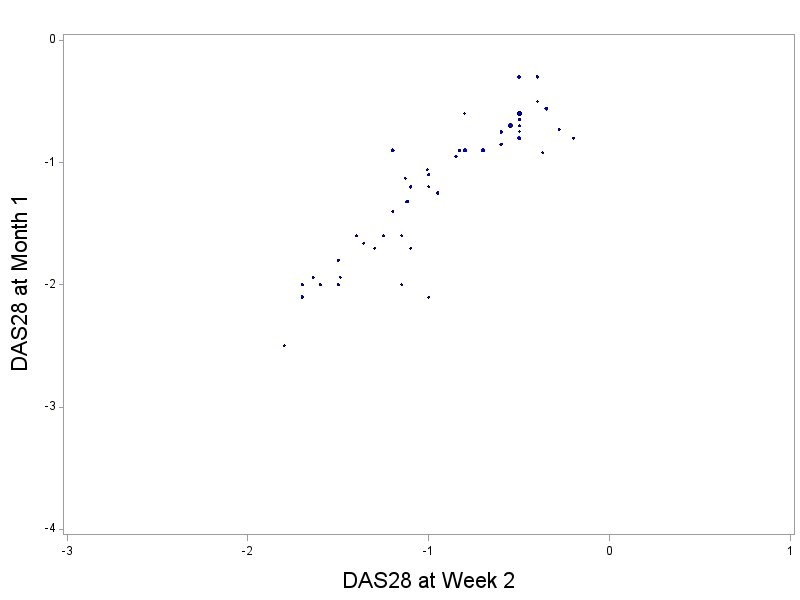
## Weighted Linear Regression ANCOVA Model Fitting and Diagnostics

The meta-analysis is analysed using weighted linear regression and the model fittings of the ANCOVA are presented in this appendix.

### Weighted Linear Regression

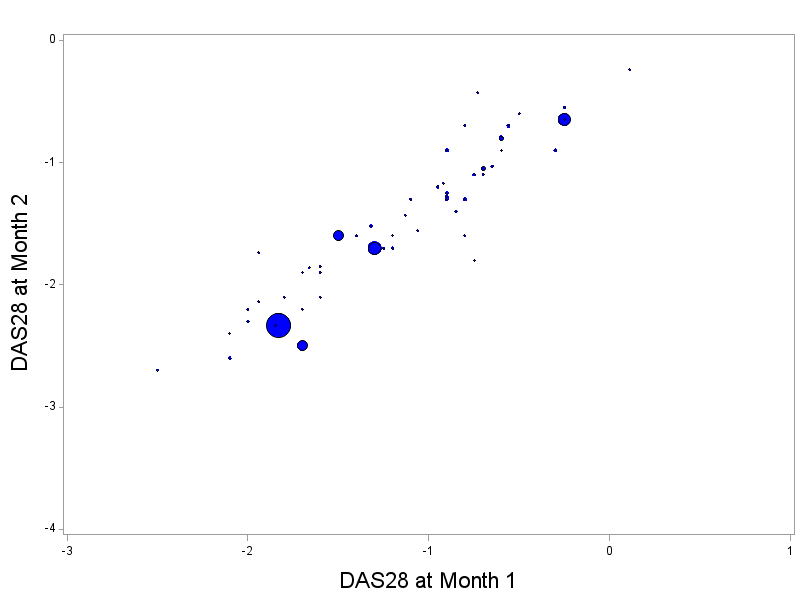
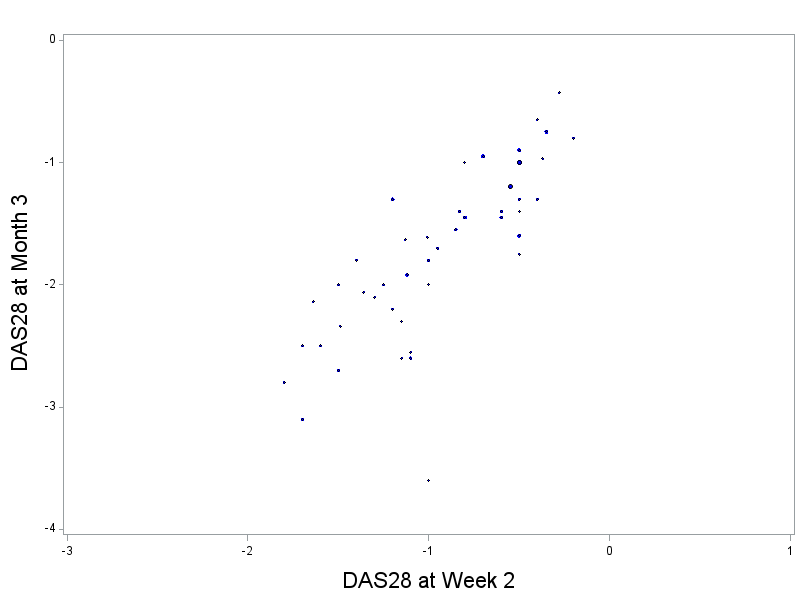
Analysis was undertaken using linear correlation of DAS28 change from baseline at both early and late time points using weighted linear regression with sample size as weight. Scatter plots are generated with the size of points representing the size of trial with a factor of square root (sqrt) of (N)/10, where N is the sample size of each treatment. The larger the sample size the bigger circles the dot in the graph. The analysis is performed using weighted linear regression analysis, using one tenth of the square root of the sample size in each trial as weight. The relationship of DAS28 at early and late time points are explored using scatter plots shown in Figure 7.8. Early time points are either Week 2 or Month 1 and late time points are Month 2, Month 3 or Month 6. Each point represented the weighted mean of DAS28 from each study treatment in the meta-analysis. In the scatter plots, all treatments are pooled.

It is evident that there is a positive linear relationship in mean DAS28 change from baseline between the Week 2 (Day 14) and Month 2 (Day 56) (graph B in Figure 7.8).



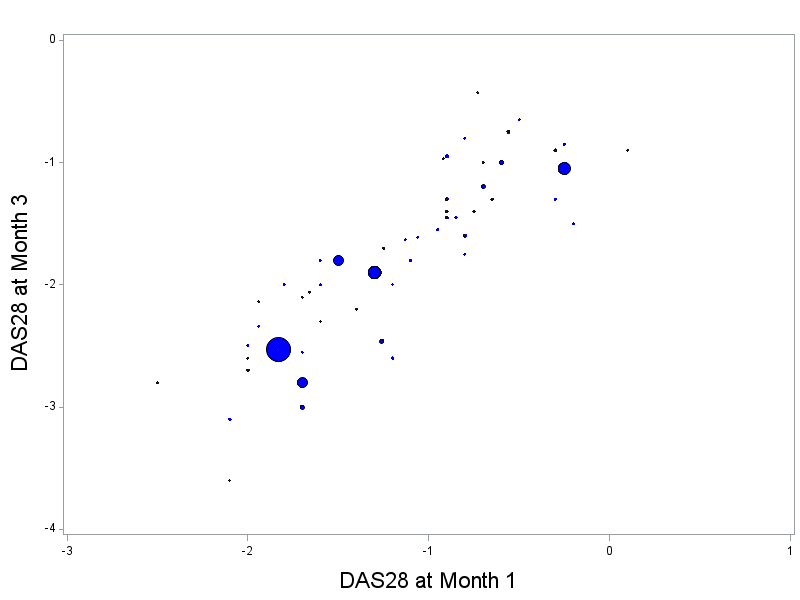
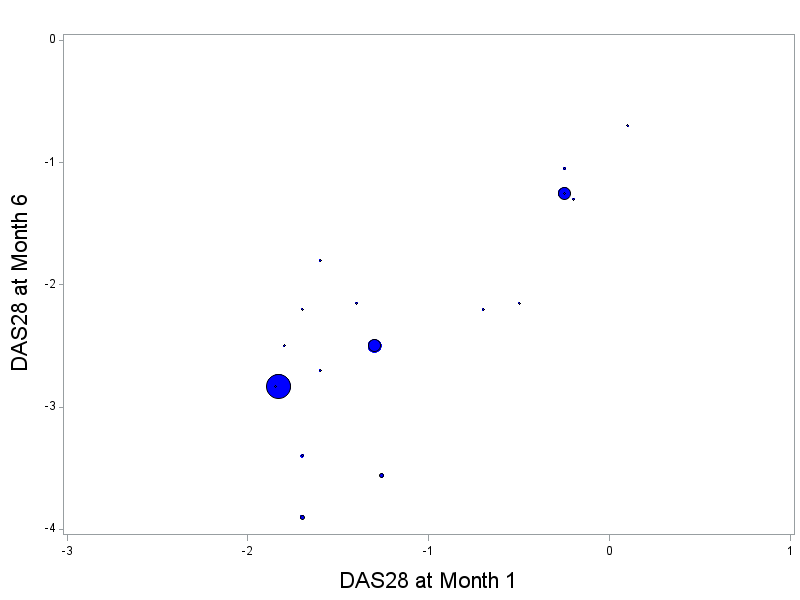
B

A



D

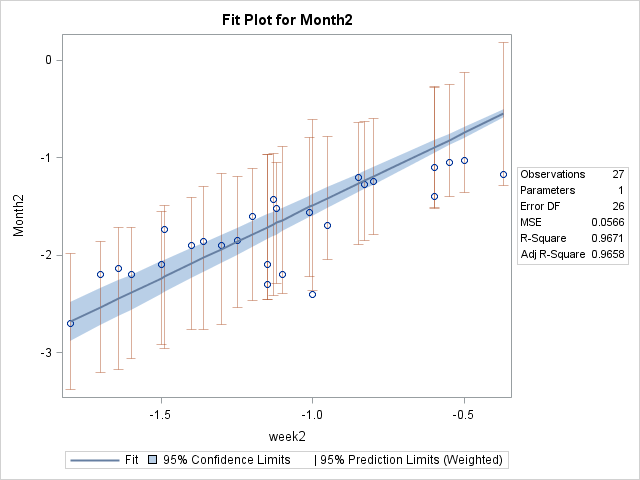
C

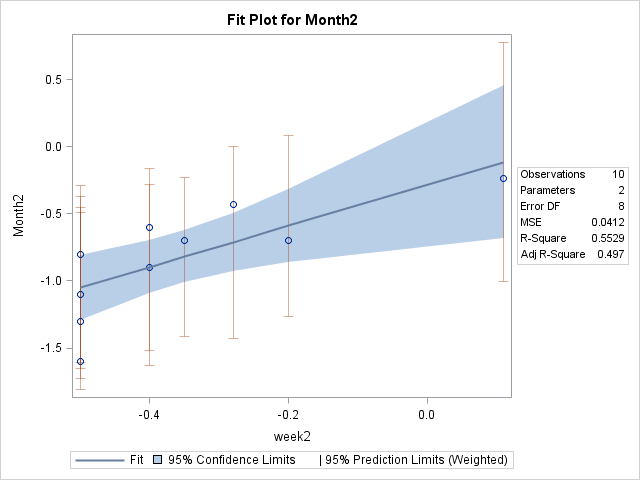
F

E

Figure 11.3 Scatter plots of DAS28 change from baseline at early time point and late time point with the size of dots representing the size of trial. Each point is the weighted mean of DAS28 from each study treatment of clinical trial in the meta-analysis. Not all the clinical trials in the meta-analysis have data for all time points and all treatments are pooled in this analysis. A: Month 1 vs. Week 2; B: Month 2 vs. Week 2; C: Month 3 vs. Week 2; D: Month 2 vs. Month 1; D: Month 3 vs. Month 1; D: Month 6 vs. Month 1.



A



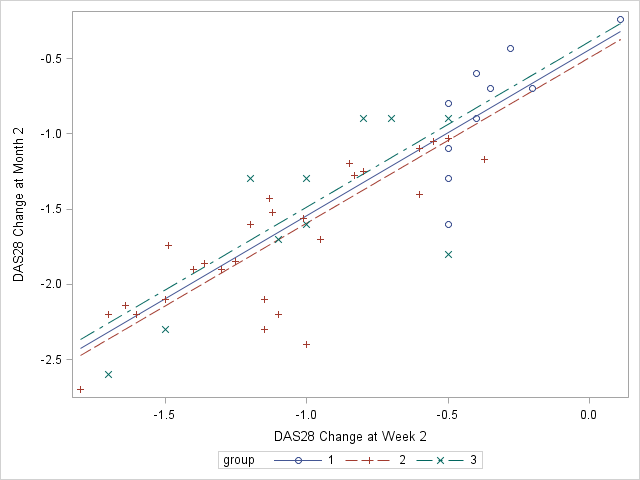
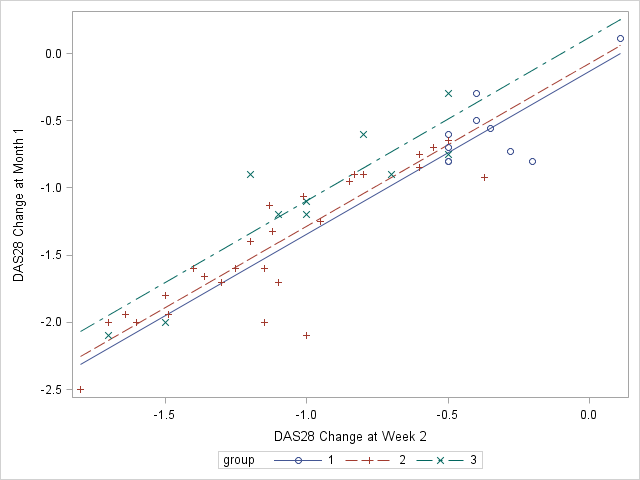
B

Figure 11.4 Scatter plot and fitted line of weighted linear regression of DAS28 change from baseline at Month 2 (Day 56) as a function of DAS28 at Week 2 (Day 14). Each open circle is the mean estimate of each individual trial. The solid line and shape areas are the linear fitted line and 95% confidence band. A: Biological treatment; B: Placebo/MTX.

Scatter plots and fitted line of weighted linear regression of DAS28 change from baseline at Month 2 (Day 56) as a function of DAS28 at Week 2 (Day 14) are shown in Figure 7.9. The confidence limit (shaded area) of the model fitting for placebo (n=10) is wider which could be a result of higher variability in placebo population, smaller sample size, or both. It is further evident that DAS28 at Day 56 is a linear function of Day 14 in the RA patients receiving biological or Placebo/MTX treatment.

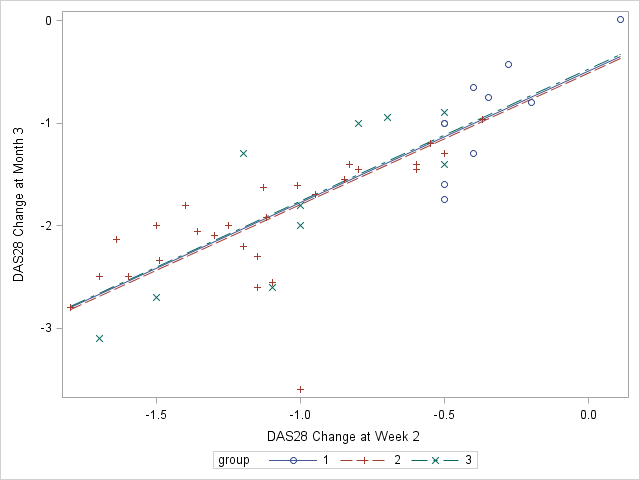
### Model Fitting

The model fittings of the ANCOVA model are displayed in Figure 11.3 and Figure 11.4 as supplemental material to Chapter 8. Selected influential values are removed prior to the analysis. It is shown that there are still a few influential values that may impact on the model fitting. The diagnostics plots of model fittings are presented in next Section 11.3.2.



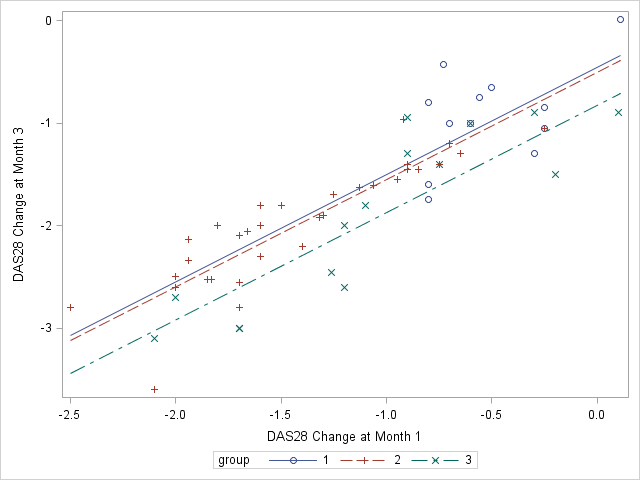
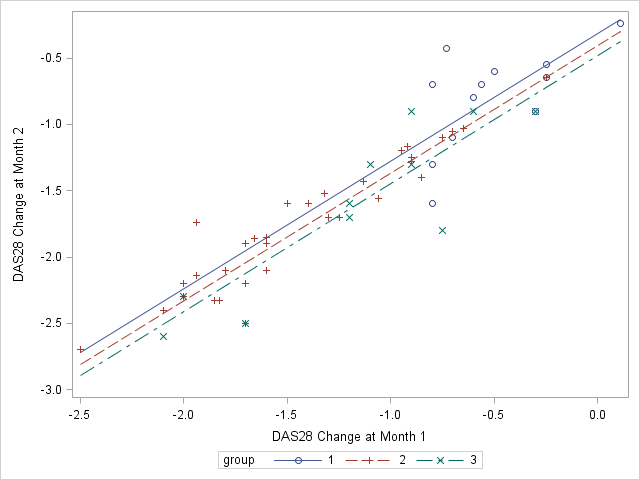
B

A



C

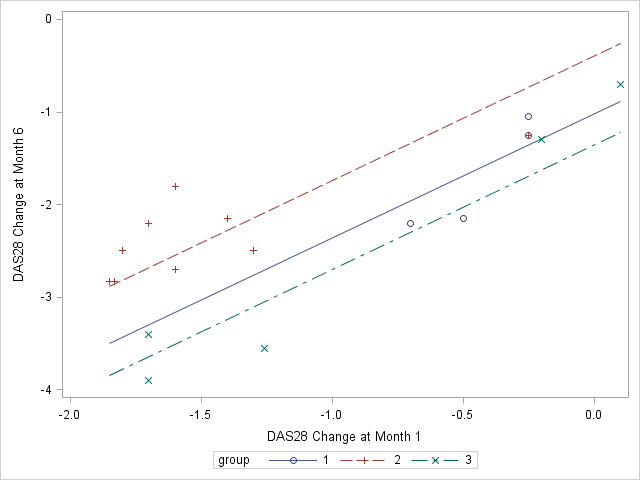
Figure 11.5. Fitting regression line along with the scatter plot of the observed DAS28 change from baseline data Month 1 (A: model 1), Month 2 (B: model 2) and Month 3 (C: model 3) vs. Week 2. Circle and solid lines denote treatment with placebo/MTX. Plus and Dashed lines are the treatment with a biological agent and x signs and dashed lines are the other RA treatments.



B

A

A



C

Figure 11.6. Fitting regression line along with the scatter plot of the observed DAS28 change from baseline data Month 2 (A: model 4), Month 3 (B: model 5) and Month 6 (C: model 6) vs. Month 1. Circle and solid lines denote treatment with placebo/MTX. Plus and Dashed lines are the treatment with a biologic agent and x signs and dashed lines are the other RA treatments.

### Model Diagnostics

Model diagnostic procedures of the ANCOVA model are performed to explore whether the assumptions of the regression model are valid. Figure 11.5 displays the observed and predicted graphs for models 1-6. In observed vs. predicted plots, diagonal lines with zero intercept are plotted and any values near or on the line indicate good model fitting. It is shown that most of the predicated values in model 1-6 are around the fitted line with few influential values which may affect the model fitting, i.e. model 3. The model fitting is performed again after removing the influential value and reported in Table 7.6.

Normal probability plots of the residuals (Figure 11.6) are used to check the normality assumption which each residual is plotted against its expected value under Normality. Both model diagnostics show the ANCOVA model fittings are reasonable.

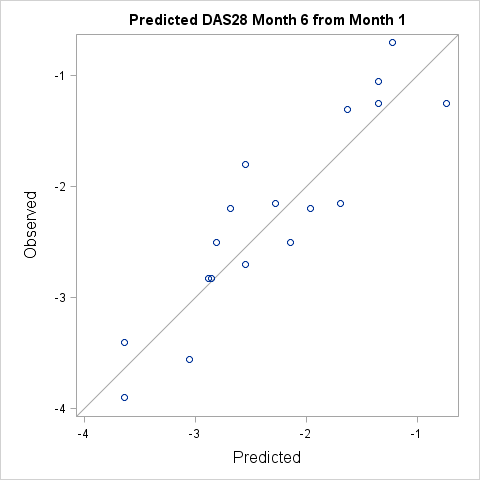
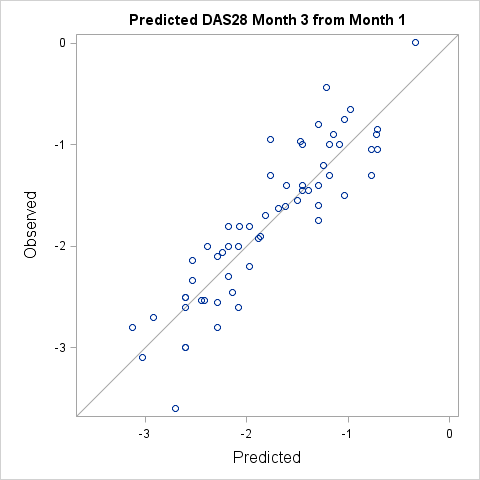
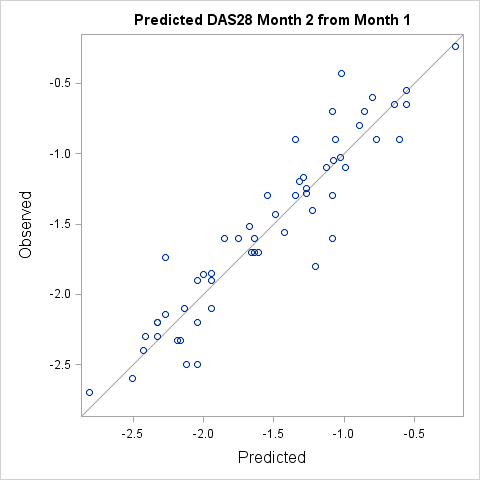
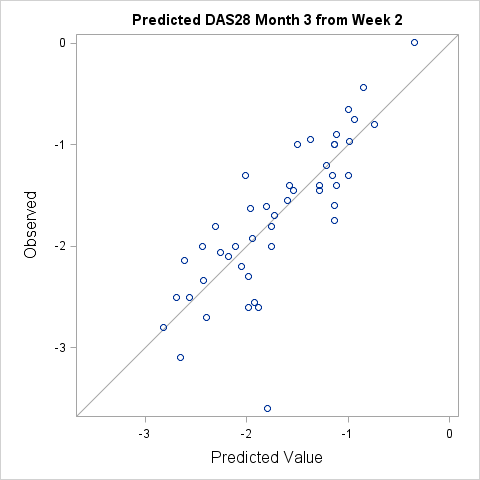
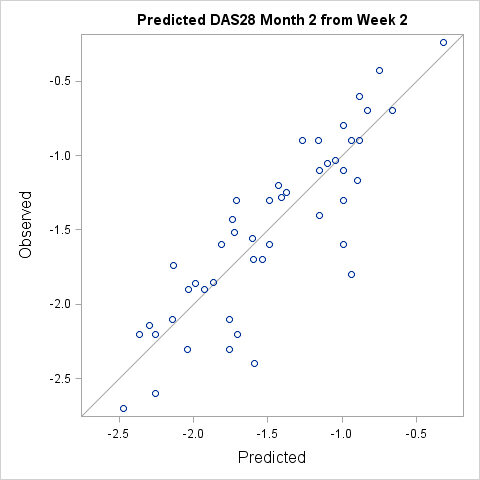
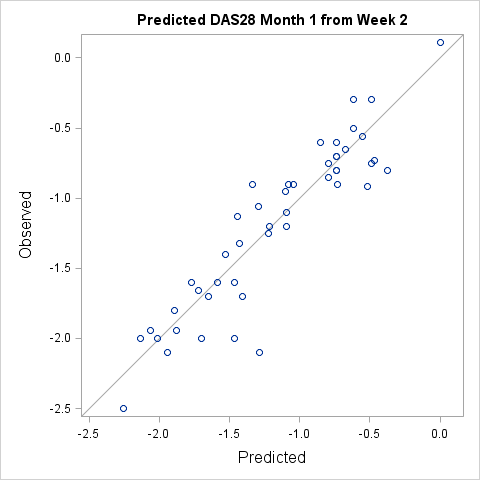


Figure 11.5. Observed DAS28 change from baseline plotted against predicted of the fitted models (model 1-6). A line with zero intercept is plotted in each graph.

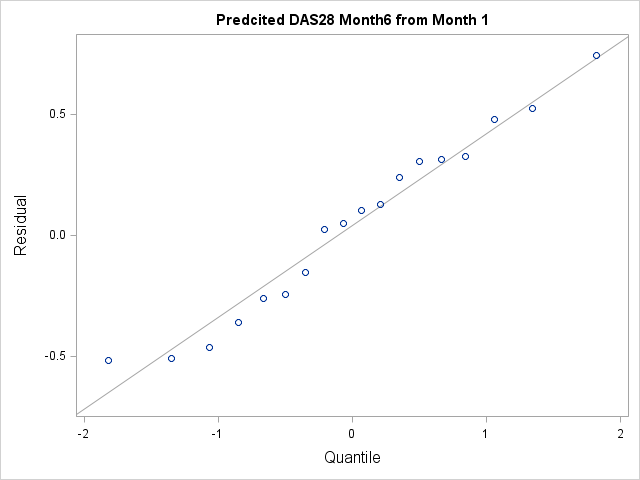
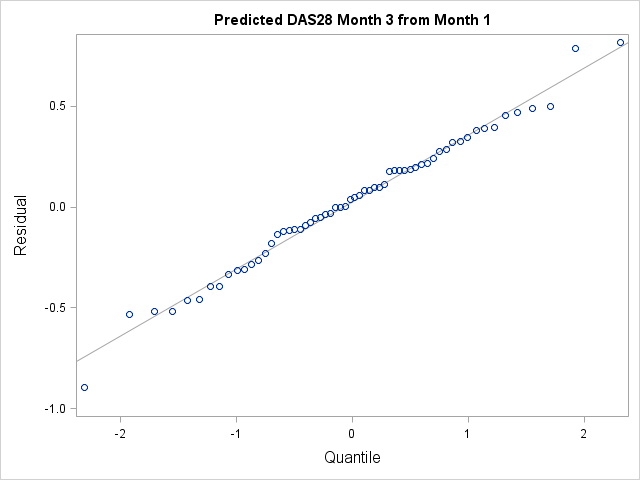
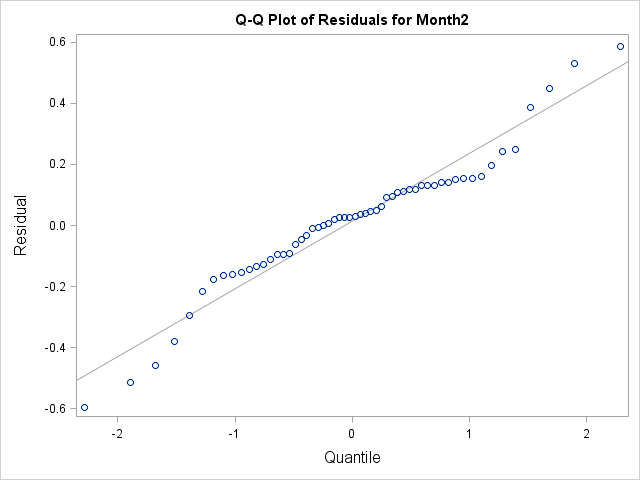
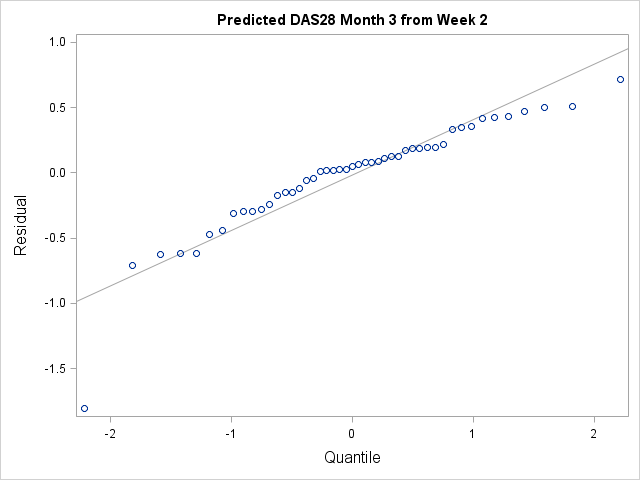
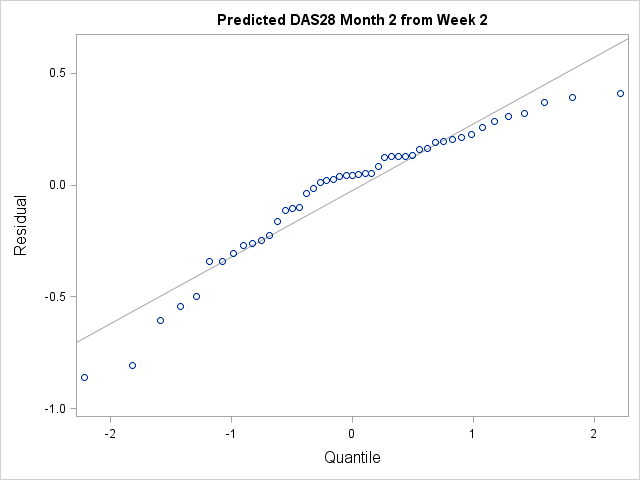
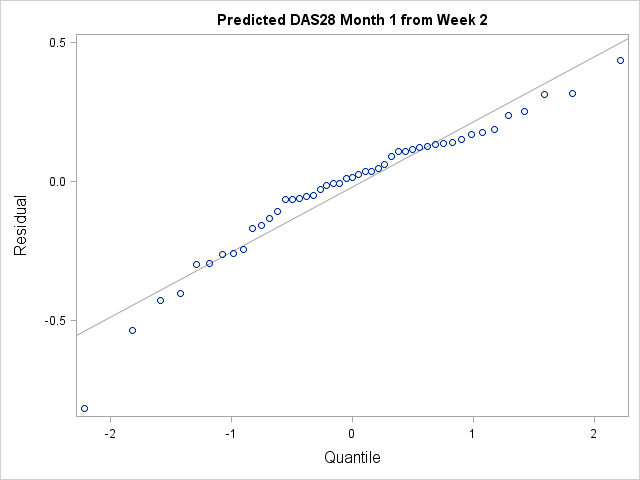


Figure 11.6. Model diagnostics: Q-Q plot of models 1-6 to predict late time point (Month 2, Month 3 or Month 6) using early data (Week 2 or Month 1).

## Supplemental Results for comparison of DRPM 1, 2, 3 and LOCF method in Chapter 9

The supplemental tables for sensitivity analysis of all DROM models and LOCF model are presented in Section 11.4.1 and additional evaluation of proportional data to be imputed is presented in Section 11.4.2.

### Supplemental Tables for Sensitivity Analysis of DRPM 1, 2, 3 and LOCF method

The supplemental tables provide complete simulated results to support the sensitivity analysis in Chapter 8 to evaluate the DRPM1, 2, 3 and LOCF model performance in comparison to true model 1) when the variance is lower than expected (or higher than expected (; 2) when correlation coefficient is lower than expected (ρ=0.1, 0.3, and 0.5) or higher ((ρ=0.9); 3) in case the relationship of late and early response are lower than expected (β=0.5, 0.75 and 1.0) or higher than expected (β=1.2). Other parameters were fixed as constant during the simulation. The complete results are presented in Table 11.2, 11.3 and 11.4 respectively.

In addition, additional sensitivity analysis are conducted when the relationship of late and early response are lower than expected (β=0.5, 0.75 and 1.0) or higher than expected (β=1.2) while fixing correlation coefficient at 0.1. The results are shown in Table 11.5 and Figure 11.7 for the *Emax* dose response under alternative hypothesis and Table 11.6 for the flat dose response under null hypothesis (placebo effect).

Table 11.2 Estimated treatment effect (treatment-placebo) and the interim futility probability of predicted model (the prediction of DAS28 at Day 56 based on Day 14), LOCF method and true data from Day 56 (Month 2) assuming data follow an *Emax* curve under alternative hypothesis. Data are simulated assuming parameters are ρ=0.7 =0.9,α1=-0.055, α2=-0.105, β1= β2=1.1 for placebo and treatment respectively.

1. **the standard deviation of DAS28 at Day 56 higher than expected (σ22=1.2)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.933 (1.026) | -0.933 (1.098) | -0.935 (1.158) | -0.890 (1.034) | -0.936 (1.182) |
| % Futility based on Cond. Power (<20%) | 24.1% | 28.5% | 30.8% | 27.1% | 31.9% |
| Bias | 0.003 | 0.003 | 0.001 | 0.046 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | 1.039 (1.108) | -1.047 (1.148) | -1.046 (1.174) | -1.028 (1.111) | -1.079 (1.179) |
| % Futility based on Cond. Power (<20%) | 14.2% | 15.3% | 16.3% | 14.1% | 13.5% |
| Bias | 0.040 | 0.032 | 0.033 | 0.051 | - |
| **Final Analysis** | | | | | |
| **Final Success (P-value <0.05)** | 57.4% | 52.5% | 49.6% | 55.0% | 52.4% |
| **Final Failure (P-value >=0.05)** | 4.2% | 3.7% | 3.3% | 3.8% | 2.0% |
| **Total Futility at Interim Analysis** | 38.3% | 43.8% | 47.1% | 41.2% | 45.4% |
| **Average Sample Size** | 52 | 50 | 49 | 51 | 49 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

1. **the standard deviation of DAS28 at Day 56 lower than expected (σ22=0.6)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.932 (0.537) | -0.932 (0.596) | -0.934 (0.644) | -0.889 (0.544) | -0.934 (0.591) |
| % Futility based on Cond. Power (<20%) | 5.5% | 8.9% | 11.8% | 6.5% | 7.8% |
| Bias | 0.002 | 0.002 | 0 | 0.045 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | 0.957 (0.569) | -0.965 (0.599) | -0.967 (0.625) | -0.939 (0.572) | -0.968 (0.593) |
| % Futility based on Cond. Power (<20%) | 1.5% | 2.3% | 3.0% | 1.7% | 1.4% |
| Bias | 0.011 | 0.003 | 0.001 | 0.029 | - |
| **Final Analysis** | | | | | |
| **Final Success (unadjusted P-value <0.05)** | 90.3% | 86.4% | 82.9% | 89.2% | 88.7% |
| **Final Failure (unadjusted P-value >0.05)** | 2.8% | 2.4% | 2.3% | 2.6% | 2.0% |
| **Total Futility at Interim Analysis** | 7.0% | 11.2% | 14.8% | 8.2% | 9.2% |
| **Average Sample Size** | 62 | 60 | 59 | 61 | 61 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

Table 11.3 Estimated treatment effect (treatment-placebo) and the interim futility probability of predicted model (the prediction of DAS28 at Day 56 based on Day 14), LOCF method and true data from Day 56 (Month 2) assuming data follow an *Emax* curve under alternative hypothesis. The data are simulated assuming parameters are =0.9, =0.9,α1=-0.055, α2=-0.105, β1= β2=1.1 for placebo and treatment respectively.

1. **Correlation coefficient of Day 14 and Day 56 response is lower than expected (ρ=0.1)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.935 (0.807) | -0.934 (0.898) | -0.936 (0.969) | -0.891 (0.817) | -0.937 (0.889) |
| % Futility based on Cond. Power (<20%) | 11.7% | 17.0% | 20.3% | 13.8% | 15.8% |
| Bias | 0.002 | 0.003 | 0.001 | 0.046 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -0.970 (0.853) | -0.973 (0.899) | -0.976 (0.937) | -0.953 (0.856) | -0.994 (0.888) |
| % Futility based on Cond. Power (<20%) | 7.0% | 8.9% | 10.6% | 7.4% | 5.7% |
| Bias | 0.024 | 0.021 | 0.018 | 0.041 | - |
| **Final Analysis** | | | | | |
| **Final Success (P-value <0.05)** | 79.5% | 72.5% | 67.8% | 77.3% | 77.3% |
| **Final Failure (P-value >=0.05)** | 1.8% | 1.6% | 1.3% | 1.6% | 0.8% |
| **Total Futility at Interim Analysis** | 18.7% | 25.7% | 30.9% | 21.2% | 21.5% |
| **Average Sample Size** | 58 | 56 | 54 | 57 | 57 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

1. **Correlation coefficient of Day 14 and Day 56 response is lower than expected (ρ=0.3)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.934 (0.807) | -0.934 (0.897) | -0.936 (0.969) | -0.891 (0.816) | -0.937 (0.888) |
| % Futility based on Cond. Power (<20%) | 11.8% | 16.9% | 20.6% | 13.6% | 15.7% |
| Bias | 0.002 | 0.003 | 0.001 | 0.048 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -0.972 (0.852) | -0.979 (0.898) | -0.981 (0.938) | -0.955 (0.856) | -0.993 (0.888) |
| % Futility based on Cond. Power (<20%) | 6.5% | 8.5% | 9.8% | 7.1% | 5.9% |
| Bias | 0.021 | 0.014 | 0.012 | 0.038 | - |
| **Final Analysis** | | | | | |
| **Final Success (P-value <0.05)** | 80.0% | 73.1% | 68.4% | 77.8% | 77.5% |
| **Final Failure (P-value >=0.05)** | 1.7% | 1.6% | 1.2% | 1.5% | 0.9% |
| **Total Futility at Interim Analysis** | 18.3% | 25.4% | 30.4% | 23.7% | 21.6% |
| **Average Sample Size** | 58 | 56 | 54 | 56 | 57 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

1. **Correlation coefficient of Day 14 and Day 56 response is lower than expected (ρ=0.5)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.933 (0.806) | -0.933 (0.896) | -0.935 (0.968) | -0.890 (0.816) | -0.936 (0.887) |
| % Futility based on Cond. Power (<20%) | 11.5% | 16.8% | 20.7% | 13.6% | 15.5% |
| Bias | 0.003 | 0.003 | 0.001 | 0.048 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -0.974 (0.852) | -0.982 (0.898) | -0.987 (0.937) | -0.958 (0.855) | -0.993 (0.887) |
| % Futility based on Cond. Power (<20%) | 6.2% | 7.9% | 9.4% | 6.5% | 5.8% |
| Bias | 0.019 | 0.011 | 0.006 | 0.035 | - |
| **Final Analysis** | | | | | |
| **Final Success (P-value <0.05)** | 80.4% | 73.8% | 68.8% | 78.4% | 77.6% |
| **Final Failure (P-value >0.05)** | 1.9% | 1.4% | 1.2% | 1.6% | 1.1% |
| **Total Futility at Interim Analysis** | 17.7% | 24.7% | 30.1% | 20.1% | 21.3% |
| **Average Sample Size** | 58 | 56 | 54 | 58 | 57 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

1. **Correlation coefficient of Day 14 and Day 56 response is higher than expected (ρ=0.9)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.931 (0.806) | -0.930 (0.893) | -0.933 (0.966) | -0.887 (0.816) | -0.932 (0.885) |
| % Futility based on Cond. Power (<20%) | 11.3% | 16.9% | 20.7% | 13.7% | 15.6% |
| Bias | 0.001 | 0.002 | -0.001 | 0.045 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -0.976 (0.853) | -0.987 (0.896) | -0.992 (0.936) | -0.962 (0.856) | -0.991 (0.888) |
| % Futility based on Cond. Power (<20%) | 5.2% | 7.1% | 8.6% | 5.4% | 5.7% |
| Bias | 0.015 | 0.004 | -0.001 | 0.029 | - |
| **Final Analysis** | | | | | |
| **Final Success (P-value <0.05)** | 81.8% | 74.6% | 69.6% | 79.5% | 77.8% |
| **Final Failure (P-value >=0.05)** | 1.7% | 1.3% | 1.1% | 1.5% | 1.0% |
| **Total Futility at Interim Analysis** | 16.5% | 24.0% | 29.3% | 19.1% | 21.3% |
| **Average Sample Size** | 59 | 56 | 54 | 58 | 57 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

Table 11.4 Estimated treatment effect (treatment-placebo) and the interim futility probability of predicted model (the prediction of DAS28 at Day 56 based on Day 14), LOCF method and true data from Day 56 (Month 2) assuming data follow an *Emax* curve under alternative hypothesis. The data are simulated assuming parameters are ρ=0.7 ==0.9,α1=-0.055, α2=-0.105 for placebo and treatment respectively.

**A: DAS28 at Day 56 is lower than expected (**β1= β2**=0.5)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.612 (0.902) | -0.612 (0.981) | -0.614 (1.048) | -0.569 (0.857) | -0.454 (0.886) |
| % Futility based on Cond. Power (<20%) | 39.7% | 44.7% | 47.3% | 41.2% | 53.5% |
| Bias | -0.158 | -0.158 | -0.160 | 0.115 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -0.654 (0.894) | -0.654 (0.938) | -0.651 (0.976) | -0.637 (0.869) | -0.627 (0.883) |
| % Futility based on Cond. Power (<20%) | 24.3% | 25.4% | 26.2% | 23.8% | 20.7% |
| Bias | 0.027 | 0.027 | -0.024 | 0.010 | - |
| **Final Analysis** | | | | | |
| **Final Success (P-value <0.05)** | 22.0% | 18.9% | 16.8% | 21.8% | 18.6% |
| **Final Failure (P-value >=0.05)** | 13.9% | 11.0% | 9.7% | 13.3% | 7.2% |
| **Total Futility at Interim Analysis** | 64.0% | 70.1% | 73.5% | 65.0% | 74.2% |
| **Average Sample Size** | 44 | 42 | 41 | 44 | 39 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

**B: DAS28 at Day 56 is lower than expected (**β1= β2**=0.75)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.749 (0.840) | -0.745 (0.925) | -0.748 (0.995) | -0.702 (0.8815) | -0.654 (0.886) |
| % Futility based on Cond. Power (<20%) | 25.1% | 30.8% | 34.4% | 27.1% | 35.7% |
| Bias | -0.095 | -0.091 | -0.094 | -0.048 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -0.781 (0.865) | -0.786 (0.910) | -0.788 (0.948) | -0.766 (0.852) | -0.773 (0.884) |
| % Futility based on Cond. Power (<20%) | 16.5% | 17.9% | 19.8% | 16.3% | 15.3% |
| Bias | -0.008 | -0.013 | -0.015 | 0.007 | - |
| **Final Analysis** | | | | | |
| **Final Success (P-value <0.05)** | 49.0% | 43.6% | 39.4% | 47.8% | 43.6% |
| **Final Failure (P-value >=0.05)** | 9.4% | 7.7% | 6.5 % | 8.8% | 5.4% |
| **Total Futility at Interim Analysis** | 41.6% | 38.7% | 54.2% | 43.4% | 51.0% |
| **Average Sample Size** | 51 | 55 | 47 | 51 | 47 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

**C: DAS28 at Day 56 is lower than expected (**β1= β2**=1.0)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.879 (0.809) | -0.879 (0.897) | -0.881 (0.969) | -0.836 (0.809) | -0.854 (0.886) |
| % Futility based on Cond. Power (<20%) | 14.6% | 19.8% | 23.8% | 16.8% | 20.2% |
| Bias | -0.025 | -0.025 | -0.027 | 0.018 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -0.918 (0.852) | -0.926 (0.898) | -0.931 (0.937) | -0.903 (0.851) | -0.927 (0.886) |
| % Futility based on Cond. Power (<20%) | 7.8% | 10.3% | 11.5% | 8.1% | 8.2% |
| Bias | 0.009 | 0.001 | -0.004 | 0.024 | - |
| **Final Analysis** | | | | | |
| **Final Success (P-value <0.05)** | 74.4% | 67.3% | 62.4% | 72.2% | 69.7% |
| **Final Failure (P-value >=0.05)** | 3.3% | 2.6% | 2.3% | 2.9% | 2.0% |
| **Total Futility at Interim Analysis** | 22.4% | 30.1% | 35.3% | 14.9% | 28.4% |
| **Average Sample Size** | 57 | 54 | 53 | 62 | 55 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

**D: DAS28 at Day 56 is greater than expected (**β1= β2**=1.2)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.986 (0.809) | -0.985 (0.897) | -0.988 (0.969) | -0.942 (0.828) | -1.014 (0.886) |
| % Futility based on Cond. Power (<20%) | 8.9% | 14.3% | 17.8% | 11.1% | 11.6% |
| Bias | 0.028 | 0.029 | 0.026 | 0.072 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -1.034 (0.854) | -1.045 (0.898) | -1.050 (0.935) | -1.019 (0.862) | -1.060 (0.888) |
| % Futility based on Cond. Power (<20%) | 3.8% | 5.4% | 6.7% | 4.1% | 3.9% |
| Bias | 0.026 | 0.015 | 0.010 | 0.041 | - |
| **Final Analysis** | | | | | |
| **Final Success (P-value <0.05)** | 86.6% | 79.6% | 75.0% | 84.3% | 84.1% |
| **Final Failure (P-value >=0.05)** | 0.6% | 0.6% | 0.5% | 0.6% | 0.4% |
| **Total Futility at Interim Analysis** | 12.7% | 19.7% | 24.5% | 15.2% | 15.5% |
| **Average Sample Size** | 60 | 57 | 56 | 59 | 59 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

Table 11.5 Estimated treatment effect (treatment-placebo) and the interim futility probability of predicted model (the prediction of DAS28 at Day 56 based on Day 14), LOCF method and true data from Day 56 (Month 2) assuming data follow an *Emax* curve under alternative hypothesis. The data are simulated assuming parameters are ==0.9,α1=-0.055, α2=-0.105 for placebo and treatment respectively.

**A: DAS28 at Day 56 is lower than expected (**β1= β2**=0.5) and correlation coefficient** ρ=0.1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.615 (0.903) | -0.614 (0.984) | -0.616 (1.050) | -0.571 (0.858) | -0.457 (0.889) |
| % Futility based on Cond. Power (<20%) | 39.0% | 44.0% | 47.0% | 40.5% | 53.6% |
| Bias | -0.158 | -0.157 | -0.159 | -0.114 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -0.632 (0.896) | -0.634 (0.940) | -0.634 (0.976) | -0.616 (0.871) | -0.633 (0.883) |
| % Futility based on Cond. Power (<20%) | 27.0% | 27.2% | 27.5% | 26.2% | 20.3% |
| Bias | 0.001 | -0.001 | -0.001 | 0.017 | - |
| Final Analysis | | | | | |
| **Final Success (P-value <0.05)** | 20.2% | 17.2% | 15.5% | 20.0% | 18.6% |
| **Final Failure (P-value >=0.05)** | 13.8% | 11.6% | 10.0% | 13.3% | 7.5% |
| **Total Futility at Interim Analysis** | 66.0% | 71.2% | 74.5% | 67.7% | 73.9% |
| **Average Sample Size** | 44 | 42 | 41 | 43 | 39 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

**B: DAS28 Response at Day 56 is greater than expected (**β1= β2=**1.2) and rho=0.1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Alternative Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.988 (0.810) | -0.988 (0.901) | -0.990 (0.971) | -0.945 (0.829) | -1.017 (0.889) |
| % Futility based on Cond. Power (<20%) | 9.5% | 14.4% | 17.7% | 11.3% | 11.5% |
| Bias | 0.029 | 0.029 | 0.037 | 0.072 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -1.034 (0.854) | -1.038 (0.898) | -1.039 (0.940) | -1.013 (0.863) | -1.060 (0.889) |
| % Futility based on Cond. Power (<20%) | 4.8% | 6.3% | 8.0% | 5.1% | 4.0% |
| Bias | 0.026 | 0.022 | 0.021 | 0.047 | - |
| **Final Analysis** | | | | | |
| **Final Success (P-value <0.05)** | 84.7% | 78.5% | 73.7% | 82.8% | 83.9% |
| **Final Failure (P-value >=0.05)** | 1.0% | 0.8% | 0.7% | 0.8% | 0.5% |
| **Total Futility at Interim Analysis** | 14.3% | 20.7% | 25.7% | 16.4% | 15.5% |
| **Average Sample Size** | 59 | 57 | 56 | 59 | 59 |

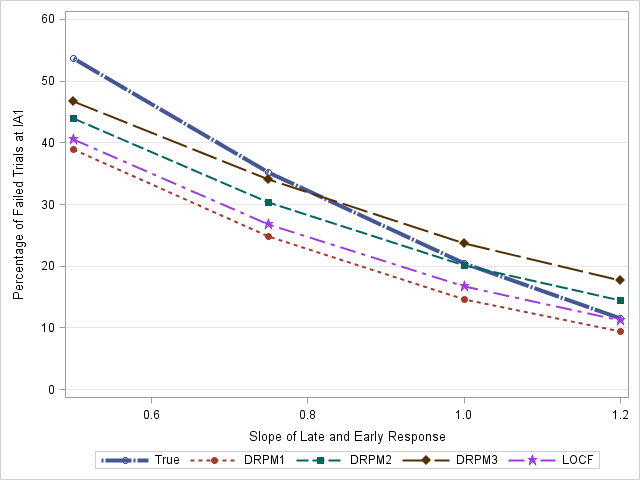
\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.

Table 11.6 Estimated treatment effect (treatment-placebo) and the interim futility probability of predicted model (the prediction of DAS28 at Day 56 based on Day 14), LOCF method and true data from Day 56 (Month 2) assuming data follow a flat curve under null hypothesis. The data are simulated assuming parameters are ==0.9,α1=-0.055, α2=-0.105 for placebo and treatment respectively.

**A: DAS28 at Day 56 is lower than expected (**β1= β2**=0.5) and correlation coefficient** ρ=0.1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interim Analysis** | **Under Null Hypothesis Effect** | | | | |
| **DRPM1** | **DRPM2** | **DRPM3** | **LOCF** | **True** |
| **Interim Analysis 1 after 3 cohorts** | | | | | |
| Average Treatment Effect (SD) | -0.055 (0.824) | -0.054 (0.913) | -0.056 (0.983) | -0.038 (0.813) | -0.057 (0.889) |
| % Futility based on Cond. Power (<20%) | 84.4% | 84.8% | 85.1% | 85.4% | 84.8% |
| Bias | -0.002 | -0.003 | -0.001 | 0.019 | - |
| **Interim Analysis 2 after 6 cohorts \*** | | | | | |
| Average Treatment Effect (SD) | -0.290 (0.858) | -0.269 (0.903) | -0.264 (0.946) | -0.292 (0.853) | -0.379 (0.884) |
| % Futility based on Cond. Power (<20%) | 13.2% | 13.4% | 13.2% | 12.3% | 12.4% |
| Bias | 0.089 | 0.110 | 0.115 | 0.087 | - |
| **Final Analysis** | | | | | |
| **Final Success (P-value <0.05)** | 0.6% | 0.4% | 0.4% | 0.6% | 1.0% |
| **Final Failure (P-value >=0.05)** | 1.8% | 1.4% | 1.3% | 1.7% | 1.8% |
| **Total Futility at Interim Analysis** | 97.6% | 98.2% | 98.3% | 97.7% | 97.2% |
| **Average Sample Size** | 28 | 28 | 28 | 28 | 28 |

\*The 2nd interim analysis is performed for all the non-stopping simulations from 1st interim analysis. Final analysis is performed for the non-stopping simulations from 1st and 2nd interim analyses. Cond. Power is the conditional power. The estimates and operating characteristics are summarized based on 10,000 simulations.



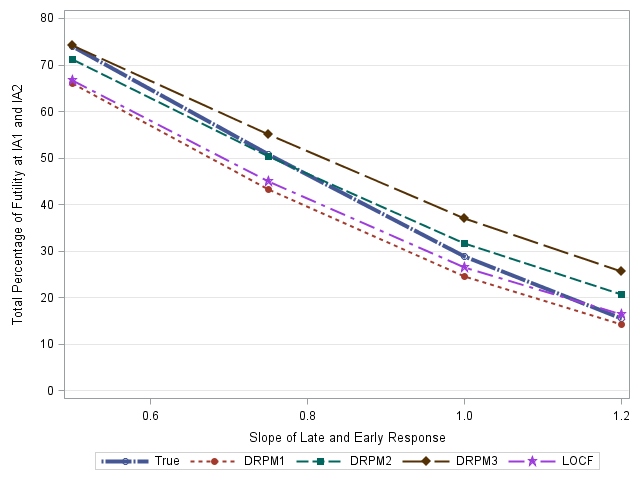
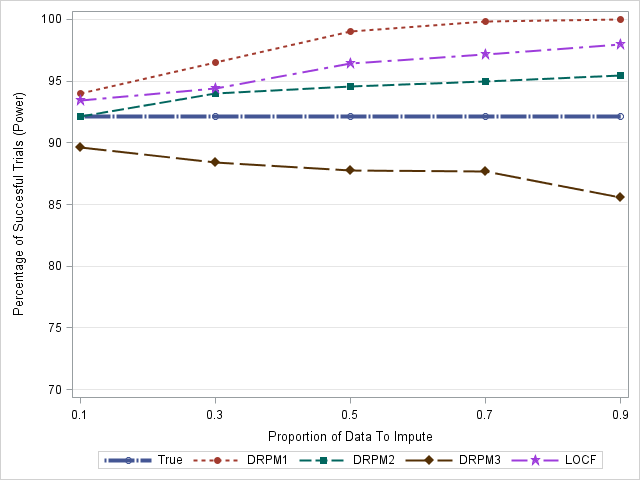


Figure 11.7 Sensitivity of percentage of futility IA1 (top)and IA1/IA2 (bottom) for DRPM1, 2, 3, LOCF, and true model if the slopes of late and early response are lower or higher than expected (β1= β2=0.5, 0.75, 1.0 and 1.2) and correlation coefficient is 0.1 (ρ=0.1). Other parameters are fixed in the simulation=0.9, =0.9,α1=-0.055, α2=-0.105 for placebo and treatment respectively.

### Sensitivity Analysis on Proportion of Data to Impute

In the model evaluation of cohort randomization design in Chapter 8, the first interim analysis was conducted after the third cohort; there were total 16 patients from first 2 cohorts completed DAS28 at Day 56 and 8 patients using predicted Day 56 response from Day 14 visit, so the proportion of data to be imputed or predicted was 33%. At the second interim analysis, the proportion of data to be imputed or predicted was 17%.

The simulation are now further repeated when the proportion of imputed data is higher than expected i.e. the proportion is 50%, 70% or 90% (Figure 11.8 and 11.9). The probability of success or statistical powers of all DRPM and LOCF methods deviate from the true data as the proportion of data being imputed increases.

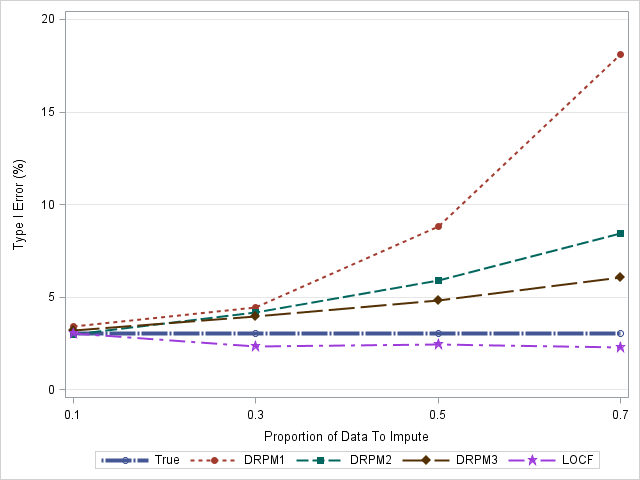


Base

Case

Figure 11.8 Sensitivity plot of statistical power displayed as proportion of data to be imputed ranging from 10% imputation to 90% imputation. The proposed DRPM1, DRPM2, DRPM3 and LOCF method were used in the imputation when the true dose response follows *Emax* curve. Other parameters are fixed in the simulated ρ=0.7 =0.9, =0.9,α1=-0.055, α2=-0.105 for placebo and treatment respectively.

The Type I error rates from the additional simulations are presented in Figure 11.8. The type I error is inflated as the proportion of imputed data increases. If the proportion of imputed data is less than or equal to 30%, the Type I error rate is below 5% for all the predictive models. When the proportion is more than 50%, Type I error rates are more than 5% in all DRPM models. The Type I error is under control for the LOCF method.



Base

Case

Figure 11.9 Type I error rate displayed as proportion of data to be imputed ranging from 10% imputation to 70% imputation. The proposed DRPM1, DRPM2, DRPM3 and LOCF method were used in the imputation when the true dose response follows flat curve under null hypothesis and other parameters are fixed in the simulated ρ=0.7 =0.9, =0.9,α1=-0.055, α2=-0.105 for placebo and treatment respectively.

## Simulation Scenarios Based on Desired Response or Placebo Like Response

The simulations of dose escalation under null effect or desired response are used to understand the percentage of dose allocated to each dose level when the drug has no effect (null effect) and the desired effect (alternative effect, the expected treatment effect assuming *Emax* dose response similar to market drug with ED90 of 3mg/kg) (van de Putte, 2004). The simulation was run in WinBUGS by the GlaxoSmithKline study team. The key results of the simulation are displayed in this section.

Figure 11.10 shows the expected frequency of dose administration in a simulated desired *Emax* effect. Dose levels refer to dose ranges by half log increase, i.e. level 1-11 is 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, and 100 mg/kg respectively. Level 4 (0.03mg/kg) is the starting level of the first cohort and the maximal dose level is level 10 (30 mg/kg). Maximal dose escalation is two levels of dose increase. The dashed line indicates a traditional parallel group design which assigns all doses from Level 4 to Level 10 with equal frequency. Once the ED90 is reached (Level 8), the frequency of this dose is twice as frequent as other doses. Compared with the traditional parallel group design, the Bayesian design, therefore, has a greater probability of assigning doses which are more efficacious (close to ED90, Level 8) but not supratherapeutic (Level 10).

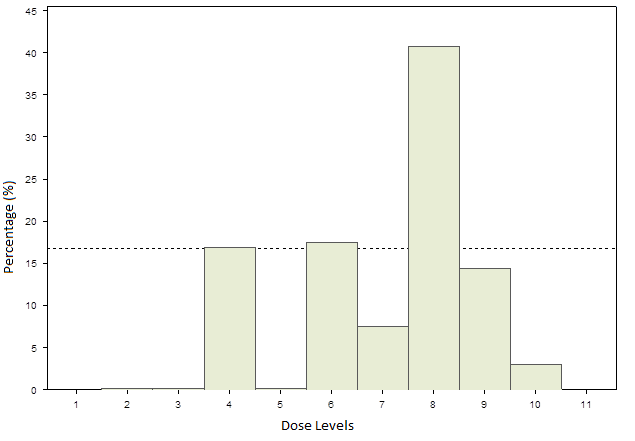


Figure 11.10 Frequency histogram of dose levels administered in a Bayesian adaptive dose-finding trial simulation for desired response (*Emax* dose response similar to market drug) (van de Putte, 2004)

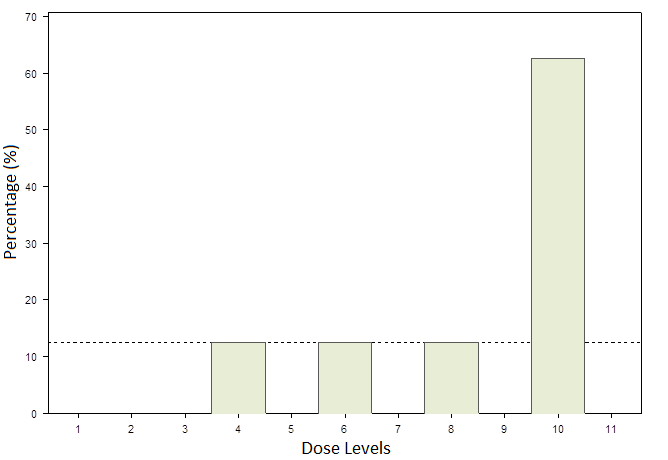


Figure 11.11 Frequency histogram of dose levels administered in a Bayesian adaptive dose-finding trial simulation under null hypothesis.

When a drug has no efficacy (under null hypothesis) (Figure 11.11), the Bayesian dose-finding design escalates quickly to highest dose, Level 10 (30mg/kg) (jumping from Level 4 to 6, Level 6 to 8, and Level 8 to 10) where there is no evidence of clinical benefit. The Bayesian designs assign subsequent cohorts to the highest dose.

Therefore, the Bayesian dose-finding approach was selected as being the most efficient method for identifying the ED90. The starting dose is 0.03mg/kg (Level 4) and dose escalation is planned to proceed to a nominal maximum dose of 30mg/kg (Level 10).

# Reference

Abdi, H., 2007. *The Bonferonni and Šidák Corrections for Multiple Comparisons*. *Encyclopedia of Measurement and Statistics*.

Abrahantes, J.C., Shkedy, Z., and Molenberghs, G., 2008. Alternative methods to evaluate trial level surrogacy. *Clinical Trials*, 5(3), pp.194–208.

Adams, C.P., and Brantner, V.V., 2010. Spending on new drug development1. *Health economics*, 19(2), pp.130–141.

Agresti, A., 2002. *Categorical Data Analysis*. *Wiley series in probability and statistics*, .

Almstrup, K., et al., 2002. Dual effects of phytoestrogens result in u-shaped dose-response curves. *Environmental health perspectives*, 110(8), pp.743–8.

Alonso, A., et al., 2004. Prentice’s Approach and the Meta-Analytic Paradigm: A Reflection on the Role of Statistics in the Evaluation of Surrogate Endpoints. *Biometrics*, 60, pp.724–728.

Alonso, A., et al., 2015. On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics*, 71(1), pp.15–24.

Annpey Pong, S.-C.C., 2010. *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*. CRC Press.

Armitage, P., McPherson, C.K., and Rowe, B.C., 1969. Repeated Significance Tests on Accumulating Data. *Journal of the Royal Statistical Society*, 132(2), pp.235–244.

Arrowsmith, J., and Miller, P., 2013. Trial watch: phase II and phase III attrition rates 2011-2012. *Nature reviews. Drug discovery*, 12(8), p.569.

Baker, D., 2008. The benefits and savings from publicly funded clinical trials of prescription drugs. *International journal of health services : planning, administration, evaluation*, 38(4), pp.731–750.

Barnett, A.G., van der Pols, J.C., and Dobson, A.J., 2004. Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, 34(1), pp.215–220.

Berry, D.A., et al., 2002. *Case Studies in Bayesian Statistics Volume V*. Lecture Notes in Statistics. New York, NY: Springer New York.

Besen, J., and Gan, S.D., 2014. A critical evaluation of clinical research study designs. *The Journal of investigative dermatology*, 134(3), p.e18.

Blance, A., Tu, Y.-K., and Gilthorpe, M.S., 2005. A multilevel modelling solution to mathematical coupling. *Statistical methods in medical research*, 14(6), pp.553–65.

Blomqvist, N., 1977. On the Relation between Change and Initial Value. *Journal of the American Statistical Association*, 72(360a), pp.746–749.

Boers, M., et al., 1997. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. *Lancet*, 350(9074), pp.309–318.

Bornkamp, B., et al., 2007. Innovative Approaches for Designing and Analyzing Adaptive Dose-Ranging Trials. *Journal of Biopharmaceutical Statistics*, 17(6), pp.965–995.

Bretz, F., et al., 2009. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*, 28(8), pp.1181–1217.

Bretz, F., Pinheiro, J.C., and Branson, M., 2005. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61(3), pp.738–48.

van Breukelen, G.J.P., 2013. ANCOVA Versus CHANGE From Baseline in Nonrandomized Studies: The Difference. *Multivariate Behavioral Research*, 48(6), pp.895–922.

Burmester, G.R., and Weinblatt, M.E., 2012. Efficacy and safety of mavrilimumab in subjects with rheumatoid arthritis. *Ann Rheum Dis*, 72(9 SRC-GoogleScholar FG-0), pp.1445–1452.

Burzykowski, T., and Buyse, M., 2006. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical statistics*, 5(3), pp.173–86.

Buyse, M., et al., 2000. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1(1), pp.49–67.

Buyse, M., 2009. Use of Meta-Analysis for the Validation of Surrogate Endpoints and Biomarkers in Cancer Trials. *The Cancer Journal*, 15(5), pp.421–425.

Buyse, M., et al., 2016. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal*, 58(1), pp.104–132.

Buyse, M., and Molenberghs, G., 1998. Criteria for the Validation of Surrogate Endpoints in Randomized Experiments. *Source : Biometrics*, 54(253), pp.1014–1029.

Calabrese, E.J., and Baldwin, L.A., 1997a. A quantitatively‐based methodology for the evaluation of chemical hormesis. *Human and Ecological Risk Assessment: An International Journal*, 3(4), pp.545–554.

Calabrese, E.J., and Baldwin, L.A., 1997b. THE DOSE DETERMINES THE STIMULATION (AND POISON):DEVELOPMENT OF A CHEMICAL HORMESIS DATABASE. *International Journal of Toxicology*, 16(6), pp.545–559.

Calabrese, E.J., and Baldwin, L.A., 2001a. U-SHAPED DOSE-RESPONSES IN BIOLOGY, TOXICOLOGY, AND PUBLIC HEALTH \*. *Annu. Rev. Public Health*, 22, pp.15–33.

Calabrese, E.J., and Baldwin, L.A., 2001b. U-Shaped Dose-Responses in Biology, Toxicology, and Public Health 1. *Annual Review of Public Health*, 22(1), pp.15–33.

Calabrese, E.J., Baldwin, L.A., and Holland, C.D., 1999. Hormesis: A Highly Generalizable and Reproducible Phenomenon With Important Implications for Risk Assessment. *Risk Analysis*, 19(2), pp.261–281.

Campbell, M.J., Walters, S.J., and Machin, D., 2007. *Medical statistics : a textbook for the health sciences*. Wiley.

Carlin, B.P., and Louis, T.A., 2008. *Bayesian methods for data analysis*. CRC Press.

Carlin, B.P., and Louis, T.A., 2010. *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition*. Taylor & Francis.

Casella, G., and Berger, R.L., 2002. *Statistical inference*. Thomson Learning.

Chevret, S., 2006. *Statistical Methods for Dose-Finding Experiments*. Wiley.

Chiolero, A., Paradis, G., Rich, B., and Hanley, J.A., 2013. Assessing the Relationship between the Baseline Value of a Continuous Variable and Subsequent Change Over Time. *Frontiers in public health*, 1, p.29.

Chow, S.-C., and Chang, M., 2008. Adaptive design methods in clinical trials - a review. *Orphanet journal of rare diseases*, 3, p.11.

Choy, E.H., et al., 2013. Safety, tolerability, pharmacokinetics and pharmacodynamics of an anti- oncostatin M monoclonal antibody in rheumatoid arthritis: results from phase II randomized, placebo-controlled trials. *Arthritis research & therapy*, 15(5), p.R132.

Ciani, O., et al., 2016. Use of surrogate end points in healthcare policy: a proposal for adoption of a validation framework. *Nature reviews. Drug discovery*, 15(7), p.516.

Coffey, C.S., and Kairalla, J.A., 2008. Adaptive Clinical Trials: Progress and Challenges. *Drugs in R and D*, 9(4), pp.229–242.

Vander Cruyssen, B., et al., 2005. DAS28 best reflects the physician’s clinical judgment of response to infliximab therapy in rheumatoid arthritis patients: validation of the DAS28 score in patients under infliximab treatment. *Arthritis research & therapy*, 7(5), pp.R1063–R1071.

Darrow, J.J., Avorn, J., and Kesselheim, A.S., 2014. New FDA Breakthrough-Drug Category - Implications for Patients. *The New England Journal of Medicine*, 370(13), pp.1252–1258.

DeMets, D.L., and Lan, K.K., n.d. Interim analysis: the alpha spending function approach. *Statistics in medicine*, 13(13–14), pp.1341-52–6.

Dhir, V., et al., 2014. Randomized controlled trial comparing 2 different starting doses of methotrexate in rheumatoid arthritis. *Clinical Therapeutics*, 36(7), pp.1005–1015.

Dmitrienko, A., Chuang-Stein, C., and D’Agostino, R.B., 2007. *Pharmaceutical Statistics Using SAS: A Practical Guide*. SAS Institute.

Dmitrienko, A., Offen, W., Wang, O., and Xiao, D., 2006. Gatekeeping procedures in dose–response clinical trials based on the Dunnett test. *Pharmaceutical Statistics*, 5(1), pp.19–28.

Dmitrienko, A., Offen, W.W., and Westfall, P.H., 2003. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in medicine*, 22(15), pp.2387–400.

Dragalin, V., Hsuan, F., and Padmanabhan, S.K., 2007. Adaptive Designs for Dose-Finding Studies Based on Sigmoid *E* max Model. *Journal of Biopharmaceutical Statistics*, 17(6), pp.1051–1070.

Dunnett, C., 1955. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50, pp.1096–1121.

Dwass, M., 1957. Modified Randomization Tests for Nonparametric Hypotheses. *The Annals of Mathematical Statistics*, 28(1), pp.181–187.

Edland, S.D., n.d. Blomqvist revisited: how and when to test the relationship between level and longitudinal rate of change. *Statistics in medicine*, 19(11–12), pp.1441–52.

Edwards, J.C.W., et al., 2004. Efficacy of B-cell-targeted therapy with rituximab in patients with rheumatoid arthritis. *The New England journal of medicine*, 350(25), pp.2572–2581.

EMEA, 2006. Reflection paper on methodolofical issues in confirmatory clinical trials with flexible design and analysis plan.

EMEA Guideline, 2005. Guideline on data monitoring committees.

EMEA Guideline, 2011. Guideline on clinical investigation of medicinal products other than NSAIDs for treatment of rheumatoid arthritis.

FDA Draft Guidance, 1994. Dose response information to support drug registration.

FDA Draft Guidance, 1999. Guidance for Industry Population Pharmacokinetics.

FDA Draft Guidance, 2013. *Guidance for Industry Rheumatoid Arthritis: Developing Drug Products for Treatment*.

Felson, D.T., et al., 1993. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. *Arthritis and rheumatism*, 36(6), pp.729–740.

Felson, D.T., et al., 1995. American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. *Arthritis and Rheumatism*, 38(6), pp.727–735.

Fiehn, C., et al., 2007. Eight versus 16-week re-evaluation period in rheumatoid arthritis patients treated with leflunomide or methotrexate accompanied by moderate dose prednisone. *Rheumatology International*, 27(10), pp.975–979.

Fisher, R.A., 1935. *The Design of Experiments, 3rd Edition*. Oliver & Boyd, London.

Fleischmann, R.M., et al., 2015. A Randomized, Double-Blind, Placebo-Controlled, Twelve-Week, Dose-Ranging Study of Decernotinib, an Oral Selective JAK-3 Inhibitor, as Monotherapy in Patients With Active Rheumatoid Arthritis. *Arthritis & Rheumatology*, 67(2), pp.334–343.

Fransen, J., Stucki, G., et al., 2003. Rheumatoid Arthritis Measures. Disease Activity Score (DAS), Disease Activity Score-28 (DAS28), Rapid Assessment of Disease Activity in Rheumatology (RADAR), and Rheumatoid Arthritis Disease Activity Index (RADAI). *Arthritis & Rheumatism*, 49, pp.S214–S224.

Fransen, J., and Van Riel, P.L.C.M., 2005. The Disease Activity Score and the EULAR response criteria. *Clinical and Experimental Rheumatology*, 23(5 SUPPL. 39), pp.S93-9.

Freedman, L.S., Graubard, B.I., and Schatzkin, A., 1992. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, 11(2), pp.167–178.

Fu, H., and Manner, D., 2010. Bayesian adaptive dose-finding studies with delayed responses. *Journal of biopharmaceutical statistics*, 20(5), pp.1055–1070.

G., L.K.K., and L, D.D., 1983. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3), pp.659–663.

Gallo, P., and Maurer, W., 2006. Challenges in implementing adaptive designs: Comments on the viewpoints expressed by regulatory statisticians. *Biometrical Journal*, 48(4), pp.591–597.

Gallo P, 2006. Operational challenges in adaptive design implementation. *Pharm Stat.* , 5(2), pp.119–24.

Genovese, M.C., et al., 2008. Interleukin-6 receptor inhibition with tocilizumab reduces disease activity in rheumatoid arthritis with inadequate response to disease-modifying antirheumatic drugs: the tocilizumab in combination with traditional disease-modifying antirheumatic drug the. *Arthritis and rheumatism*, 58(10), pp.2968–80.

Genovese, M.C., et al., 2013. Efficacy and safety of secukinumab in patients with rheumatoid arthritis: a phase II, dose-finding, double-blind, randomised, placebo controlled study. *Annals of the rheumatic diseases*, 72(6), pp.863–9.

Genovese, M.C., et al., 2014. Efficacy and safety of olokizumab in patients with rheumatoid arthritis with an inadequate response to TNF inhibitor therapy: outcomes of a randomised Phase IIb study. *Annals of the rheumatic diseases*, 73(9), pp.1607–15.

Geusens, P., and Lems, W., 2008. Efficacy and tolerability of lumiracoxib, a highly selective cyclo-oxygenase-2 (COX2) inhibitor, in the management of pain and osteoarthritis. *Therapeutics and Clinical Risk Management*, 4(2), pp.337–344.

Good, P., 1994. *Permutation Tests*. Springer Series in Statistics. New York, NY: Springer New York.

Grieve, A.P., and Krams, M., 2005. ASTIN: a Bayesian adaptive dose-response trial in acute stroke. *Clinical trials (London, England)*, 2(4), pp.340-351-358, 364–378.

Grigor, C., et al., 2004. Effect of a treatment strategy of tight control for rheumatoid arthritis (the TICORA study): a single blind randomised controlled trial. *Lancet*, 354(9430), pp.263–9.

Gruttola, V. De, and Tu, X., 1994. Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*.

Haines, L.M., Perevozskaya, I., and Rosenberger, W.F., 2003. Bayesian optimal designs for Phase I clinical trials. *Biometrics*, 59(3), pp.591–600.

Hardwick, J., Oehmke, R., and Stout, Q.F., 2006. New adaptive designs for delayed response models. *Journal of Statistical Planning and Inference*, 136(6), pp.1940–1955.

Hay, M., et al., 2014. Clinical development success rates for investigational drugs. *Nature biotechnology*, 32(1), pp.40–51.

van der Heijde, D.M., et al., 1992. Validity of single variables and composite indices for measuring disease activity in rheumatoid arthritis. *Annals of the rheumatic diseases*, 51(2), pp.177–181.

Hills, R.K., and Burnett, A.K., 2011. Applicability of a ‘Pick a Winner’ trial design to acute myeloid leukemia. *Blood*, 118(9), pp.2389–94.

Hochberg, M.C., Silman, A.J., Smolen, J.S., and Weinblatt, Michael E. Weisman, M.H., 2009. *Rheumatoid Arthritis*.

Hsu, J.C., and Berger, R.L., 1999. Stepwise Confidence Intervals without Multiplicity Adjustment for Dose-Response and Toxicity Studies. *Journal of the American Statistical Association*, 94(446), pp.468–469.

Hung, H.M.J., Cui, L., Wang, S.-J., and Lawrence, J., 2005. Adaptive statistical analysis following sample size modification based on interim review of effect size. *Journal of biopharmaceutical statistics*, 15(4), pp.693–706.

Hung, H.M.J., Neill, R.T.O., Wang, S., and Lawrence, J., 2006. A Regulatory View on Adaptive / Flexible Clinical Trial Design. *Biometrical Journal*, 48, pp.565–573.

ICH, 1994. E4: Dose-response information to support drug registration.

ICH, 1997. E6: Good Clinical Practice: Consolidated guideline.

ICH, 1998a. E9:Statistical Principles for Clinical Trials.

ICH, 1998b. Topic S 6: Preclinical Safety Evaluation of Biotechnology-Derived Pharmaceuticals.

Jennison, C., and Turnbull, B.W., 1999. *Group Sequential Methods with Applications to Clinical Trials*. CRC Press.

Julious, S., 2015. Case Studies and Exemplars of Efficient Study Designs in Clinical Trials. In: *Society for Clincal Trials Meeting*.

Kairalla, J.A., Coffey, C.S., Thomann, M.A., and Muller, K.E., 2012. Adaptive trial designs: a review of barriers and opportunities. *Trials*, 13(1), p.145.

Kimpel, D., 2007. Effects of Abatacept in Patients with Methotrexate-Resistant Active Rheumatoid Arthritis: A Randomized Trial Kremer JM, Genant HK, Moreland LW, et al ( …. *Year Book of Medicine*, 144(12), pp.865–876.

Koopmeiners, J.S., and Modiano, J., 2014. A Bayesian adaptive Phase I-II clinical trial for evaluating efficacy and toxicity with delayed outcomes. *Clinical trials (London, England)*, 11(1), pp.38–48.

Krams, M., et al., 2003. Acute Stroke Therapy by Inhibition of Neutrophils (ASTIN): an adaptive dose-response study of UK-279,276 in acute ischemic stroke. *Stroke; a journal of cerebral circulation*, 34(11), pp.2543–8.

Krams, M., et al., 2007. Adaptive designs in clinical drug development: opportunities, challenges, and scope reflections following PhRMA’s November 2006 workshop. *Journal of biopharmaceutical statistics*, 17(6), pp.957–964.

Kutner, M.H., Nachtsheim, C., Neter, J., and Li, W., 2005. *Applied linear statistical models*. McGraw-Hill Irwin.

Leonov, S., and Miller, S., 2009. An adaptive optimal design for the E(max) model and its application in clinical trials. *Journal of biopharmaceutical statistics*, 19(2), pp.360–385.

Li, G., Shih, W.J., Xie, T., and Lu, J., 2002. A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics (Oxford, England)*, 3(2), pp.277–287.

Li, J., and Fu, H., 2013. Bayesian adaptive D-optimal design with delayed responses. *Journal of biopharmaceutical statistics*, 23(3), pp.559–68.

Li, W., and Fu, H., 2011. Bayesian optimal adaptive designs for delayed-response dose-finding studies. *Journal of biopharmaceutical statistics*, 21(5), pp.888–901.

Lionberger, R.A., 2008. FDA critical path initiatives: opportunities for generic drug development. *The AAPS journal*, 10(1), pp.103–9.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D., 2000. WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, pp.325–337.

Marcus, R., Eric, P., and Gabriel, K.R., 1976. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3), pp.655–660.

Matsui, T., et al., 2007. Disease Activity Score 28 (DAS28) using C-reactive protein underestimates disease activity and overestimates EULAR response criteria compared with DAS28 using erythrocyte sedimentation rate in a large observational cohort of rheumatoid arthritis patients . *Annals of the rheumatic diseases*, 66(9), pp.1221–1226.

Mattson, M.P., and Calabrese, E.J., 2010. *Hormesis : a revolution in biology, toxicology, and medicine*. Springer.

McKenzie, J.E., Herbison, G.P., and Deeks, J.J., 2016. Impact of analysing continuous outcomes using final values, change scores and analysis of covariance on the performance of meta-analytic methods: a simulation study. *Research Synthesis Methods*, 7(4), pp.371–386.

Mease, P., et al., 2012. A phase II, double-blind, randomised, placebo-controlled study of BMS945429 (ALD518) in patients with rheumatoid arthritis with an inadequate response to methotrexate. *Annals of the Rheumatic Diseases*, 71(7), pp.1183–1189.

Milligan, P. a, et al., 2013. Model Based Drug Development : a Rational Approach to Efficiently Accelerate Drug Development. *Clinical pharmacology and therapeutics*, 93(March), pp.502–14.

Möttönen, T., et al., 1996. Outcome in patients with early rheumatoid arthritis treated according to the ‘sawtooth’ strategy. *Arthritis and rheumatism*, 39(6), pp.996–1005.

Möttönen, T., et al., 1998. Only high disease activity and positive rheumatoid factor indicate poor prognosis in patients with early rheumatoid arthritis treated with ‘sawtooth’ strategy. *Annals of the rheumatic diseases*, 57(9), pp.533–539.

Newman, K.B., 2014. *Modelling Population Dynamics: Model Formulation, Fitting and Assessment Using State-space Methods*. Springer.

Newman, S.P., 1996. *Understanding Rheumatoid Arthritis*. Psychology Press.

NHS Specification Document, 2014, Funding for Primary Research Using Efficient Study Designs to Evaluate Clinical and Public Health Interventions for the NHS

Nishimoto, N., et al., 2007. Study of active controlled monotherapy used for rheumatoid arthritis, an IL-6 inhibitor (SAMURAI): evidence of clinical and radiographic benefit from an x ray reader-blinded randomised controlled trial of tocilizumab. *Annals of the rheumatic diseases*, 66(9), pp.1162–1167.

Nishimoto, N., and Takagi, N., 2010. Assessment of the validity of the 28-joint disease activity score using erythrocyte sedimentation rate (DAS28-ESR) as a disease activity index of rheumatoid arthritis in the efficacy evaluation of 24-week treatment with tocilizumab: Subanalysis of the SAT. *Modern Rheumatology*, 20(6), pp.539–547.

Nixon, R.M., et al., 2009. Using short-term evidence to predict six-month outcomes in clinical trials of signs and symptoms in rheumatoid arthritis. *Pharmaceutical Statistics*, 8(2), pp.150–162.

O’Brien, P.C., and Fleming, T.R., 1979. A multiple testing procedure for clinical trials. *Biometrics*, 35(3), pp.549–556.

O’Hagan, A., Stevens, J.W., and Campbell, M.J., 2005. Assurance in clinical trial design. *Pharmaceutical Statistics*, 4(3), pp.187–201.

Oldham, P.D., 1962. A note on the analysis of repeated measurements of the same subjects. *Journal of chronic diseases*, 15, pp.969–77.

Owen, S.C., et al., 2014. Colloidal Drug Formulations Can Explain ‘Bell-Shaped’ Concentration–Response Curves. *ACS Chemical Biology*, 9(3), pp.777–784.

Petris, G., Campagnoli, P., and Petrone, S., 2009. *Dynamic Linear Models with R*. New York, NY: Springer New York.

Phipson, B., Smyth Walter, G.K., and Hall, E., 2010. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology on Stat. Appl. Genet. Molec. Biol*, 31(1).

Piedbois, P., and Buyse, M., 2004. Meta-Analyses Based on Abstracted Data: A Step in the Right Direction, but Only a First Step. *Journal of Clinical Oncology*, 22(19), pp.3839–3841.

Piedbois, P., and Buyse, M., 2008. Endpoints and surrogate endpoints in colorectal cancer: a review of recent developments. *Current Opinion in Oncology*, 20(4), pp.466–471.

Pincus, T., and Sokka, T., 2004. Clinical trials in rheumatic diseases: Designs and limitations. *Rheumatic Disease Clinics of North America*, 30(4), pp.701–724.

Pinheiro, J., Bornkamp, B., and Bretz, F., 2006. Design and Analysis of Dose-Finding Studies Combining Multiple Comparisons and Modeling Procedures. *Journal of Biopharmaceutical Statistics*, 16(5), pp.639–656.

Pinheiro, J.C., and DeMets, D.L., 1997. Estimating and reducing bias in group sequential designs with Gaussian independent increment structure. *Biometrika*, 84(4), pp.831–845.

Pocock, S.J., 1977. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2), pp.191–199.

Pocock, S.J., 1982. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics*, 38(1), pp.153–162.

Prentice, R.L., 1989. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4), pp.431–440.

Proschan, M. a, and Hunsberger, S. a, 1995. Designed extension of studies based on conditional power. *Biometrics*, 51(4), pp.1315–1324.

Quinn, M.A., et al., 2005. Very early treatment with infliximab in addition to methotrexate in early, poor-prognosis rheumatoid arthritis reduces magnetic resonance imaging evidence of synovitis and damage, with sustained benefit after infliximab withdrawal: Results from a twelve-m. *Arthritis and Rheumatism*, 52(1), pp.27–35.

Rader, R.A., 2008. (Re)defining biopharmaceutical. *Nature Biotechnology*, 26(7), pp.743–751.

Reynolds, A.R., 2009. Potential relevance of bell-shaped and u-shaped dose-responses for the therapeutic targeting of angiogenesis in cancer. *Dose-response : a publication of International Hormesis Society*, 8(3), pp.253–84.

Rintelen, B., Andel, I., Sautner, J., and Leeb, B.F., 2006. Leflunomide/chloroquin combination therapy in rheumatoid arthritis: A pilot study. *Clinical Rheumatology*, 25(4), pp.557–559.

Schiff, M.H., et al., 2014. Rheumatoid arthritis secondary non-responders to TNF can attain an efficacious and safe response by switching to certolizumab pegol: a phase IV, randomised, multicentre, double-blind, 12-week study, followed by a 12-week open-label phase. *Annals of the rheumatic diseases*, 73(12), pp.1–4.

Senn, S., 2006. Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, 25(24), pp.4334–4344.

Sherman, R.E., et al., 2013. Expediting Drug Development — The FDA’s New ‘Breakthrough Therapy’ Designation. *New England Journal of Medicine*, 369(20), pp.1877–1880.

Sheskin, D., 2007. *Handbook of parametric and nonparametric statistical procedures 4th Edition*.

Shih, W.J., and Zhao, P.L., 1997. Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes. *Statistics in Medicine*, 16(17), pp.1913–1923.

Shirley, E., 1977. A non-parametric equivalent of Williams’ test for contrasting increasing dose levels of a treatment. *Biometrics*, 33(2), pp.386–389.

Shuster, J.J., 1990. *CRC Handbook of Sample Size Guidelines for Clinical Trials*. CRC Press.

Silva V, Grande AJ, Porfírio GJM, Puga MES, Martimbianco ALC, Torres MFS, Costa MB, Carvalho MR, Fioretti BTDS, Costa CS, Peccin MS, Riera R, Torloni MR, A.A., 2013. Statistical simulation to assess results of meta-analyses using post-intervention, change from baseline and mixed methods | Colloquium Abstracts. *In: Better Knowledge for Better Health | Un meilleur savoir pour une meilleure santé. Abstracts of the 21st Cochrane Colloquium;* .

Simon, R., 1989. Optimal two-stage designs for phase II clinical trials. *Controlled clinical trials*, 10(1), pp.1–10.

Smith, M.K., et al., 2006. Implementation of a Bayesian adaptive design in a proof of concept study. *Pharmaceutical Statistics*, 5(1), pp.39–50.

Smolen, J.S., et al., 2013. Maintenance, reduction, or withdrawal of etanercept after treatment with etanercept and methotrexate in patients with moderate rheumatoid arthritis (PRESERVE): A randomised controlled trial. *The Lancet*, 381(9870), pp.918–929.

Smolen, J.S., et al., 2014. Sirukumab, a human anti-interleukin-6 monoclonal antibody: a randomised, 2-part (proof-of-concept and dose-finding), phase II study in patients with active rheumatoid arthritis despite methotrexate therapy. *Annals of the rheumatic diseases*, 73(9), pp.1616–25.

Spiegelhalter, D.J., Abrams, K.R., and Myles, J.P., 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons.

Stallard, N., 2011. Group-sequential methods for adaptive seamless phase II/III clinical trials. *Journal of biopharmaceutical statistics*, 21(4), pp.787–801.

Stohl, W., et al., 2013. Efficacy and Safety of Belimumab in Patients with Rheumatoid Arthritis: A Phase II, Randomized, Double-blind, Placebo-controlled, Dose-ranging Study. *The Journal of Rheumatology*, 40(5), pp.579–589.

Temple, J., 2012. *Adaptive Designs for Dose-Finding Trials*. University of Bath.

Temple, J., 2013. *Bayesian Adaptive Design, Design and Analysis*. Jennison C.

Thomas, N., 2006. Hypothesis testing and Bayesian estimation using a sigmoid Emax model applied to sparse dose-response designs. *Journal of biopharmaceutical statistics*, 16(5), pp.657–677.

Thomas, N., Sweeney, K., and Somayaji, V., 2014. Meta-Analysis of Clinical Dose–Response in a Large Drug Development Portfolio. *Statistics in Biopharmaceutical Research*, 6(4), pp.302–317.

Ting, N., 2006. *Dose Finding in Drug Development*. Springer Science & Business Media.

Tu, Y.-K., and Gilthorpe, M.S., 2007. Revisiting the relation between change and initial value: a review and evaluation. *Statistics in Medicine*, 26(2), pp.443–457.

Vanags, D., et al., 2006. Therapeutic efficacy and safety of chaperonin 10 in patients with rheumatoid arthritis: a double-blind randomised trial. *Lancet*, 368(9538), pp.855–863.

Verrier, D., Sivapregassam, S., and Solente, A.-C., 2014. Dose-finding studies, MCP-Mod, model selection, and model averaging: Two applications in the real world. *Clinical Trials*, 11(4), pp.476–484.

Vickers, A.J., and Altman, D.G., 2001. Statistics notes: Analysing controlled trials with baseline and follow up measurements. *BMJ (Clinical research ed.)*, 323(7321), pp.1123–4.

van Vollenhoven, R., Harju, a, Brannemark, S., and Klareskog, L., 2003. Treatment with infliximab (Remicade) when etanercept (Enbrel) has failed or vice versa: data from the STURE registry showing that switching tumour necrosis factor alpha blockers can make sense. *Annals of the rheumatic diseases*, 62(12), pp.1195–1198.

Weir, C.J., Spiegelhalter, D.J., and Grieve, A.P., 2007. Flexible design and efficient implementation of adaptive dose-finding studies. *Journal of biopharmaceutical statistics*, 17(6), pp.1033–1050.

Westfall, P.H., et al., 2000. *Multiple Comparisons and Multiple Tests Text and Workbook Set*. SAS INST (NC).

Williams, D.A., 1971. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, 27(1), pp.103–117.

Williams, D.A., 1972. The comparison of several dose levels with a zero dose control. *Biometrics*, 28(2), pp.519–531.

Wittes, J., and Brittain, E., 1990. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in medicine*, 9(1–2), pp.65-71–2.

Woodcock, J., 2005. FDA introductory comments: clinical studies design and evaluation issues. *Clinical trials (London, England)*, 2(4), pp.273–275.

Wu J, Banerjee A, Jin B, Menon SM, Martin SW, Heatherington AC. Clinical dose–response for a broad set of biological products: A model-based meta-analysis. *Stat. Methods Med. Res.* [Internet]. 2017 Jan

Yamanaka, H., et al., 2011. Efficacy and tolerability of tocilizumab in rheumatoid arthritis patients seen in daily clinical practice in Japan: Results from a retrospective study (REACTION study). *Modern Rheumatology*, 21(2), pp.122–133.

Yazici, Y., et al., 2012. Efficacy of tocilizumab in patients with moderate to severe active rheumatoid arthritis and a previous inadequate response to disease-modifying antirheumatic drugs: the ROSE study. *Annals of the Rheumatic Diseases*, 71(2), pp.198–205.