Optimal Designs for Matching Adjusted Indirect Comparison (MAIC)

by

by Xiang Zheng

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Statistics
University of Manitoba
Winnipeg

Abstract

The thesis aimed to develop a method for optimizing design subject to matching the pre-defined baseline characteristics in clinical trials. As part of a clinical trial, a new treatment must be compared with a competitor treatment in order to determine its effect on the patient before the new treatment is launched. Ideally, we can directly compare the new treatment with competitor treatment in randomized controlled trials (RCTs). However, direct comparison is difficult to achieve due to various factors, such as time, price, regulation, and patents. A matching-adjusted indirect comparison (MAIC) method leverages all available data by adjusting average patient characteristics in trials with Individual patient data (IPD) to match those reported in the aggregate trials data (AgD). MAIC is a reweighting method in which the weights are calculated by deriving the propensity scores in the Individual patient's data. This can reduce the bias.

As IPD matches to the pre-defined baseline characteristics, we make use of optimal design theory and convert this into a constrained optimization problem. The Lagrangian method is used to determine the optimal design subject to satisfying the constraints of baseline characteristics. We formulate the Lagrangian and then transform the constrained problem to one where we simultaneously maximize several functions of the design weights. These functions have a common maximum of zero. In order to find the optimal design, we used the software R and a class of multiplicative algorithms. We then perform a sensitivity analysis and compare the Lagrangian method and the MAIC method by calculating the effective sample sizes (ESS). The higher the value of ESS the less information is lost due to reweighting.

The Lagrangian method performs better than the MAIC method.

Owing to the broad applicability of optimal design, we have tried to make use of this theory in order to obtain a better methodology in MAIC. The proposed methodology is quite flexible and can be applied to different types of constraints. The methodology can be applied to situation where there is a lack of direct comparison. It will also reduce the time and cost of running experiments.

Acknowledgements

First of all, I would like to express my deepest and heartily gratitude to my supervisor, Dr. Saumen Mandal, for his huge support of my study and research, for his patience, suggestions, and immense knowledge. Without his guidance and persistent help this thesis would not have been successfully completed.

Secondly, I sincerely thank Anh Nam Tran, who inspired me in my research topics, and for his generous support and assistance for my thesis.

Thirdly, I would like to thank my committee members, Dr. Po Yang, Dr. Yuliya Martsynyuk, for monitoring my work, spending time in reading and providing me with valuable comments to widen my view and deepen my understanding of the research topic.

I also would like to thake this opportunity to give deep thanks to all the professors and instructors of the Department of Statistics at University of Manitoba for their help and wonderful course teaching. Besides, I am thankful to the support staff of the department Ms. Margaret Smith, Ms. Rosa Di Noto for taking care of the administrative work.

Finally, I take this opportunity to express my profound gratitude to my mom and my wife for providing me with support and continuous encouragement throughout the process of researching and writing this thesis. Without their confirmed support, I would never have enough abilities to finish my thesis.

Dedication

This dissertation is dedicated to my mom, Li Xu, my wife, Krystal Zhang and my daughter to be born baby Caroline Zheng.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Methodologies of experimental design were first proposed by Sir Ronald Fisher in the 1920s and early 1930s. Since then, they have been applied broadly in many areas (Montgomery, 2017). Researchers seek answers for various problems in their research areas of interest based on experimental designs. A design is characterized by variables of interest. For example, we compare two drugs for a specific disease, recruit patients and assign them to a treatment group for each drug. For each patient, we consider baseline characteristics, such as age, gender, weight, etc. How to specify values or levels of those variables such that the design gives less bias and best estimation of the parameters of interest is an important question to design an experiment. After specifying the values or levels of the variables (called the design variables), we need to decide how many observations are assigned to each value or level. Answering these two questions is fundamental in optimal design, which is the main focus of this study.

1.1.1 Optimal Design

When designing an experiment, it is essential to ensure the design provides good estimation of the parameters of the model. A good design guarantees capturing the actual treatment effect precisely without needlessly wasting materials and subjects. In statistics, there are many problems calculating one or more optimizing probability distributions. We can obtain a high-efficient design utilizing methodologies of optimal design theories (see, e.g., Berger and Wong (2009), Silvey (2013), Pukelsheim (2006), Berger and Wong (2009), Fedorov (2013), Torsney (1977), Mandal and Torsney (2006)). It is an important aspect of many statistical research areas to study how to construct optimal designs. Examples include optimal response-adaptive design, maximum likelihood estimation, response surface design, model selection and discrimination, stratified sampling, image processing and optimal structure design. One of the essential aspects is constructing optimal design subject to specific constraints of interest. It evolved into a very crucial topic because the constrained optimization technique can be applied to any circumstances where the restrictions are needed. Once a statistical model is specified based on the objectives of the study, the optimal design is identified based on pre-defined criteria. We denote an optimal design criterion by ϕ, which could be a function of the variance of a parameter estimator of a statistical model, or the generalized variance of the parameter estimators or simply the average variance. Reducing variance yields good estimates, which is desirable. Thus, selecting a design with a minimal variance is an important objective which is also achieved via optimizing functions called optimality criteria. Atkinson et al. (2007) categorized optimal design criteria based on a range of perspectives, which is

explained in detail in the next chapter, comprising of D-, A-, G- and E-optimality.

1.1.2 Matching-Adjusted Indirect Comparison

In clinical trials, optimal design can help investigators achieve higher quality results for the given resource constraints (Sverdlov et al., 2020). Before a new medical treatment comes out, it requires comparing the treatment effect between the new treatment with the competitor treatment. Ideally, we can apply direct comparison between new treatment and the competitor treatment in randomized controlled trials (RCTs). However, direct comparison is difficult to achieve due to various factors, such as time, price, regulation, patents and other constraints. Bucher et al. (1997) first presented an indirect statistical method that adjusted the indirect comparison of treatment effects without head-to-head randomized trials. The method compares the magnitude of the treatment effects between two treatments relative to a common competitor, which serves as a link between the two treatments. Consequently, indirect treatment comparisons are becoming more common in the field (Deodhar, 2018). When a common competitor arm between RCTs (typically placebo) is available, the network meta-analysis method can be performed. Network meta-analysis deals with multiple treatment comparison, and it combines information from all randomized comparisons among a set of treatments that are considered in a given medical condition. In fact, network meta-analysis is a generalization of a pairwise meta-analysis, and has been a popular research topic recently. This method has been very useful in evidence-based medicine (Nietert et al., 2013). However, this method relies on analyzing aggregate data, and it must assume that the trial populations are similar.

This can lead to potential bias in the presence of uneven treatment effect modifiers (Deodhar, 2018). Signorovitch et al. (2010) proposed a matching-adjusted indirect comparison method that leverages all available data by adjusting average patient characteristics in trials with individual patient's data (IPD) in order to match those reported for trials without IPD. This method can avoid some limitations, such as disparities between populations, lack of common competitors, sensitivity to modelling assumptions and different definitions or reporting methods for outcome measures (Dias et al., 2013). Matching-adjusted indirect comparison (MAIC) is the most commonly used approach for population adjustment, and it has been acknowledged as a valid and robust method when comparing differences between trial populations. For more information, we refer to Rugo et al. (2021), Signorovitch et al. (2010), Deodhar (2018), Phillippo et al. (2016), and Phillippo et al. (2018).

1.2 Motivation and Objective

In practice, people are not only interested in minimizing covariance matrix to construct optimal design but also want to match their baseline characteristics to be balanced with those of the patients from the aggregate data of the competitor's trails. However, in practice, individual competitor data is hard to obtain. For new medical treatments and pharmaceutical product studies, due to cost or competition issues, there is always only a small amount of reference data available to assess a product's or treatment's impact on patient outcomes. There are many reasons for lack of available data on a common competitor's treatment against two competing

interventions. Planning clinical trials can be very expensive, investigators need to run significant lengths of time to finish the plan, execution and analysis. Due to policy and many reasons that may change in the course of running a lengthy clinical, there will be a lot of uncertainty that can hinder to identify the common competitor. Also, in some cases, it cannot be assured that assigning patients randomly to a particular treatment arm of clinical trials is ethical. Ideally, the new medical treatments and pharmaceutical products will follow current standards of care to evaluate and to test against all relevant alternatives from marketing in randomized trials. However, this process may take several years before a direct study is performed, and it will delay healthcare decision-makers in approving new treatments for reimbursement. The new technique, Matching-Adjusted Indirect Comparison (MAIC), was published in 2010 (Signorovitch et al., 2010) to solve this problem. Matching-adjusted indirect comparison is a reweighting method in which the weights are calculated by deriving propensity scores in the individual patient's data (IPD) to match summaries reported (typically using mean) in the aggregate trials data (AgD). This can reduce the bias in the results. IPD are such as age and gender characteristics data. In AgD, the summaries reported data can be the mean of age and proportion of gender of aggregate trials data. After matching, each IPD is assigned a re-weighting weight that we denote by w^*. The MAIC method uses individual patient data (IPD) to equalize baseline characteristics across trials from different studies. In one popular criterion (known as D-optimality) in optimal design theory, we minimize the determinant of the covariance matrix of the estimators (generalized variance) of the parameters in a pre-defined statistical model. Equivalently, we maximize the determinant of the

information matrix. In each design point, we assign an optimal design weight, say p^*. Optimal design reduces the time and costs of experimentation. The MAIC method is applied for indirect comparison by matching IPD to AgD. The optimal design theory is applied to reduce the cost of the experiment with fewer experiment runs. We did not find any study that combines these two areas, namely the MAIC and the optimal design.

This motivates us to study this novel unified approach to match the IPD to AgD. We want to create a design that not only satisfies the IPD match to AgD but also the design is optimal. This new methodology can be applied to situation where there is a lack of direct comparison.

Let us consider a problem. Assuming that we have treatment A, we want to know whether treatment A is efficient; we already know that treatment B is efficient, and treatment B is verified its effectiveness through a control group experiment by comparing it with control treatment C. In the ideal case, we know the individual patient data baseline characteristics of AC trials, so we can apply indirect comparisons through a common competitor. However, we do not have individual baseline characteristics of AC trials, such as Age, Gender, Region, etc. Only the aggregated characteristics data of AC trials, such as the mean age of patients and the proportion of males attending, can be found. For easy understanding, we denote the aggregated data as AgD. The available individual patient data is IPD. Under this condition, how can we make a design? The matching adjusted indirect comparison (MAIC) method can be applied in this situation. The idea is to leverage all available individual patient data by adjusting average patient characteristics in our trial to match with AgD. In

optimal design theory, this is called a constraint. We can construct an optimal design subject to this constraint. Now, the problem becomes a constrained optimal design problem.

Construction of constrained optimal design is an important topic from both theoretical and practical point of view. Torsney and Mandal (2001) and Mandal et al. (2005) proposed some approaches that can transform the constrained optimization problem into the problem of maximizing a number of functions of the design weights. Torsney and Mandal (2006) extended their work by applying a class of multiplicative algorithms indexed by a function $f(.)$ to construct optimal design with constraints.

For the MAIC method re-weighted IPD to match the distribution at the aggregate data (AgD), we refer to Rugo et al. (2021) and Alsop and Pont (2022). For the Lagrangian method we refer to the works of Torsney and Mandal (2001), and Mandal et al. (2005). In this thesis, we develop a unified approach by combining the Lagrangian approach and the MAIC method.

1.3 Overview of the Dissertation

This thesis is organized into five chapters. Below is a brief description of each of the remaining four chapters.

Chapter 2 is concerned with optimal design and matching-adjusted indirect comparison methodologies. The first part reviews the existing optimization design methods, related basic knowledge of optimal design framework, and a brief discussion on discrete optimization and related algorithm. The second part reviews the basic

knowledge of matching adjusted indirect comparison procedure.

Chapter 3 shows in detail how to formulate the problems using the Lagrangian approach and how to construct an optimal design subject to the constraints of matching the pre-defined values.

Chapter 4 performs a sensitivity analysis and compares the methods using the effective sample sizes (ESS).

Chapter 5 summarizes the thesis and discusses the limitation and future work.

Chapter 2

Overview of Methodologies and Concepts

2.1 Optimal Design Concepts

Linear models were initially used to develop optimal design theory, and with the development of computer science, there have been significant advances in generating optimal design using algorithms. The optimal design has been implemented in many applications of research areas, such as medical science and marketing research in business (Montgomery, 2017).

Let us begin with the problem of experimental design with a model of the type:

$$y \sim \pi(y|\boldsymbol{x}, \boldsymbol{\theta}, \sigma) \tag{2.1}$$

where y is the response variable and $\boldsymbol{x} = (x_1, x_2, ..., x_m)^T$ are design variables that

can be chosen by the experimenter, $\boldsymbol{x} \in \boldsymbol{\chi} \subseteq \boldsymbol{R}, \boldsymbol{\chi}$ is referred to as the design space. This is typically continuous, but it can also be discrete. $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)^T$ are k-dimensional unknown parameters, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \boldsymbol{R^k}$. σ is a nuisance parameter, which is fixed and unknown and not of primary interest. $\pi(.)$ is a probability model. For each $\boldsymbol{x} \in \boldsymbol{\chi}$, an experiment is performed and the outcome is the response variable $y = y(\boldsymbol{x})$, where $var(y(\boldsymbol{x})) = \sigma^2$ and σ is assumed to be independent with the experimental condition $\boldsymbol{x}$.

In linear model, $y(\boldsymbol{x})$ can be written in the explicit form

$$\boldsymbol{E}(y|\boldsymbol{x}, \boldsymbol{\theta}, \sigma) = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\theta} \tag{2.2}$$

where $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), ..., f_k(\boldsymbol{x}))^T$ is a vector of k real-valued functions. The function $\boldsymbol{f}$ is known to the experimenter.

Now, the question naturally arises what values of $\boldsymbol{x}$ should we select to have the best inference for the parameters $\boldsymbol{\theta}$. Let us suppose that it is on point estimation. It needs first to select n supports points $(x_1, x_2, x_3, ..., x_n) \in \boldsymbol{\chi}$. Suppose the estimator $\hat{\boldsymbol{\theta}}$ is unbiased estimator for $\boldsymbol{\theta}$. In order to have an accuracy of $\hat{\boldsymbol{\theta}}$, we need to ensure the variance is as small as possible. We can write the $k \times k$ dispersion matrix (variance-covariance matrix) as $\boldsymbol{D}(\hat{\boldsymbol{\theta}}) = \boldsymbol{E}([\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}]^T)$. The dispersion contains information about accuracy of not only the diagonal elements of $\hat{\boldsymbol{\theta}}$ but also its off-diagonal elements. In general, smaller $\boldsymbol{D}(\hat{\boldsymbol{\theta}})$ means better accuracy of $\hat{\boldsymbol{\theta}}$.

Assume the model (2.2) is true, and that $\boldsymbol{y_i}$'s are independent random variables

with equal variance σ^2, then the standard linear model is given by:

$$\boldsymbol{E(Y)} = \boldsymbol{X\theta},\ \boldsymbol{D(Y)} = \sigma^2 \boldsymbol{I_n} \tag{2.3}$$

where $\boldsymbol{Y} = (y_1, y_2, ..., y_n)$, $\boldsymbol{X}$ is $n \times k$ design matrix, $\boldsymbol{I_n}$ is $n \times n$ identity matrix. $\boldsymbol{D(Y)}$ denotes the variance-covariance matrix of $\boldsymbol{Y}$.

In the linear model, we can obtain the best linear unbiased estimator (BLUE) using the least squares estimation method. The estimators $\hat{\boldsymbol{\theta}}$ are the solutions of:

$$(\boldsymbol{X^T X})\hat{\boldsymbol{\theta}} = \boldsymbol{X^T Y} \tag{2.4}$$

where $\boldsymbol{X^T X}$ is the information matrix. The larger the matrix $\boldsymbol{X^T X}$, the greater is the information in the experiment. If we suppose all the parameters $\boldsymbol{\theta}$ are interested, then we must ensure that the matrix $\boldsymbol{X^T X}$ is non-singular. Then the unique solution for (2.4) is expressed as:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y} \tag{2.5}$$

with

$$E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta},\quad \boldsymbol{D}(\hat{\boldsymbol{\theta}}) = \sigma^2 (\boldsymbol{X^T X})^{-1}.$$

The predicted response at $\boldsymbol{x}$ is

$$\hat{\boldsymbol{Y}}(\boldsymbol{x}) = \boldsymbol{f_1}(\boldsymbol{x})\hat{\boldsymbol{\theta}}_1 + \boldsymbol{f_2}(\boldsymbol{x})\hat{\boldsymbol{\theta}}_2 + ... + \boldsymbol{f_k}(\boldsymbol{x})\hat{\boldsymbol{\theta}}_k = \boldsymbol{f}^T(\boldsymbol{x})\hat{\boldsymbol{\theta}}.$$

From the above equation, the dispersion matrix $\boldsymbol{\theta}$ depends proportionately on σ^2, not on $\boldsymbol{\theta}$. Therefore, to obtain a better inference for $\boldsymbol{\theta}$, we need more information in an experiment. That is, we need to have a maximum information matrix. Then the goal is to maximize the information matrix $\boldsymbol{X^T X}$ or equivalently we can minimize the inverse of the information matrix.

2.1.1 Exact and Approximate Designs

The linear model in (2.2) can be expressed as:

$$E(y|\boldsymbol{v}, \boldsymbol{\theta}, \sigma) = \boldsymbol{v}^T\boldsymbol{\theta} \tag{2.6}$$

where

$$\boldsymbol{v} = (\boldsymbol{f_1}(\boldsymbol{x}), \boldsymbol{f_2}(\boldsymbol{x}), ...\boldsymbol{f_k}(\boldsymbol{x}))^T, \boldsymbol{v} \in \boldsymbol{\mathcal{V}};$$

$$\boldsymbol{\mathcal{V}} = \{\boldsymbol{v} \in R^k : \boldsymbol{v} = (\boldsymbol{f_1}(\boldsymbol{x}), \boldsymbol{f_2}(\boldsymbol{x}), ...\boldsymbol{f_k}(\boldsymbol{x}))^T, \boldsymbol{x} \in \chi\}$$

From the above, it is apparent that the choice of a vector $\boldsymbol{x}$ in the design space χ is equivalent to the choice of a k-vector $\boldsymbol{v}$ in the closed bounded k-dimensional space $\boldsymbol{\mathcal{V}} = \boldsymbol{f}(\boldsymbol{x})$, where $\boldsymbol{\mathcal{V}}$ is the vector function $(\boldsymbol{f_1}, \boldsymbol{f_2}, ..., \boldsymbol{f_k})^T$ and is the image under $\boldsymbol{f}$ of $\boldsymbol{\chi}$. That is $\boldsymbol{\mathcal{V}}$ is the induced design space. This design space is normally continuous, but it can be assumed to be discrete. It can be justified using Caratheodory's theorem.

For assumed discrete design space $\boldsymbol{\mathcal{V}}$ consisting of J distinct vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_J$, we must first choose a value of $\boldsymbol{v}$ from the J elements of $\boldsymbol{\mathcal{V}}$ to obtain observation on y.

After discretization, the design problem becomes more concise. Then we need to consider what points $\boldsymbol{v}_j$ should be taken, and the number of observations, that is, n_j at these points that should be taken to determine the best estimator of $\boldsymbol{\theta}$. Given n observations in total, we have now to choose how many observations of n_j we must take at $\boldsymbol{v_j}$, subject to $\sum_{j=1}^{J} n_j = n$. Therefore, we can express the information matrix in the form:

$$\boldsymbol{X^T X} = \boldsymbol{M(n)} = \sum_{j=1}^{J} n_j \boldsymbol{v}_j \boldsymbol{v}_j^T = \boldsymbol{V N V}^T \tag{2.7}$$

where $\boldsymbol{n} = (n_1, n_2, .., n_J)^T, \boldsymbol{N} = diag(n_1, n_2, .., n_J)$.

Suppose we would like to determine the values of the vector $\boldsymbol{n}$ to make the matrix $\boldsymbol{M(n)}$ as large as possible. It is an integer programming problem to find the exact integer values n_j, and so it becomes an exact design problem. However, the use of calculus theory cannot be accustomed to find the optimal solutions. Even without additional constraints, it is still hard to solve integer programming problems. In order to solve this problem, we can find a proportion p_j as defined by $p_j = n_j/n$. That is, $p_j \geq 0$ and $\sum_j^J p_j = 1$. So (2.7) can be expressed as:

$$\boldsymbol{M(n)} = \sum_{j}^{J} n_j \boldsymbol{v}_j \boldsymbol{v}_j^T = n \sum_{j}^{J} p_j \boldsymbol{v}_j \boldsymbol{v}_j^T = \boldsymbol{n M(p)} \tag{2.8}$$

Thus, our problem is selecting design point p to make information $\boldsymbol{M(p)}$ as large

as possible, subject to $p_j = n_j/n$. Now, it has become an approximate design problem. The corresponding weight is continuous for approximate design, and the sum of corresponding weights is 1. Whereas for exact design, the weight is discrete, and the sum of corresponding numbers of subjects of an exact design is n. Approximate design is a preferred choice over the original design and is given by np^*, rounded to the nearest exact design.

Let us look at the expression of information matrix closely. We notice the expectation of $\boldsymbol{vv}^T$ by viewing $\boldsymbol{p}$ as probability distribution on $\boldsymbol{\mathcal{V}}$ that is $\boldsymbol{M(p)} = E_p[\boldsymbol{vv^T}]$, where $\boldsymbol{P}(\boldsymbol{v} = \boldsymbol{v}_j) = p_j$. The design can be thought of by a set of probabilities or weights $p_j,\ p_j$ being assigned to $\boldsymbol{v}_j$. Note the points that are not support points of the design are assigned to zero weights.

2.1.2 Design measure

The form of a design measure can write as:

$$\xi = \begin{pmatrix} x_1 & x_2 & ... & x_J \\ p_1 & p_2 & ... & p_J \end{pmatrix} \tag{2.9}$$

where the first line is the location of the design points x_j, the second line is the design weights $p_j^{'}$s. $x_j \in \boldsymbol{\chi}$ and $\sum_{j=1}^{J} p_j = 1,\ 0 \leq p_j \leq 1\ for\ all\ j$. We also can use notation p instead of ξ.

2.1.3 Support of a Design Measure

The support of design measure ξ in the design space $\mathcal{V}$ is defined as a set of vectors $\boldsymbol{v_j}$ that we consider to have nonzero weights under p. It is given by:

$$Supp(\xi) = \{\boldsymbol{v}_j \in \boldsymbol{\mathcal{V}} : p_j > 0, j = 1, 2, ...J\}. \tag{2.10}$$

Assuming p^* is an optimal design, the support of the design measure under design p^* may be a strict subset of $\boldsymbol{\mathcal{V}}$.

2.1.4 Standardized Variance of the Predicted Response

The standardized variance of the predicted response on y at $\boldsymbol{x}$ for the design (2.9) represented as:

$$d(\boldsymbol{x}, \boldsymbol{p}) = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{M}^{-1}(\boldsymbol{p})\boldsymbol{f}(\boldsymbol{x}) \tag{2.11}$$

where $\boldsymbol{M^{-1}(p)}$ is the inverse of the information matrix, that is, the covariance matrix.

2.1.5 The information matrix $\boldsymbol{M(p)}$

The information matrix $\boldsymbol{M(p)}$ can be written as:

$$\boldsymbol{M(p)} = \sum_{j=1}^{J} p_j \boldsymbol{v}_j \boldsymbol{v}^T = \boldsymbol{V}\boldsymbol{P}\boldsymbol{V}^T. \tag{2.12}$$

where $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_J]$ and $\boldsymbol{P} = \text{diag}(p_1, p_2, \ldots, p_J)$.

There are two essential properties of an information matrix that are symmetric

and non-negative definite. The symmetry property follows from its definition. The non-negativeness property can be easily verified as:

$$\boldsymbol{x}^T \boldsymbol{M}(\boldsymbol{p})\boldsymbol{x} = \boldsymbol{x}^T E_p[\boldsymbol{v}\boldsymbol{v}^T]\boldsymbol{x} = E_p[\boldsymbol{x}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{x}] = E_p[(\boldsymbol{x}^T\boldsymbol{v})^2] \geq 0.$$

The information matrix is used widely in optimal experimental design. The inverse of the variance-covariance matrix (dispersion matrix) is the information matrix. Minimizing the variance corresponds to maximizing the information.

2.2 Criteria in Optimal Design and Properties

2.2.1 D-optimality

D-optimality is the most popular and important design criterion in the literature. The criterion function is defined as:

$$\phi_D(\boldsymbol{p}) = \psi_D\{\boldsymbol{M}(\boldsymbol{p})\} = logdet\{\boldsymbol{M}(\boldsymbol{p})\} = -logdet\{\boldsymbol{M^{-1}}(\boldsymbol{p})\} \tag{2.13}$$

In D-optimality, we maximize the determinant of information matrix $\boldsymbol{M}(\boldsymbol{p})$, or its logarithm $\boldsymbol{logdet\{M(p)\}}$. Because of the reciprocity property of the covariance matrix and information matrix, maximizing the determinant of the information matrix is equivalent to minimizing the determinant of the covariance matrix. Hence, we minimize the generalized variance of the parameter estimates in D-optimality.

D-optimality has some useful properties. It has a connection between D-optimality and the standardized variance of the predicted response.

Suppose we have a given model with design variable $\boldsymbol{x}$ and have a D-optimal design, that is p^*, then

$$sup_x d(\boldsymbol{x}, \boldsymbol{p}^*) = k \tag{2.14}$$

where $\boldsymbol{d(x, p)}$ is the standardized variance of the predicted response as defined in (2.11), and k is the number of parameters in the model (Kiefer and Wolfowitz, 1960).

In the linear model, assume the normality of the errors. The volume of the confidence ellipsoid for the parameters is proportional to the D-optimal criterion. The joint confidence region for the vectors of unknown parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is given by:

$$\{\boldsymbol{\theta} : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{M(p)} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq c)\}, \text{ for some critical value } c \tag{2.15}$$

where $\hat{\boldsymbol{\theta}}$ is the least squares estimate or the maximum likelihood estimate of $\boldsymbol{\theta}$. Generally speaking, the smaller the volume of the confidence interval the more accurate estimators. In other words, D-optimality is equivalent to minimizing the volume of the ellipsoid, where the volume is proportional to $\boldsymbol{[det(M(p)^{-1}]^{1/2}}$.

The D-optimality can be also explained in terms of the eigenvalues of the information matrix $\boldsymbol{M(p)}$. Let us denote $(\lambda_1, \lambda_2, .., \lambda_k)$ are the eigenvalues of $\boldsymbol{M(p)}$. The eigenvalues of inverse of $\boldsymbol{M(p)}$ that is $\boldsymbol{M(p)}^{-1}$ would be $(1/\lambda_1, 1/\lambda_2, .., 1/\lambda_k)$. The half length of the axes of the confidence ellipsoid is the form of $c\sqrt{(1/\lambda_k)}$. So, the eigenvalues of $\boldsymbol{M(p)}^{-1}$ are proportional to the squared of the lengths of the axes of the confidence ellipsoid. In other words, the D-optimal design can be obtained by minimizing the product of the eigenvalues of $\boldsymbol{M(p)}^{-1}$, i.e. $\prod_{\boldsymbol{i=1}}^{\boldsymbol{k}} \frac{1}{\lambda_i}$.

The D-optimality criterion is a concave function of the positive definite symmetric matrices. The criterion function $\boldsymbol{\phi_D}$ is differentiable whenever it is finite, and the first partial derivatives are given by:

$$\frac{\partial \phi_D}{\partial p_j} = \boldsymbol{v}_j^T \boldsymbol{M^{-1}}(\boldsymbol{p}) \boldsymbol{v}_j. \tag{2.16}$$

Moreover, another property of the D-optimality criterion is invariant under a non-singular linear transformation of the design place. The most extensive references of this area can be found in Kiefer and Wolfowitz (1959), Fedorov (2013), Silvey (2013), Berger and Wong (2009), Atkinson et al. (2007), Mandal and Torsney (2006), Mandal et al. (2005).

2.2.2 D_A-optimality

D_A-optimality is only interested in some of the unknown parameters or some linear combinations of the parameters of the linear model. This criterion has very similar properties to the D-optimality criterion. Sibson (1972) called this as D_A-optimality in order to emphasize the dependence of the design on the matrix of coefficients A.

The criterion function of D_A-optimality is given by:

$$\phi_{DA}(\boldsymbol{p}) = \psi_{DA}\{\boldsymbol{M}(\boldsymbol{p})\} = -logdet\{\boldsymbol{A}\boldsymbol{M}^{-1}(\boldsymbol{p})\boldsymbol{A}^T\} \tag{2.17}$$

Suppose s linear combinations of the parameters $\theta_1, \theta_2, .., \theta_k$ we are interested in. The elements of s linear combinations of the vector $\boldsymbol{\alpha} = \boldsymbol{A\theta}$, where $\boldsymbol{A}$ is $s \times k$ matrix

of rank $s \leq k$. If we assume $\boldsymbol{M(p)}$ is non-singular, then the covariance matrix of the least squares estimator of $\boldsymbol{A\theta}$ is proportional to the matrix $\boldsymbol{AM^{-1}(p)A^T}$.

The partial derivatives of D_A-optimality criterion are given by:

$$\frac{\partial \phi_{DA}}{\partial p_j} = \boldsymbol{v_j^T M^{-1}(p) A^T [AM^{-1}(p) A^T]^{-1} AM^{-1}(p) v_j.} \tag{2.18}$$

2.2.3 D_s-optimality

D_s-optimality is an important special case of D_A-optimality. In D_s-optimality, we are interested in a subset of s parameters. In D_A-optimality, if $\boldsymbol{A} = [\boldsymbol{I_s} : \boldsymbol{O}]$, where $\boldsymbol{I_s}$ is the identity matrix with dimension $s \times s$ and $\boldsymbol{O}$ is zero matrix with $s \times (k-s)$ dimension. If we are only interested in estimating the first s parameters. Then the criterion becomes a D_s-optimality criterion. In this case, without loss of generality, considering the first s parameters θ_1 to θ_s in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the information matrix of $\boldsymbol{M(p)}$ can be written as:

$$\boldsymbol{M(p)} = \begin{bmatrix} \boldsymbol{M_{11}}^{s \times s} & \boldsymbol{M_{12}}^{s \times (k-s)} \\ \boldsymbol{M_{12}}^{T} & \boldsymbol{M_{22}}^{(k-s) \times (k-s)} \end{bmatrix} \tag{2.19}$$

From Rohde (1965), the inverse of $\boldsymbol{AM^{-1}(p)A^T}$ can be expressed as:

$$(\boldsymbol{AM^{-1}(p)A}^T)^{-1} = \boldsymbol{M_{11} - M_{12} M_{22}^{-1} M_{12}^T}. \tag{2.20}$$

Therefore, D_s design criterion becomes that selecting the design p to maximize the determinant of $(\boldsymbol{AM^{-1}(p)A}^T)^{-1} = \boldsymbol{M_{11} - M_{12} M_{22}^{-1} M_{12}^T}$. It is easy to see, that maximizing ϕ_{D_A} is equivalent to maximizing the ϕ_{Ds} criterion:

$$\phi_{D_s}(\boldsymbol{p}) = logdet\{\boldsymbol{M_{11}} - \boldsymbol{M_{12}}\boldsymbol{M_{22}^{-1}}\boldsymbol{M_{12}^{T}}\} \tag{2.21}$$

For more details we refer to Karlin and Studden (1966), Atwood (1969), and Silvey (2013).

2.2.4 A-optimality

A-optimality is defined by maximizing criterion function:

$$\phi_A(\boldsymbol{p}) = \psi_A\{\boldsymbol{M(p)}\} = -trace\{\boldsymbol{M}^{-1}(\boldsymbol{p})\} \tag{2.22}$$

Because of the property of the covariance matrix and the information matrix, the A-optimal design minimizes the sum of the variances of all the parameter estimates or the average variance but ignores the correlation structures of these estimators.

The partial derivatives of the criterion are given by:

$$\frac{\partial \phi_A}{\partial p_j} = \boldsymbol{v_j}^T \boldsymbol{M(p)}^{-2} \boldsymbol{v_j} \tag{2.23}$$

Unlike D-optimality, the A-optimality criterion may not be invariant under the linear transformation of the scale of the independent variables. Compared with other criteria from the computational aspect, it only requires the addition of the k diagonal elements of the $\boldsymbol{M(p)^{-1}}$. The more discussion can be found in Elfving (1952), Atkinson et al. (2007) and Berger and Wong (2009).

2.2.5 G-optimality

In G-optimality (Global optimality), we maximize the criterion function given by:

$$\phi_G(\boldsymbol{p}) = \psi_G\{\boldsymbol{M}(\boldsymbol{p})\} = -Max_{v\in\mathcal{V}}\boldsymbol{v}^T\boldsymbol{M}^{-1}(\boldsymbol{p})\boldsymbol{v}. \tag{2.24}$$

This criterion seeks to minimize the maximum value of standadrdized variance of predicted response, $\boldsymbol{v}^T\boldsymbol{M}(\boldsymbol{p})^{-1}\boldsymbol{v}$, which is proportional to the variance of $\boldsymbol{v}^T\hat{\boldsymbol{\theta}}$. Kiefer and Wolfowitz (1960) prove that this criterion and D-optimal criterion are equivalent.

To understand the connection between G-optimality and D-optimality. We say that under the optimal design p^*, the standardized variance for a G-optimal design is less than or equal to k, where k is the number of parameters in the model. That is, $d(\boldsymbol{x}, p^*) \leq k$ with equality at the support points. Thus, we can use this inequality to check whether a design is D-optimal or not. G-optimality is also invariant under a non-singular linear transformation of the scale of the independent variables. Suppose uniquely $\boldsymbol{v_j}^T\boldsymbol{M}(\boldsymbol{p})^{-1}\boldsymbol{v_j} = Max_t\boldsymbol{v_t}^T\boldsymbol{M}(\boldsymbol{p})^{-1}\boldsymbol{v_t}$, then the partial derivative of ϕ_G can write as:

$$\frac{\partial\phi_G}{\partial p_j} = [\boldsymbol{v_j}^T\boldsymbol{M}^{-1}(\boldsymbol{p})\boldsymbol{v_j}]^2. \tag{2.25}$$

2.2.6 Linear Optimality

A linear or L-optimal design criterion function is given by:

$$\phi_L(\boldsymbol{p}) = \psi_L\{\boldsymbol{M}(\boldsymbol{p})\} = -tr\{\boldsymbol{M}^{-1}(\boldsymbol{p})\boldsymbol{L}\} \tag{2.26}$$

It is linear in the elements of the covariance matrix $\boldsymbol{M(p)^{-1}}$, where $\boldsymbol{L}$ is $k \times k$ matrix of coefficients. $\boldsymbol{L}$ can be writen as $\boldsymbol{L = A^T A}$ where $\boldsymbol{A}$ is a matrix of dimension $s \times k$ with rank s. There is a relationship between L-optimal design and D_A-optimal design:

$$\begin{aligned} \phi_L(p) = \psi_L\{\boldsymbol{M(p)}\} &= -tr\{\boldsymbol{M^{-1}(p)L}\} \\ &= -tr\{\boldsymbol{M}^{-1}(p)\boldsymbol{A}^T\boldsymbol{A}\} \\ &= -tr\{\boldsymbol{AM}^{-1}(\boldsymbol{p})\boldsymbol{A}^T\} \end{aligned} \tag{2.27}$$

We minimize the trace of the matrix $\boldsymbol{AM^{-1}(p)A}$ in L-optimal design while we minimize the log determinant of the matrix $\boldsymbol{AM^{-1}(p)A}$ in D_A-optimal design. If this matrix of coefficients L is an identity matrix $\boldsymbol{I}$, then this criterion is simply an A-optimality criterion.

The first partial derivatives of ϕ_L are given by:

$$\frac{\partial \phi_L}{\partial p_j} = \boldsymbol{v_j}^T\boldsymbol{M^{-1}(p)A}^T\boldsymbol{AM^{-1}(p)v_j} \tag{2.28}$$

2.3 A Class of Algorithms in Optimal Design

In the previous section, we discussed different optimality criterion function $\phi(p)$ which is a function of the information matrix $\boldsymbol{M(p)}$. Recall an approximate design

in one design variable:

$$\xi = \begin{Bmatrix} x_1 & x_2 & \cdots & x_J \\ p_1 & p_2 & \cdots & p_J \end{Bmatrix} \tag{2.29}$$

The first line are the locations of the design points, the second line are associated design weights p_j satisfying the constraints $p_j \geq 0,\ \sum_{j=1}^{J} p_j = 1$. Generally, our problem is to maximize the criterion function $\phi(p)$ subject to satisfying the constraints $p_j \geq 0, \sum_{j=1}^{J} p_j = 1$.

In order to find an optimal design, we use the directional derivative $F_\phi\{\boldsymbol{p}, \boldsymbol{q}\}$ tool to determine optimality conditions. The multiplicative algorithm is used to construct the optimal design. We will now review the directional derivations, multiplicative algorithms and the required function $f(.)$.

2.3.1 Directional Derivatives

The directional derivatives $F_\phi\{\boldsymbol{p}, \boldsymbol{q}\}$ of a criterion function $\phi(.)$ at p in the direction of q is defined as:

$$F_\phi\{\boldsymbol{p}, \boldsymbol{q}\} = \lim_{\varepsilon \downarrow 0} \frac{\phi\{(1-\varepsilon)\boldsymbol{p} + \varepsilon \boldsymbol{q}\} - \phi(\boldsymbol{p})}{\varepsilon} \tag{2.30}$$

Denoting j^{th} unit vector in $\boldsymbol{R^j}$ by $\boldsymbol{e_j}$, $d_j = \frac{\partial \phi}{\partial \boldsymbol{p}}$, the partial derivative of ϕ respect to p_j, we can simplify the directional derivatives

$$
\begin{aligned}
F_j = F_\phi\{\boldsymbol{p}, \boldsymbol{e_j}\} &= \sum_{i=1}^{J}(q_i - p_i)d_i \\
&= \frac{\partial \phi}{\partial p_j} - \sum_{i=1}^{J} p_i \frac{\partial \phi}{\partial p_j} \\
&= d_j - \sum_{i=1}^{J} p_i d_i
\end{aligned}
\tag{2.31}
$$

This derivative exists even if $\phi(.)$ is not differentiable. F_j is called the vertex directional derivative of $\boldsymbol{\phi(.)}$ at p.

2.3.2 Optimality Conditions

If $\phi(\boldsymbol{p})$ is differentiable at p^*, then the first order conditions for $\phi(p^*)$ to be a local maximum of $\phi(\boldsymbol{p})$ in the feasible region of the general problem are

$$
F_j^* = F_\phi\{\boldsymbol{p}, \boldsymbol{e_j}\} \begin{cases} = 0 & \text{if } p_j^* > 0 \\ \leq 0 & \text{if } p_j^* = 0 \end{cases}
\tag{2.32}
$$

Based on the General Equivalence Theorem, if $\phi(\boldsymbol{p})$ is concave on its feasible region, then the first-order stationarity conditions are both necessary and sufficient for optimality (Kiefer, 1974).

2.3.3 Multiplicative Algorithm

Because optimal solution typically is not always obtained analytically, it often requires algorithms to solve. A class of algorithms with neatly satisfying the basic

constraints of the design is the multiplicative algorithm:

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(d_j^{(r)})}{\sum_{i=1}^{J} p_j^r f(d_i^{(r)})} \tag{2.33}$$

where $d_j^{(r)} = \frac{\partial \phi}{\partial p_j}$ at r^{th} iterate $p = p^{(r)}$, and the function $f(.)$ satisfies following conditions and depend on a free positive parameter δ. This type of iteration by taking $f(d) = d^{\delta}$ was first proposed by Torsney (1977). The choice of the free positive parameter δ can affect the convergence rates. The more references for choosing the function $f(.)$ and the free parameter δ can be found at Torsney and Alahmadi (1992), Torsney and Mandal (2001), Mandal and Torsney (2006) and Mandal et al. (2017).

2.4 Matching Adjusted Indirect Comparison (MAIC)

As discussed in the Chapter 1, matching adjusted indirect comparison is a re-weighting method where the weights are calculated by deriving propensity scores in the individual patient data (IPD) and matching the summaries reported (typically mean and variance) in the Aggregate data (AgD). See figure 2.1.

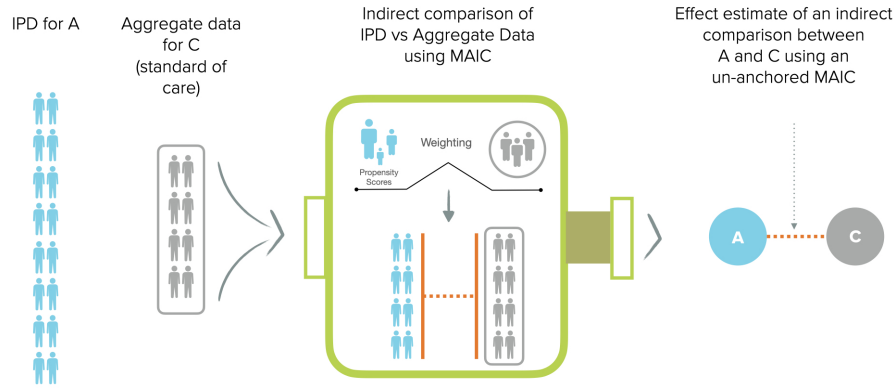
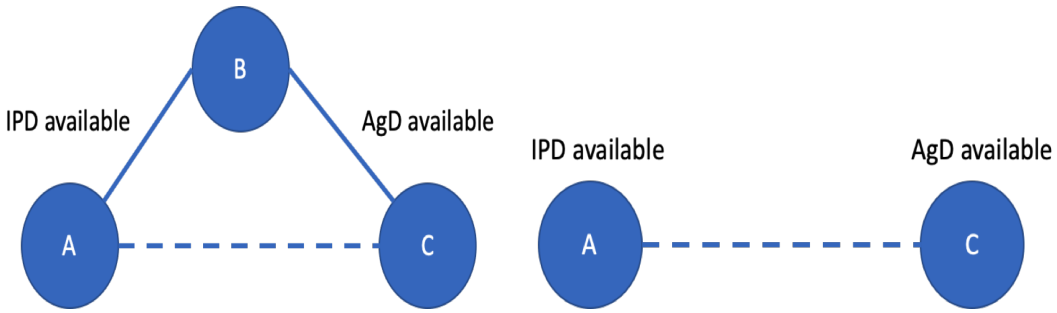


Figure 2.1: Idea of MAIC

There are two types of MAIC, if IPD and AgD are linked by a common treatment, it is called anchored MAIC, otherwise, it is called unanchored MAIC. This is depicted in the image below:

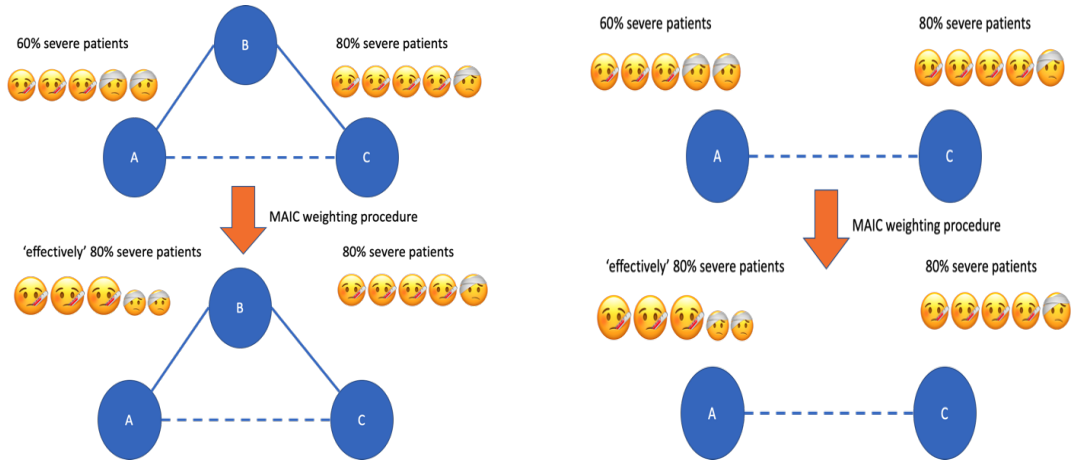


(a) Anchord MAIC (b) Unanchored MAIC

Figure 2.2: Two case of MAIC

In general, to explain the MAIC procedure, we need to make sure the studies are satisfied by an important assumption. The important assumption of MAIC is that the IPD and AgD are from a common underlying patient population. Let us denote A, B, C by three treatments. We will give an example of an anchored MAIC here.

We do not consider the Unanchored cases here. It can be shown below:



(a) Anchored MAIC (b) Unanchored MAIC

Figure 2.3: MAIC weighting procedure

Let us describe the situation in figure 2.3a. Treatment A and treatment C have common treatment B. The IPD is only available from AB trials. For BC trials we have AgD. The emoji face with a head bandage is a severe patient, face with a thermometer is a non-severe patient. In AB trials, 60 percent of patients are served patients compared to 80 percent of served patients in BC trials. There may be severe patient effects of different treatments. Without any adjustment comparison of AB, trials will cause bias. By using MAIC, re-weight the percentage of patients in AB trials, increase the contribution of the effect sizes of severe patients and decrease the contribution of the non-severe patients that matches the actual percentage of served patients in the indirect comparison AC trials.

The main idea behind MAIC is to derive propensity scores in the IPD group AB trials and use the propensity scores to calculate weights for each patient to balance prognostic variables and effect modifiers between AC trials. This reweighting

approach was originally published by Signorovitch et al. (2010).

2.4.1 MAIC Methodology

After a general explanation, we now give a theoretical example. Let us suppose that IPD is available for treatment t=0 for our own trial. But only AgD is available for treatment t=1, which is the competitor's trial. The purpose of MAIC here is to reweight IPD to match AgD characteristics for treatment t=1. Suppose that the baseline characteristics are age which is an effect modifier and gender, which is a prognostic variable. Individual patient i is assigned to treatment t and has baseline characteristics X_{it}, X_{i1} is baseline characteristics for treatment t=1 (AgD), which is for AgD. X_{i0} is individual patient data for treatment t=0 (IPD). The weights are derived from a propensity score equation that will be applied to the patients. The weights are defined by

$$w_{it} = \frac{P(t_i = 1|x_{it})}{P(t_i = 0|x_{it})} \tag{2.34}$$

The probabilities indicated here refer to the probability of the patient assigned treatment t=1 vs treatment t=0, and w_{it} is the odd ratio that patient i receives treatment t=1 to t=0. In other words, re-weight the patients assigned treatment 0 to match the distribution of patients assigned treatment 1. In order to find the re-weight weight, the weights are calculated by estimating the logistics propensity score model:

$$log(w_{it}) = \alpha + x'_{it}\beta \tag{2.35}$$

We also can re-write as:

$$w_{it} = e^{\alpha + x'_{it}\beta} = e^{\alpha} e^{x'_{it}\beta}$$

where α and β are the parameters.

In order to estimate w_{it}, we need to estimate the parameters first. The weight for subject i is: $w_{it} = exp(\alpha + x'_{it}\beta)$.

The maximum likelihood estimation(MLE) is normally used to estimate the parameters for the logistic regression model. However, that is not possible here, because there is no individual patient data (IPD) for t=1; we only have aggregate data (AgD). In this case, to estimate β, we use the method of moment, which is a key feature for the MAIC method (Signorovitch et al., 2010). A detailed description can be found in Appendix B of the Supplemental Digital Content of the reference. We want the mean of the weighted IPD is equal to the mean of the baseline characteristics for the AgD. The mean of the weighted IPD can be represented as when $t = 0$:

$$\frac{\sum_{i:t=0} x_{it}\hat{w_{it}}}{\sum_{i:t=0} \hat{w}_{it}} = \frac{\sum_{i:t=0} x_{it} e^{\alpha} e^{x'_{it}\hat{\beta}}}{\sum_{i:t=0} e^{\alpha} e^{x'_{it}\hat{\beta}}} = \frac{\sum_{i:t=0} x_{it} e^{x'_{it}\hat{\beta}}}{\sum_{i:t=0} e^{x'_{it}\hat{\beta}}} \tag{2.36}$$

Thus we see that we only need to estimate the β. Now β can be estimated by solving the equation:

$$\frac{\sum_{i:t=0} x_{it} exp(x'_{it}\hat{\beta})}{\sum_{i:t=0} exp(x'_{it}\hat{\beta})} - \bar{x}_1 = 0$$

The mean of the weighted single-arm IPD will be matched to the mean of the given AgD. The mean of the weighted IPD is set to equal $\bar{x_1}$.

$$\sum_{i:t=0}(x_{it} - \bar{x}_1))exp(x'_{it}\hat{\beta}) = 0$$

Without loss generality, assume that $\bar{x_1} = 0$, so we get

$$\sum_{i:t=0} x_{it}exp(x'_{it}\beta) = 0$$

We see that this is a derivative of an objective function

$$Q(\beta) = \sum_{i:t=0} exp(x'_{it}\beta).$$

The objective function is convex, therefore, the minimum of the objective function has a unique solution. We will get an estimate $\hat{\beta}$ by minimizing the objective function $Q(\beta)$. Note that $\hat{\beta}$ can be obtained by minimizing $Q(\beta)$ using Newton-Raphson method. The BFGS algorithm is used in a Technical Support Document (TSD) produced for the National Institute for Health and Care Excellence (NICE) (Phillippo et al., 2018).

We can now calculate the weight ($\hat{w}_{it}$) of each individual that is related to design weight p. That is, the mean of the weighted IPD is equal to the mean of the baseline characteristics for the AgD. Then the MAIC weight will be calculated as:

$$\frac{\hat{w}_{it}}{\sum_i \hat{w}_{it}}. \tag{2.37}$$

Chapter 3

Construction of the Optimal Designs subject to Matching Pre-defined Baseline Characteristics

3.1 Formulation of the Constrained Optimization Problem

In this chapter, we develop the methodologies and provide a step by step approach on constructing optimal designs subject to matching pre-defined baseline characteristics in MAIC by using a Lagrangian method.

We consider D-optimality as our main criterion function. We consider age and

gender in the baseline characteristics. As discussed earlier, the available individual patient data are IPD, and the aggregated data that needs to be matched are AgD.

We start with the linear model as discussed in Chapter 2. We consider the design variables $\boldsymbol{x}$ as the baseline characteristics variables in MAIC. The model can be written as

$$E(y|\boldsymbol{v}, \boldsymbol{\theta}, \sigma) = \boldsymbol{v}^T\boldsymbol{\theta} \tag{3.1}$$

where

$$\boldsymbol{v} = (\boldsymbol{f_1}(\boldsymbol{x}), \boldsymbol{f_2}(\boldsymbol{x}), ...\boldsymbol{f_k}(\boldsymbol{x}))^T, \boldsymbol{v} \in \boldsymbol{\mathcal{V}};$$

$$\boldsymbol{\mathcal{V}} = \{\boldsymbol{v} \in R^k : \boldsymbol{v} = (\boldsymbol{f_1}(\boldsymbol{x}), \boldsymbol{f_2}(\boldsymbol{x}), ...\boldsymbol{f_k}(\boldsymbol{x}))^T, \boldsymbol{x} \in \chi\}$$

In this context, the approximate design is characterized by a probability measure p, the design space must be discretized. p is characterized by a set of weights $p_1, p_2, ..., p_J$ satisfying $p_J \geq 0, j = 1, 2, ..., J$ and $\sum p_j = 1$ with weights p_j assigned to $\boldsymbol{v}_j$. The design is given by

$$\xi = \begin{Bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_J \\ p_1 & p_2 & \cdots & p_J \end{Bmatrix} \tag{3.2}$$

The objective is to construct optimal design subject to matching pre-defined baseline characteristics which are denoted by $\boldsymbol{\mu} = \{\mu_1, ..., \mu_J\}^T$. As mentioned above, we take D-optimality ($\phi(\boldsymbol{p})$, the log-determinant of the information matrix) as our main

criterion function. Thus, we wish to maximize

$$\phi(\boldsymbol{p}) \text{ subject to } \sum_{j=1}^{J} p_j \boldsymbol{x}_j = \boldsymbol{\mu} \tag{3.3}$$

with $p_j \geq 0$ and $\sum p_j = 1$.

Here, the baseline characteristics constraints $\sum_{j=1}^{J} p_j \boldsymbol{x}_j = \boldsymbol{\mu}$ can be written as $\boldsymbol{Vp} = \boldsymbol{\mu}$, where $\boldsymbol{V}$ consists of the rows the values of the baseline characteristics variable $\boldsymbol{x}_i$, and $\boldsymbol{p}$ is the vector $(p_1, p_2, \cdots, p_J)^T$. An example with two variable model is given in Section (3.2).

As we have $\boldsymbol{Vp} = \boldsymbol{\mu}$, we can write

$$\boldsymbol{(Vp - \mu)^T (Vp - \mu)} = 0$$

Let

$$D(\boldsymbol{p}) = \boldsymbol{(Vp - \mu)}^T \boldsymbol{(Vp - \mu)} = \boldsymbol{p}^T \boldsymbol{V}^T \boldsymbol{Vp} - 2(\boldsymbol{V}^T \boldsymbol{\mu})^T \boldsymbol{p} + \boldsymbol{\mu}^T \boldsymbol{\mu}$$

So, the baseline characteristics constraints are written as $D(\boldsymbol{p}) = 0$. In general, $D(\boldsymbol{p})$ could be equal to a constant, say c. In our case, $c = 0$. So, in general, the constraints can be $D(\boldsymbol{p}) = c$. Thus, our problem becomes:

Maximize

$$\begin{gathered} \phi(\boldsymbol{p}) \\ \text{subject to } D(\boldsymbol{p}) = c \end{gathered} \tag{3.4}$$

with $p_j \geq 0$ and $\sum p_j = 1$.

We now formulate the Lagrangian function as given by

$$L(\phi, \boldsymbol{p}, \lambda, \gamma) = \phi(\boldsymbol{p}) + \lambda[D(\boldsymbol{p}) - c] + \gamma(\sum p_j - 1) \tag{3.5}$$

where $\phi(\boldsymbol{p}) = logdet(\boldsymbol{M(p)})$, constraint function is $D(\boldsymbol{p})$, the Lagrange multipliers λ and γ are the rates of change of function being maximized as a function of the constraint parameter.

We consider first partial derivatives with respect to p for $j = 1, 2, ..., J$:

$$\begin{aligned} \frac{\partial L}{\partial p_j} &= d_j^L \\ &= \frac{\partial \phi}{\partial p_j} + \lambda \frac{\partial D}{\partial p_j} + \gamma \\ &= d_j^\phi + \lambda d_j^D + \gamma \end{aligned} \tag{3.6}$$

L is the Lagrange function, so d_j^L must be zero. It implies that we must have:

$$\begin{aligned} d_j^L &= 0 \\ i.e. \sum_{j=1}^{J} p_j d_j^L &= 0 \\ i.e. \sum_{j=1}^{J} p_j (d_j^\phi + \lambda d_j^D + \gamma) &= 0 \end{aligned}$$

Thus,

$$\gamma = -\sum_{j=1}^{J} p_j (d_j^\phi + \lambda d_j^D) \tag{3.7}$$

So, the vertex directional derivatives of L are obtained as:

$$\begin{aligned} F_j^L &= d_j^L - \sum_{j=1}^{J} p_j d_j^L \\ &= d_j^\phi + \lambda d_j^D - [\sum_{j=1}^{J} p_j d_j^\phi + \sum_{j=1}^{J} p_j d_j^D] \\ &= (d_j^\phi - \sum_{j=1}^{J} p_j d_j^\phi) + \lambda [d_j^D - \sum_{j=1}^{J} p_j d_j^D] \\ &= F_j^\phi + \lambda F_j^D \\ &\equiv 0 \end{aligned} \tag{3.8}$$

where $F_j^\phi = (d_j^\phi - \sum_{j=1}^{J} p_j d_j^\phi)$, $F_j^D = d_j^D - \sum_{j=1}^{J} p_j d_j^D$.

Equivalently, we can write in matrix form:

$$\boldsymbol{F^L = F^\phi + \lambda F^D = 0} \tag{3.9}$$

Hence,

$$\boldsymbol{F^D \lambda = -F^\phi} \tag{3.10}$$

where $\boldsymbol{F^D} = [F_1^D, F_2^D, .., F_J^D]^T$ and $\boldsymbol{F^\phi} = [F_1^\phi, F_2^\phi, .., F_J^\phi]^T$.

That is, we have a system of linear equations on λ, say $\boldsymbol{A\lambda = b}$, where $\boldsymbol{A = F^D}$, $\boldsymbol{\lambda} = \lambda$, and $\boldsymbol{b} = -\boldsymbol{F}^\phi$.

The set of solutions is given by:

$$\boldsymbol{\lambda = A^- b + (I - A^- A)z} \text{ for any } \boldsymbol{z}$$

where $\boldsymbol{A^-}$ is any generalized inverse of A. If $\boldsymbol{A^T A}$ is nonsingular, then one choice is the Moore-Penrose inverse which is given by:

$$\begin{aligned} \boldsymbol{A^-} &= (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T \\ &= ((\boldsymbol{F^D})^T\boldsymbol{F^D})^{-1}(\boldsymbol{F^D})^T \end{aligned} \tag{3.11}$$

Then

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{A^-b} = -((\boldsymbol{F^D})^T\boldsymbol{F^D})^{-1}(\boldsymbol{F^D})^T\boldsymbol{F^\phi} \tag{3.12}$$

Now, our optimal design, say p^*, must satisfy:

$$\begin{aligned} \boldsymbol{F^D}\hat{\boldsymbol{\lambda}} &= -\boldsymbol{F^\phi} \\ i.e.\ \boldsymbol{F^D}\hat{\lambda} + \boldsymbol{F^\phi} &= \mathbf{0} \end{aligned} \tag{3.13}$$

Substituting the value $\hat{\boldsymbol{\lambda}}$, we obtain $\boldsymbol{u}$:

$$\begin{aligned} \boldsymbol{u} &= [(\boldsymbol{F^D})^T\boldsymbol{F^D}]\boldsymbol{F^\phi} - [((\boldsymbol{F^D})^T\boldsymbol{F^\phi}]\boldsymbol{F^D} = \mathbf{0} \\ i.e.\ \ \boldsymbol{u}^T\boldsymbol{u} &= 0 \end{aligned} \tag{3.14}$$

So the optimal design p^* must satisfy:

$$\boldsymbol{u} = (u_1, u_2, ...u_J)^T = \mathbf{0} \tag{3.15}$$

where

$$u_i = [(\boldsymbol{F^D})^T\boldsymbol{F^D}]F_i^\phi - [((\boldsymbol{F^D})^T\boldsymbol{F^\phi}]F_i^D \tag{3.16}$$

That is, the elements of $\boldsymbol{u}$ are functions of the directional derivatives of ϕ and D.

Equivalently, we can say that the optimal design p^* must satisfy $\boldsymbol{u}^T\boldsymbol{u} = 0$.

As $\boldsymbol{u}^T\boldsymbol{u} \geq 0$, p^* should minimize $\boldsymbol{u}^T\boldsymbol{u}$ or equivalently p^* should maximize the function $U = -[\boldsymbol{u}^T\boldsymbol{u}]$ with a maximum value of zero, because $-[\boldsymbol{u}^T\boldsymbol{u}] \leq 0$.

After simplification, we can write down U as:

$$U = -[(\boldsymbol{F^D})^T\boldsymbol{F^D}]^2[(\boldsymbol{F^\phi})^T\boldsymbol{F^\phi}][1 - R] \tag{3.17}$$

where

$$\begin{aligned} R &= \frac{[(\boldsymbol{F^D})^T\boldsymbol{F^\phi}]^2}{[(\boldsymbol{F^\phi})^T\boldsymbol{F^\phi}][(\boldsymbol{F^D})^T\boldsymbol{F^D}]} \\ &= \frac{\zeta_1^2}{\zeta_2\zeta_3} \end{aligned} \tag{3.18}$$

and

$$\begin{aligned} \zeta_1 &= (\boldsymbol{F^D})^T\boldsymbol{F^\phi} \\ \zeta_2 &= (\boldsymbol{F^\phi})^T\boldsymbol{F^\phi} \\ \zeta_3 &= (\boldsymbol{F^D})^T\boldsymbol{F^D} \end{aligned} \tag{3.19}$$

Substituting the R above, we obtain U:

$$\begin{aligned} U &= -[(\boldsymbol{F^D})^T\boldsymbol{F^D}]^2[(\boldsymbol{F^\phi})^T\boldsymbol{F^\phi}][1 - R] \\ &= \zeta_1^2\zeta_3 - \zeta_2\zeta_3^2 \end{aligned} \tag{3.20}$$

Ensuring the constraint can be done by transforming to a maximization problem,

we maximize

$$T = -[D(\boldsymbol{p}) - c]^2 = -[D(\boldsymbol{p})]^2 \text{ (without loss of generality we can take } c = 0) \quad (3.21)$$

Hence, our optimal design p^* should simultaneously maximize both U and T with a common maximum of zero.

Because functions U and T are negative and have a common maximum of zero. The sum of two functions $U + T$ will also have a maximum of zero at the common optimizing p^*. Also, it is equivalent to maximizing the minimum of U and T, that is, $min\{U, T\}$. We use algorithms outlined in Chapter 2. The form of the algorithm at $(r+1)^{th}$ step is

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(z_j^{(r)}, \delta),}{\sum_{i=1}^{J} p_j^{(r)} f(z_i^{(r)}, \delta)} \quad (3.22)$$

where $f(z, \delta)$ is a positive and strictly increasing function in z, and it may depend on a positive parameter δ.

The choice of the function $f(z, \delta)$ depends on what type of optimization problem is used.

In our problem, the choices of z are:

$$z_j^{(r+1)} \begin{cases} = F_j^U + F_j^T & \text{if maximize } U + T \\ = F_j^{min\{U,T\}} & \text{if maximize } \min\{U, T\} \end{cases} \quad (3.23)$$

where F_j^U, F_j^T are the directional derivatives of U, T, respectively as given by:

$$F_j^U = d_j^U - \sum_{j=1}^{J} p_j d_j^U$$
$$F_j^T = d_j^T - \sum_{j=1}^{J} p_j d_j^T \tag{3.24}$$

Here d_j^U and d_j^T are the first order partial derivative of the functions U and T respectively.

Refer to optimality conditions in chapter 2, the criterion has both positive and negative vertex directional derivatives, so the $f(.)$ needs to be defined for positive and negatives $F_j's$. The directional derivatives must satisfy the optimality conditions.

We now provide the steps for the calculations in the following.

3.1.1 Calculation Procedure

The calculation procedures are as follows:

$$d_j^\phi \quad d_{ij}^\phi \; ; \quad d_j^D \quad d_{ij}^D \; ; \quad d_j^U \quad F_j^U \; ; \quad d_j^T \quad F_j^T$$

1. Calculate Derivatives of D-optimality: $\phi(\boldsymbol{p}) = logdet\boldsymbol{M}(\boldsymbol{p})$

   (a) First derivative

   $$d_j^{\boldsymbol{\phi}} = \boldsymbol{v}_j^T \boldsymbol{M}^{-1} \boldsymbol{v}_j$$

   (b) Second derivative

   $$d_{ij}^{\boldsymbol{\phi}} = -[\boldsymbol{v}_j^T \boldsymbol{M}^{-1} \boldsymbol{v}_i]^2$$

2. Calculate Derivatives of Constraint Function: $D(\boldsymbol{p})$

   (a) First derivative

   $$d_j^D$$

   (b) Second derivative

   $$d_{ij}^D$$

   The exact expressions of d_j^D and d_{ij}^D are given in the following section.

3. Calculate Derivatives of U Function: $U = -\boldsymbol{u}^T\boldsymbol{u} = \zeta_1^2\zeta_3 - \zeta_2\zeta_3^2$

   (a) First derivative

   $$d_j^U = 2\zeta_1\zeta_3\frac{\partial\zeta_1}{\partial p_j} + \zeta_1^2\frac{\partial\zeta_3}{\partial p_j} - \zeta_3^2\frac{\partial\zeta_2}{\partial p_j} - 2\zeta_2\zeta_3\frac{\partial\zeta_3}{\partial p_j}$$

   where $\zeta_1 = (\boldsymbol{F^D})^T\boldsymbol{F^\phi}$, $\zeta_2 = (\boldsymbol{F^\phi})^T\boldsymbol{F^\phi}$, $\zeta_3 = (\boldsymbol{F^D})^T\boldsymbol{F^D}$,

   $$\boldsymbol{F^D} = \begin{bmatrix} F_1^D \\ F_2^D \\ .. \\ F_J^D \end{bmatrix},\ F_j^D = d_j^D - \sum_{i=1}^{J} p_i d_i^D$$

   $$\boldsymbol{F^\phi} = \begin{bmatrix} F_1^\phi \\ F_2^\phi \\ .. \\ F_J^\phi \end{bmatrix},\ F_j^\phi = d_j^\phi - \sum_{i=1}^{J} p_i d_i^\phi$$

$$\frac{\partial \zeta_1}{\partial p_j} = \sum_{i=1}^{J} [F_i^{\boldsymbol{\phi}} \frac{\partial F_i^D}{\partial p_i} + F_i^D \frac{\partial F_i^{\phi}}{\partial p_i}]$$

$$\frac{\partial \zeta_2}{\partial p_j} = 2 \sum_{i=1}^{J} F_i^{\boldsymbol{\phi}} \frac{\partial F_i^{\phi}}{\partial p_i}$$

$$\frac{\partial \zeta_3}{\partial p_j} = 2 \sum_{i=1}^{J} F_i^D \frac{\partial F_i^D}{\partial p_i}$$

$$\frac{\partial F_j^{\phi}}{\partial p_i} = \frac{\partial^2 \phi}{\partial p_i p_j} - [\frac{\partial \phi}{\partial p_i} + \sum_{i=1}^{J} p_t \frac{\partial^2 \phi}{\partial p_t p_i}]$$

$$\frac{\partial F_j^D}{\partial p_i} = \frac{\partial^2 D}{\partial p_i p_j} - [\frac{\partial D}{\partial p_i} + \sum_{i=1}^{J} p_t \frac{\partial^2 D}{\partial p_t p_i}]$$

(b) Directional derivative

$$F_j^U = d_j^U - \sum_{i=1}^{J} p_i d_i^U$$

4. Calculate Derivatives of T Function: $T = -[D(\boldsymbol{p})]^2$

(a) First derivative

$$d_j^T = -2[D(\boldsymbol{p}) - c] d_j^D$$

(b) Directional derivative

$$F_j^T = d_j^T - \sum p_i d_i^T$$

We now consider some models of interest. We first consider a regression model with two variables in the following section.

3.2 Two Variable Model without Interaction Term

The model is

$$\begin{aligned} E(y|x_1, x_2) &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 \\ &= \boldsymbol{v}^T \boldsymbol{\theta} \end{aligned} \tag{3.25}$$

where

$$\boldsymbol{v_x} = (1, x_1, x_2)^T, \boldsymbol{v} \in \mathcal{V};$$

$$\boldsymbol{\mathcal{V}} = \{\boldsymbol{v} \in R^k : \boldsymbol{v} = (1, x_1, x_2)^T, \boldsymbol{x} \in \chi\}.$$

For optimal design in two variables we consider the discretized design space consisting of all pairs (x_{1i}, x_{2j}) of the values for each x_1 and x_2. That is, the space will consists of $i \times j = J$ pairs of (x_1, x_2). We set up J as the number of total pairs. Variable x can be continuous variable or binary variable $(0, 1)$. Then, we can write the design with arbitrary weights as:

$$p = \begin{Bmatrix} (x_{11}, x_{21}) & (x_{12}, x_{22}) & \cdots & (x_{1J}, x_{2J}) \\ p_1 & p_2 & \cdots & p_J \end{Bmatrix} \tag{3.26}$$

After we define the model and the corresponding optimal design in two variables, we can go back to our objective problem. We wish to construct optimal design in which we maximize $\phi(\boldsymbol{p})$ subject $p_j \geq 0, \sum p_j = 1$ and p_j also match the AgD defined values $\boldsymbol{\mu}$.

3.2.1 Formulation of the Problem

Let y_i be a continuous or binary outcome for individual i. We consider x_1 as age characteristic variable, x_2 as gender characteristic variable. In AgD μ_1 is aggregate baseline characteristic age, μ_2 is the proportion of aggregate baseline characteristic male.

- x_1: Age
- x_2: Gender, male=0, female=1
- $\mu = (\mu_1, \mu_2)$
- μ_1: mean of age
- μ_2: proportion of male

The objective is to construct optimal design subject to matching pre-defined baseline characteristics, $\boldsymbol{\mu} = \{\mu_1, ..., \mu_J\}$. That is, the constraints are

$$\begin{aligned} \sum_{j=1}^{J} p_j x_{1j} &= p_1 x_{11} + p_2 x_{12} + p_3 x_{13} + ... + p_J x_{1J} = \mu_1 \\ \sum_{j=1}^{J} p_j x_{2j} &= p_1 x_{21} + p_2 x_{22} + p_3 x_{23} + ... + p_J x_{2J} = \mu_2 \end{aligned} \tag{3.27}$$

As discussed earlier in this chapter, the matching pre-defined baseline characteristics could formulate a constraint function as $D(\boldsymbol{p})$. Thus, the problem is:

Maximize

$$\begin{gathered} \phi(\boldsymbol{p}) \\ \text{subject to } D(\boldsymbol{p}) = c \end{gathered} \tag{3.28}$$

with $p_j \geq 0$ and $\sum p_j = 1$.

For the above model, the matrix $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_J]$ of (2.12) is given by

$$\boldsymbol{V} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1J} \\ x_{21} & x_{22} & \cdots & x_{2J} \end{pmatrix} \tag{3.29}$$

where $\boldsymbol{v_i} = (1, x_{1i}, x_{2i})^T$. $\boldsymbol{\mu}$ and $\boldsymbol{p}$ are given by

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ \mu_1 \\ \mu_2 \end{pmatrix} \tag{3.30}$$

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ \ldots \\ p_J \end{pmatrix}$$

.

Thus, using the above we can obtain $D(\boldsymbol{p})$:

$$D(\boldsymbol{p}) = (\boldsymbol{V}\boldsymbol{p} - \boldsymbol{\mu})^T(\boldsymbol{V}\boldsymbol{p} - \boldsymbol{\mu}) = \boldsymbol{p}^T\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{p} - 2(\boldsymbol{V}^T\boldsymbol{\mu})^T\boldsymbol{p} + \boldsymbol{\mu}^T\boldsymbol{\mu}$$

Then we can formulate the Lagrangian function:

$$L(\phi, \boldsymbol{p}, \boldsymbol{\lambda}, \gamma) = \phi(\boldsymbol{p}) + \boldsymbol{\lambda}[D(\boldsymbol{p}) - c] + \gamma(\sum p_j - 1) \tag{3.31}$$

Ensuring the constraints can be done by transforming into a maximization problem. Taking $c = 0$, we maximize

$$T = -[D(\boldsymbol{p}) - c]^2 = -[D(\boldsymbol{p})]^2 \tag{3.32}$$

At this point, we need to make sure the constraints satisfy the first order conditions with respect to the Lagrange multipliers. Therefore, optimal design p^* simultaneously maximizes both U and T with a common maximum of zero. The algorithm is

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(z_j^{(r)}, \delta)}{\sum_{i=1}^{J} p_j^{(r)} f(z_i^{(r)}, \delta)} \tag{3.33}$$

where $f(z, \delta)$ is strictly increasing and positive function in z, and depends on a free positive parameter δ.

As discussed earlier, we have two choices for the maximization problem:

$$\begin{aligned} \phi_L(\boldsymbol{p}) &= U(\boldsymbol{p}) + T(\boldsymbol{p}) \\ or, \quad \phi_L(\boldsymbol{p}) &= min\{U(\boldsymbol{p}), T(\boldsymbol{p})\} \end{aligned} \tag{3.34}$$

We need directional derivatives of both F^T and F^U:

$$F_j^U = d_j^U - \sum_{j=1}^{J} p_j d_j^U$$
$$F_j^T = d_j^T - \sum_{j=1}^{J} p_j d_j^T$$

where d_j^T and d_j^U are first partial derivatives for the two function T and U respectively.

$$T(\boldsymbol{p}) = -D(\boldsymbol{p})^2$$

$$D(\boldsymbol{p}) = \boldsymbol{p}^T \boldsymbol{V}^T \boldsymbol{V} \boldsymbol{p} - 2(\boldsymbol{V}^T \boldsymbol{\mu})^T \boldsymbol{p} + \boldsymbol{\mu}^T \boldsymbol{\mu}$$

$$d_j^D = 2\boldsymbol{V}^T \boldsymbol{V} \boldsymbol{p} - 2\boldsymbol{V}^T \boldsymbol{\mu}$$

$$\begin{aligned} d_j^T &= \frac{\partial T}{\partial p_i} \\ &= \frac{\partial(-[D(\boldsymbol{p})]^2)}{\partial p_i} \\ &= -2[D(\boldsymbol{p})] d_j^D \\ &= -2[\boldsymbol{p}^T \boldsymbol{V}^T \boldsymbol{V} \boldsymbol{p} - 2(\boldsymbol{V}^T \boldsymbol{\mu})^T \boldsymbol{p} + \boldsymbol{\mu}^T \boldsymbol{\mu}] d_j^D \\ &= -2[(\boldsymbol{V}\boldsymbol{p} - \boldsymbol{\mu})^T (\boldsymbol{V}\boldsymbol{p} - \boldsymbol{\mu})] 2\boldsymbol{V}^T \boldsymbol{V} \boldsymbol{p} - 2\boldsymbol{V}^T \end{aligned}$$

For $U = -\boldsymbol{u}^T \boldsymbol{u} = \zeta_1^2 \zeta_3 - \zeta_2 \zeta_3^2$, d_j^U is given by:

$$\begin{aligned} d_j^U &= \frac{\partial U}{\partial p_i} \\ &= 2\zeta_1 \zeta_3 \frac{\partial \zeta_1}{\partial p_j} + \zeta_1^2 \frac{\partial \zeta_3}{\partial p_j} - \zeta_3^2 \frac{\partial \zeta_2}{\partial p_j} - 2\zeta_2 \zeta_3 \frac{\partial \zeta_3}{\partial p_j} \end{aligned}$$

ζ_i are given by $\zeta_1 = (\boldsymbol{F^D})^T \boldsymbol{F}^\phi$, $\zeta_2 = (\boldsymbol{F}^\phi)^T \boldsymbol{F}^\phi$, $\zeta_3 = (\boldsymbol{F^D})^T \boldsymbol{F}^D$.

$\frac{\partial \zeta_i}{\partial p_j}$ are given by

$$\frac{\partial \zeta_1}{\partial p_j} = \sum_{i=1}^{J} [F_i^{\phi} \frac{\partial F_i^D}{\partial p_i} + F_i^D \frac{\partial F_i^{\phi}}{\partial p_i}]$$

$$\frac{\partial \zeta_2}{\partial p_j} = 2 \sum_{i=1}^{J} F_i^{\phi} \frac{\partial F_i^{\phi}}{\partial p_i}$$

$$\frac{\partial \zeta_3}{\partial p_j} = 2 \sum_{i=1}^{J} F_i^D \frac{\partial F_i^D}{\partial p_i}$$

and

$$\frac{\partial F_j^{\phi}}{\partial p_i} = \frac{\partial^2 \phi}{\partial p_i p_j} - [\frac{\partial \phi}{\partial p_i} + \sum_{i=1}^{J} p_t \frac{\partial^2 \phi}{\partial p_t p_i}]$$

$$\frac{\partial F_j^D}{\partial p_i} = \frac{\partial^2 D}{\partial p_i p_j} - [\frac{\partial D}{\partial p_i} + \sum_{i=1}^{J} p_t \frac{\partial^2 D}{\partial p_t p_i}]$$

First and second derivatives of $D(\boldsymbol{p})$:

$$\frac{\partial D}{\partial p_i} = d_i^D = 2\boldsymbol{V}^T \boldsymbol{V} \boldsymbol{p} - 2\boldsymbol{V}^T \boldsymbol{\mu}$$

$$\frac{\partial^2 D}{\partial p_t p_i} = d_{ij}^D = 2\boldsymbol{V}^T \boldsymbol{V}$$

First and second derivatives of ϕ depend on the choice of the criterion function ϕ.

For D-optimality,

$$\frac{\partial \phi}{\partial p_i} = d_j^{\phi} = \boldsymbol{v_j}^T \boldsymbol{M}^{-1} \boldsymbol{v}_j$$

$$\frac{\partial^2 \phi}{\partial p_i p_j} = d_{ij}^{\phi} = -[\boldsymbol{v_j}^T \boldsymbol{M}^{-1} \boldsymbol{v}_i]^2$$

We finally make sure that we simultaneously maximize U and T (that is, we maximize either $U(\boldsymbol{p})+T(\boldsymbol{p})$ or $min\{U(\boldsymbol{p}), T(\boldsymbol{p})\}$) with a common maximum of zero.

In the following sub-section, we will set up the problem and report the results.

3.2.2 Example: Problem Specification

We assume IPD are generated 10 values for x_1 (age), from 15 to 30. The length out is 10. For x_2 (gender), we have 0 and 1 (0 for male and 1 for female). For AgD, we take μ_1 and μ_2 as 24 and 0.3 respectively. We choose D-optimalily as our main criterion.

age	15, 16.66667, 18.33333, 20, 21.66667, 23.33333, 25, 26.66667, 28.33333, 30
gender	0 1

So we have

$$IPD: \quad \begin{aligned} x_1 &= (15, 16.66667, 18.3333320, ..., 28.33333, 30) \\ x_2 &= (0, 1) \end{aligned}$$

and

$$AgD: \quad \begin{aligned} \mu_1 &= 24 \\ \mu_2 &= 0.3 \end{aligned}$$

As we mentioned earlier, the design space consisting of all pairs (x_1, x_2), x_1 are the values of age, x_2 is binary variable 0 or 1. The space consists of $10 \times 2 = 20$ pairs (x_1, x_2). So, we can write the design as:

$$p = \begin{Bmatrix} (x_{11}, x_{21}) & (x_{12}, x_{22}) & \cdots & (x_{120}, x_{220}) \\ p_1 & p_2 & \cdots & p_{20} \end{Bmatrix}$$

Baseline characteristics constraints are given in the following:

$$\sum_{j=1}^{20} p_j x_{1j} = p_1(15) + p_2(16.66667) + p_3(18.33333) + ... + p_{20}(30) = 24$$
$$\sum_{j=1}^{20} p_j x_{2j} = p_1(0) + p_2(0) + p_3(0) + ... + p_{20}(1) = 0.3$$

According to the above setup, $\boldsymbol{V}$ matrix is given by

$$\boldsymbol{V} = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 & 1 \\ 15.00000 & 16.66667 & \cdots & 26.66667 & 28.33333 & 30.00000 \\ 0 & 0 & \cdots & 1 & 1 & 1 \end{pmatrix}$$

The matrix $\boldsymbol{V}$ is (3 x 20). $\boldsymbol{\mu}$ is a (3×1) vector and $\boldsymbol{p}$ is a (20×1) vector.

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 24 \\ 0.3 \end{pmatrix}$$

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ \dots \\ p_{19} \\ p_{20} \end{pmatrix}$$

3.2.3 Results

The choice of $f(.)$ plays an essential role in the convergence of the algorithms. The convergence rates depends on function $f(.)$. The free positive parameter δ could also affect the convergence rates. Three functions would be considered in this work. A suitable choice is a function which is centred on zero and changes reasonably quickly about zero.

$$\begin{aligned} f(z) &= exp(\delta z) \\ f(z) &= exp(\delta z)/(1 + exp(\delta z)) \quad \text{the logistic c.d.f.} \\ f(z) &= \Phi(\delta z) \quad \text{the normal c.d.f} \end{aligned} \tag{3.35}$$

We start with choice of $f(z) = exp(\delta z)$, with z as the sum for the directional derivatives U and T, where $z_j^{(r+1)} = d_j^U + d_j^T$, $\delta = 0.01$. We take the initial weights as: $p_j^0 = 1/J$, $J = 20$. The above setup gives the starting values: $U = -264102410, T = -5.2441$, $-103.966 \leq F_j^T \leq 103.966$, $0 \leq F_j^U \leq 0$.

After 141120 iterations, the direction derivatives $F_j's$ of U, T and z are given by

$$T = -0.0000001551191$$

$$U = -0.002132791$$

$$-0.00002126505 \leq F_j^T \leq 0.000009999988$$

$$F_j^U = 0$$

$$-0.00002126505 \leq F_j^z \leq 0.000009999988$$

The optimal design p^* leads to:

$$\sum_{j=1}^{20} p_j^* x_{1j} = 24$$
$$\sum_{j=1}^{20} p_j^* x_{2j} = 0.3198457$$

We obtain the optimal design as given by:

Table 3.1: Optimal design points on without interaction model

x_1	x_2	p^*
15.00000	0	0.03917533
16.66667	0	0.04378956
18.33333	0	0.04894728
20.00000	0	0.05471249
21.66667	0	0.06115676

23.33333	0	0.06836005
25.00000	0	0.07641178
26.66667	0	0.08541188
28.33333	0	0.09547204
30.00000	0	0.10671713
15.00000	1	0.01842238
16.66667	1	0.02059224
18.33333	1	0.02301768
20.00000	1	0.02572880
21.66667	1	0.02875925
23.33333	1	0.03214663
25.00000	1	0.03593299
26.66667	1	0.04016533
28.33333	1	0.04489617
30.00000	1	0.05018422

To ensure the method provides good solutions, there are prerequisites that need to be checked. We first check the optimality conditions. The directional derivatives must satisfy the first order conditions. Second, we check the constraint conditions. The design must match the pre-defined baseline characteristics.

From the above result, we can see the maximum directional derivatives are less than equal to 0.00000999998, almost equal to 0. The range of directional derivatives

values is almost close to zero. This satisfies the optimality conditions. Substituting the design weights p^* into constraint functions, the sum of IPD on Age of design weights is equal to the AgD μ_1, the sum of proportion IPD on male is equal to 0.3199457, which is close to the AgD μ_2. In other words, the design is not only optimal but also matches with the pre-defined baseline characteristics $\boldsymbol{\mu}$. The methodology performed really well. Convergence rates of algorithm depends on the choice of the function $f(z)$ and the positive parameter δ.

We chose $f(z) = exp(\delta z)$, $f(z) = \Phi(\delta z)$, and $f(z) = exp(\delta z)/(1 + exp(\delta z))$, with δ values $\delta = 0.001$, $\delta = 0.007$ and $\delta = 0.01$. We record the number of iterations (for n = 1, 2, 3, 4, 5) to achieve the first order condition $max\{F_j\} \leq 10^n$, for $j = 1, 2, ..., J$. The initial design was $p_j^0 = 1/J, j = 1, 2, ..., J$.

Now compare the result of Table 3.2, Table 3.3 and Table 3.4, we can see that the function $f(z) = exp(\delta z)$ gives the fast convergence result. For example, with $f(z) = exp(\delta z)$ and $\delta = 0.001$, the number of iterations needed to achieve the first order condition for n=5 is 1412109. For $f(z) = \Phi(\delta z)$ with $\delta = 0.001$, the number of iterations needed to achieve the first order condition for n=5 is 1769995. For $f(z) = exp(\delta z)/(1+exp(\delta z))$, with $\delta = 0.001$ the number of iterations needed for n=5 is 2824377. Now consider the suitable choice of δ. For example $f(z) = exp(\delta z)$ with $\delta = 0.01$, the number of iterations needed to achieve the condition for n=5 is 141120, whereas for $\delta = 0.001$ is 1412109, for $\delta = 0.007$ is 201700. Therefore, if we compare the number of iterations in this case, the best choice is the function $f(z) = exp(\delta z)$ with $\delta = 0.01$.

Table 3.2: Without interaction term model, $f(z) = exp(\delta z)$

$f(z) = exp(\delta z)$ Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$					
δ	n=1	n=2	n=3	n=4	n=5
0.001	292	3440	55384	294321	1412109
0.007	54	466	7884	42017	201700
0.01	22	262	5450	29343	141120

Table 3.3: Without interaction term model, $f(z) = \Phi(\delta z)$

$f(z) = \Phi(\delta z)$ Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$					
δ	n=1	n=2	n=3	n=4	n=5
0.001	367	4309	69418	368911	1769995
0.007	48	656	9975	52843	253400
0.01	53	396	6934	36998	177657

Table 3.4: Without interaction term model, $f(z) = exp(\delta z)/(1 + exp(\delta z))$

$f(z) = exp(\delta z)/(1 + exp(\delta z))$ Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$					
δ	n=1	n=2	n=3	n=4	n=5
0.001	586	6864	110760	588664	2824377
0.007	79	1022	15874	84198	403836
0.01	2	755	11159	59023	282949

Let us consider one more example with the initial design using the observed data referred to the above example. We change the value δ to 0.0001, in function $f(z\delta) = exp(\delta z)$, we take n=6 in the first order condition $F_j \leq 10^{-n}$. After 66389674 iterations, the direction derivatives $F_j's$ of U, T and z we get the following results:

$$T = 0.000000007528679$$

$$U = -0.000106369$$

$$0.000002232951 \leq F_j^T \leq 0.000001$$

$$F_j^U = 0$$

$$0.000002232951 \leq F_j^z \leq 0.000001$$

And the p^* leads to

$$\sum_{j=1}^{20} p_j^* x_{1j} = 24$$
$$\sum_{j=1}^{20} p_j^* x_{2j} = 0.3093149$$

Here we see the directional derivatives closely satisfy the first order condition, the U is more close to 0. The design gives more closer match to the pre-defined baseline characteristics $\mu_2 = 0.3$ for gender, where $\sum_{j=1}^{20} p_j^* x_{2j} = 0.3093149$. In the previous example we had $\sum_{j=1}^{20} p_j^* x_{2j} = 0.3198457$.

In order to make an honest comparison, we report the algorithm running time in Tables 3.5 to 3.7 for the three choices of the function in the multiplicative algorithm, namely the exponential function, the normal cdf and the logistic cdf. As we see in

Table 3.5: Algorithm running time: without interaction term model, $f(z) = exp(\delta z)$

$f(z) = exp(\delta z)$ Algorithm running time needed to achieve $max\{F_j\} \leq 10^{-n}$					
δ	n=1	n=2	n=3	n=4	n=5
0.001	0.367733 secs	2.607782 secs	40.39536 secs	3.616298 mins	12.57052 mins
0.007	0.191402 secs	0.4745774 secs	5.804641 secs	31.71404 secs	3.041113 mins
0.01	0.1424165 secs	0.3463619 secs	3.936565 secs	21.35258 secs	1.874617 mins

Table 3.6: Algorithm running time: without interaction term model, $f(z) = \Phi(\delta z)$

$f(z) = \Phi(\delta z)$ Algorithm running time needed to achieve $max\{F_j\} \leq 10^{-n}$					
δ	n=1	n=2	n=3	n=4	n=5
0.001	0.8483758 secs	4.212314 secs	51.99371 secs	4.696498 mins	21.1393 mins
0.007	0.4318831 secs	0.6111541 secs	7.703218 secs	39.40134 secs	3.177345 mins
0.01	0.1894691 secs	0.4503889 secs	5.021205 secs	28.18651 secs	2.148628 mins

Table 3.7: Algorithm running time: without interaction term model, $f(z) = exp(\delta z)/(1 + exp(\delta z))$

$f(z) = exp(\delta z)/(1 + exp(\delta z))$ Algorithm running time needed to achieve $max\{F_j\} \leq 10^{-n}$					
δ	n=1	n=2	n=3	n=4	n=5
0.001	1.029904 secs	5.191191 secs	1.385527 mins	7.361828 mins	33.91572 mins
0.007	0.5444775 secs	1.286603 secs	12.44664 secs	1.092825 mins	5.250727 mins
0.01	0.1392734 secs	0.7136967 secs	8.283533 secs	44.18676 secs	3.71339 mins

the tables, running time depends on both the choice and the function $f(z)$ and the tuning parameter δ. When δ takes the value 0.01, running time is minimum for the exponential function. It is important to note here is that running time also depends for different criteria, for example, if we choose a different criterion (other than D-optimality) in our main criterion in the Lagrangian formulation. Running time also depends on the number of constraints we have in the Lagrangian formulation.

3.3 Two Variable Model with Interaction Term

Now we consider the two-variable model and add the interaction between the two variables x_1 and x_2. The model is

$$\begin{aligned} E(y|x_1, x_2) &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 \\ &= \boldsymbol{v}^T \boldsymbol{\theta} \end{aligned} \tag{3.36}$$

where

$$\boldsymbol{v}_x = (1, x_1, x_2, x_1 x_2)^T, \boldsymbol{v} \in \mathcal{V};$$

$$\mathcal{V} = \{\boldsymbol{v} \in R^k : \boldsymbol{v} = (1, x_1, x_2, x_1 x_2)^T, \boldsymbol{x} \in \chi\}$$

We consider the discretized design space consisting of all pairs (x_1, x_2). The space

will consists of $i \times j = J$ pairs (x_1, x_2). The design is given by

$$p = \begin{Bmatrix} (x_{11}, x_{21}) & (x_{12}, x_{22}) & \cdots & (x_{1J}, x_{2J}) \\ p_1 & p_2 & \cdots & p_J \end{Bmatrix}. \tag{3.37}$$

3.3.1 Formulation of the Problem

As we did before, we define x_1 as the age characteristic variable, x_2 as the gender characteristic variable. In AgD, μ_1 is the aggregate baseline characteristic age, and μ_2 is the proportion of aggregate baseline characteristic male. Here the $\boldsymbol{V}$ matrix is given by

$$\boldsymbol{V} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1J} \\ x_{21} & x_{22} & \cdots & x_{2J} \\ x_{11}x_{21} & x_{12}x_{22} & \cdots & x_{1J}x_{2J} \end{pmatrix} \tag{3.38}$$

In two variable without interaction model, matching pre-defined baseline characteristic constraint functions satisfy $\sum_{j=1}^{J} p_j x_{1j} = \mu_1$, and $\sum_{j=1}^{J} p_j x_{2j} = \mu_2$. In the two-variable model with interaction, adding an interaction is like adding one more variable, so the constraint function needs to satisfy for the interaction term as well. As we mentioned earlier, we use the D-optimality criterion, so we maximize the D-optimality criterion subject to the constraints as given below:

Maximize $\phi(\boldsymbol{p}) = logdet(\boldsymbol{M(p)})$, subject to

$$\begin{aligned}
&\sum_{j=1}^{J} p_j x_{1j} = p_1 x_{11} + p_2 x_{12} + p_3 x_{13} + ... + p_J x_{1J} = \mu_1 \\
&\sum_{j=1}^{J} p_j x_{2j} = p_1 x_{21} + p_2 x_{22} + p_3 x_{23} + ... + p_J x_{2J} = \mu_2 \\
&\sum_{j=1}^{J} p_j x_{1j} \sum_{j=1}^{J} p_j x_{2j} = \mu_1 \mu_2
\end{aligned} \tag{3.39}$$

Set up $\boldsymbol{\mu}$ and $\boldsymbol{p}$:

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ \mu_1 \\ \mu_2 \\ \mu_1 \mu_2 \end{pmatrix} \tag{3.40}$$

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ \cdots \\ p_J \end{pmatrix}$$

Then for $\boldsymbol{Vp = \mu}$, the constraint function would be:

$$D(\boldsymbol{p}) = (\boldsymbol{Vp - \mu})^T(\boldsymbol{Vp - \mu}) = \boldsymbol{p}^T\boldsymbol{V}^T\boldsymbol{Vp} - 2(\boldsymbol{V}^T\boldsymbol{\mu})^T\boldsymbol{p} + \boldsymbol{\mu}^T\boldsymbol{\mu} \tag{3.41}$$

We now choose the specifications for the variables and the constraints and construct the design in the following.

3.3.2 Example: Problem Specification

We use the same data as we used in the model without interaction. We assume IPD are generated for 10 values for x_1 (age), from 15 to 30. For x_2 (gender), we have 0 and 1 (0 for male and 1 for female). For AgD, we take μ_1 and μ_2 as 24 and 0.3 respectively. We choose the D-optimalily as our main criterion.

age	15, 16.66667, 18.33333, 20, 21.66667, 23.33333, 25, 26.66667, 28.33333, 30
gender	0 1

IPD :

$$x_1 = (15, 16.66667, 18.3333320, ..., 28.33333, 30)$$
$$x_2 = (0, 1)$$

and

AgD :

$$\mu_1 = 24$$
$$\mu_2 = 0.3$$

Where vector matrix is (4×20):

$$\boldsymbol{V} = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 & 1 \\ 15.00000 & 16.66667 & \cdots & 26.66667 & 28.33333 & 30 \\ 0 & 0 & \cdots & 1 & 1 & 1 \\ 0 & 0 & \cdots & 26.66667 & 28.33333 & 30 \end{pmatrix}$$

$\boldsymbol{\mu}$ is a (1×4) matrix

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 24 \\ 0.3 \\ 7.2 \end{pmatrix}$$

and $\boldsymbol{p}$ is a (20×1) matrix with elements $p_1, p_2, \ldots, p_{20}$.

Baseline characteristic constraints are given below.

$$\sum_{j=1}^{20} p_j x_{1j} = p_1(15) + p_2(16.66667) + p_3(18.33333) + ... + p_{20}(30) = 24$$
$$\sum_{j=1}^{20} p_j x_{2j} = p_1(0) + p_2(0) + p_3(0) + ... + p_{20}(1) = 0.3$$
$$\sum_{j=1}^{20} p_j x_{1j} \sum_{j=1}^{20} p_j x_{2j} = 7.2$$

We now report the results in the following.

3.3.3 Results

We first choose $f(z\delta) = exp(\delta z)$, $\delta = 0.0001$. We take n=5 in the directional derivatives $F_j \leq 10^{-n}$ at p^*. The initial weights start at $p_j^0 = 1/20$. After 1185605 iterations, the directional derivatives $F_j's$ of U, T and z are given by

$$T = -0.000000005409875$$

$$U = -0.064057$$

$$-0.00001552158 \leq F_j^T \leq 0.00001$$

$$F_j^U = 0$$

$$-0.00001552158 \leq F_j^z \leq 0.00001$$

And the p^* lead

$$\sum_{j=1}^{20} p_j^* x_{1j} = 23.99456$$
$$\sum_{j=1}^{20} p_j^* x_{2j} = 0.3066285$$
$$\sum_{j=1}^{20} p_j x_{1j} \sum_{j=1}^{20} p_j x_{2j} = 7.357416$$

We can write the optimal design as given below:

Table 3.8: Optimal design points on with interaction model

x_1	x_2	p^*
15.00000	0	0.03632907
16.66667	0	0.04131745
18.33333	0	0.04699079
20.00000	0	0.05344314
21.66667	0	0.06078147
23.33333	0	0.06912744
25.00000	0	0.07861939
26.66667	0	0.08941470
28.33333	0	0.10169232
30.00000	0	0.11565579
15.00000	1	0.02171471
16.66667	1	0.02333478
18.33333	1	0.02507571
20.00000	1	0.02694653
21.66667	1	0.02895693
23.33333	1	0.03111731
25.00000	1	0.03343887
26.66667	1	0.03593364
28.33333	1	0.03861453
30.00000	1	0.04149544

To ensure the methodology performs well, the prerequisites need to be checked. First is the optimality conditions. The directional derivatives must be less than or equal to 0. The second is the constraint conditions. The design weight must match the pre-defined baseline characteristics.

From the above result, we can see the maximum directional derivatives of z, F_j^z are less than equal to 0.00001, almost equal to 0. The range of directional derivatives values is almost close to zero. This satisfies the optimality conditions. Substitute design weights p^* into constraint functions, the sum of IPD Age of design weights is equal to 23.99456, the sum of proportion IPD male is equal to 0.3066285, the AgD $\mu_1 = 24$ and $\mu_2 = 0.3$. In other words, the optimal design approximately matches the pre-defined baseline characteristics μ.

Table 3.9: With interaction term model, $f(z) = exp(\delta z)$

$f(z) = exp(\delta z)$ Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$					
δ	n=1	n=2	n=3	n=4	n=5
0.00007	3238	15179	70941	336294	1693723
0.0001	2267	10625	49659	235405	1185606
0.0002	1115	5293	24810	117683	592784

Now we look at the results of Table 3.6, Table 3.7 and Table 3.8. Choice of $\delta = 0.002$ gives the fast convergence result. In Table 3.6 the function $f(z) = exp(\delta z)$ with $\delta = 0.0002$ gives the fast convergence result. The number of iterations needed to achieve the first order conditions for n=5 is 592784. In Table 3.7, with function $f(z) = \Phi(\delta z)$, $\delta = 0.0002$, the number of iterations needed to achieve the first order

Table 3.10: With interaction term model, $f(z) = \Phi(\delta z)$

$f(z) = \Phi(\delta z)$ Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$					
δ	n=1	n=2	n=3	n=4	n=5
0.00007	4059	19025	88926	421864	2129195
0.0001	2842	13313	62235	295590	1496572
0.0002	1408	6620	31045	149105	768019

Table 3.11: With interaction term model, $f(z) = exp(\delta z)/(1 + exp(\delta z))$

$f(z) = exp(\delta z)/(1 + exp(\delta z))$ Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$					
δ	n=1	n=2	n=3	n=4	n=5
0.00007	6476	30363	141914	672844	3390305
0.0001	4534	21253	99341	471139	2376039
0.0002	2268	10620	49656	236020	1197201

conditions for n=5 is 768019. In table 3.8, with $f(z) = exp(\delta z)/(1 + exp(\delta z))$, $\delta = 0.0002$ the number of iterations needed to achieve the first order condition for n=5 is 1197201. Comparing the three functions, the $f(z) = exp(\delta z)$ gives fastest convergence result as we see the number of iterations needed to achieve the first order condition for n=5 has minimum iterations.

Chapter 4

Sensitivity Analysis and Comparison of the Methods

4.1 Introduction

We discussed using the Lagrangian method to construct the optimal designs subject to matching pre-defined baseline characteristics in Chapter 3. This method provides the desired result. As we mentioned earlier, the research problem was inspired by applying the optimal design theory to matching-adjusted indirect comparison method. The objective was to construct optimal design subject to independent patient data (IPD) matches to baseline characteristics of AgD. In optimization theory, minimum variance is the goal achieved by optimizing the variance function. We add constraints that the design needs to match the baseline characteristics of AgD. We use Lagrangian method to achieve this combined goal. The Lagrangian method is a strategy for finding the local maxima and minima of a function subject to constraints. The optimal

design obtained by the Lagrangian method is denoted by p^*. The MAIC method reweighs IPD to match the distribution at the aggregate data (AgD), most often using the mean and standard deviation of baseline characteristics (Alsop and Pont, 2022). The design weight obtained by the MAIC method is denoted by w.

In this chapter, we use Kish's (approximate) effective sample size (ESS) to compare the two methods. For MAIC method, the ESS is usually used to measure the accuracy and precision of matching. More references can be found at Alsop and Pont (2022), and Signorovitch et al. (2010). Kish's (approximate) effective sample size (ESS) is a simplified formula used for unequal weights of effective sample size (Kish, 1965). Effective sample size is the original sample size divided by the design effect, that is, $\frac{n}{D_{eff}}$. Here D_{eff} is design effect, which is the ratio of two variances for estimators of some parameters, and n is the sample size. Optimal design minimizes the variance. Because of the variance connection between ESS and optimal design, we can calculate the ESS for the corresponding optimal design. The formula of ESS is given by

$$ESS = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2} \tag{4.1}$$

where w_i is the estimated weight for i^{th} patient. For the Lagrangian method, we replace weight w_i to weight p_i^*. The ESS represents the number of patients included in the analysis after weighting. The sum of the weights is constrained to be the number of patients n and all weights are non-negative. The effective sample size (ESS) is maximum at n when all w_i are equal to 1, in which case we weigh each patient equally in the analysis and there is no loss of information due to re-weighting.

However, weighting always reduces the effective sample size. The higher the value of ESS, the less information will be lost due to reweighting. That is, the larger the effective sample size, the better effective performance will be (Alsop and Pont, 2022).

The estimated weights can be achieved by multiple sets. For example, suppose we want to match IPD from 4 patients to AgD baseline on a mean of 4.3. We denote four patients with baseline responses: 2, 3, 4, and 7. How do we choose the four weights (w_1, w_2, w_3, w_4)?

One set of weights

$$w_1 = 0.19, w_2 = 1.23, w_3 = 1.66, w_4 = 0.92$$

The ESS is

$$ESS = \frac{(0.19 + 1.23 + 1.66 + 0.92)^2}{0.19^2 + 1.23^2 + 1.66^2 + 0.92^2} = 3.1$$

Alternative set of weights

$$w_1 = 0.57, w_2 = 0.17, w_3 = 2.38, w_4 = 0.88$$

The ESS is

$$ESS = \frac{(0.57 + 0.17 + 2.38 + 0.88)^2}{0.57^2 + 0.17^2 + 2.38^2 + 0.88^2} = 2.4$$

Comparing the two sets of weights, the alternative set of weights should be avoided, because this choice results in a loss of information.

4.2 Illustrative Examples for the Model without Interaction Term

Now, let us give some examples in model: $E(y|x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$. Suppose age in IPD generates from a normal distribution. We consider age as a symmetrical and continuous distribution. So normal distribution is a natural choice. Another possibility is the uniform distribution. We wish to match IPD from n patients to AgD on baseline characteristics age μ_1 and gender proportion μ_2. Consider an example with number of patients $n = 200$, and $Age \sim N(\mu = 30, \sigma = 2)$. Gender is defined by 1 or 0. The design space consists of 200 pairs. The design with equal initial weights is given by

$$p = \begin{Bmatrix} (x_{11}, x_{21}) & (x_{12}, x_{22}) & \cdots & (x_{1200}x_{2200}) \\ 0.005 & 0.005 & \cdots & 0.005 \end{Bmatrix}$$

The match value of AgD must be within the range of IPD in order to have a numerically meaningful result. For example, when there is only one baseline variable to match, if AgD is 40, whereas the IPD ranges from minimum value 18 to 35, then MAIC will not be able to match the IPD data onto the AgD. In this example, the sample mean age is 29.97436, the sample proportion of male 0.5. We set $\mu_1 = 31.5$, the proportion of male is $\mu_2 = 0.2$.

Let us first obtain the effective sample size (ESS) associated with the MAIC method. Let x_i be the individual patients data, $i = 1, 2, ..n$. From chapter 2, given

IPD, the odds that patients assigned for AgD versus patients assigned to IPD is

$$wi = \frac{P(AgD|x_i)}{P(IPD|x_i)}.$$

Now, we are ready to proceed with our analysis. We use the designed R package to calculate the MAIC weight. The R code is formulated from the Technical Support Document of Signorovitch et al. (2010).

4.2.1 Calculation process

Simulated IPD

We tabulate the IPD and we aggregate the IPD to obtain the summaries. We printed the results shown below. In the IPD table, each row represents baseline characteristics of patients, 0 represents male, and 1 represents female. There is a total of 200 patients. The mean age in IPD is 29.97436, and the proportion of males is 0.5.

```
# Generate age
 age<-rnorm(100,30,2)
 # Define gender
 gender<-c(0,1)
 # Candidate set for age and gender assigned to patient
 t = expand.grid(age,gender)
 #  Create IPD data set
 IPD<-data.frame( age = t$Var1, # Generate ages
                  gender = t$Var2) # Generate genders)
```

```r
# Summary IPD data set
Summarise_IPD <- IPD<-data.frame(mean(age),mean(gender))
```

```
> IPD
# A tibble: 200 × 2
     age gender
   <dbl>  <dbl>
 1  30.5      0
 2  29.3      0
 3  29.4      0
 4  25.4      0
 5  29.7      0
 6  30.3      0
 7  27.0      0
 8  28.0      0
 9  28.1      0
10  29.0      0
# … with 190 more rows
> Summarise_IPD
  mean.age. mean.gender.
1  29.97436          0.5
```

Simulated AgD

After simulated IPD, we can define AgD, $\mu_1 = 31.5$ and $\mu_2 = 0.3$ in AgD.

```r
# Define value of AgD
mu_1<-31.5 # aggregate age
mu_2 <-0.3 # aggregate proportion of male
# Create IPD data set
AgD<-data.frame(mu_1,mu_2)
```

Calculate MAIC weights

We first estimate the individual weights by estimating the logistics propensity score model:

$$log(w_i) = \alpha + \alpha_1^T X^{EM}$$

It is equivalent to minimizing

$$\hat{w_i} = exp(\alpha_1^T X^{EM}) \tag{4.2}$$

when $\bar{X}^{EM} = 0$.

To solve this, we define the objective for minimizing the above function. The gradient function will be used by the minimization algorithm.

```
# Objective function
objfn <- function(a1, X){
  sum(exp(X %*% a1))
}

# Gradient function
gradfn <- function(a1, X){
 colSums(sweep(X, 1, exp(X %*% a1), "*"))
}
```

To make sure the $\bar{X}^{EM} = 0$, we subtract mean of IPD to AgD.

```

X.EM.0<-sweep(with(IPD, cbind(IPD$age,IPD$gender)),2,
```

```
          with(AgD,
               c(mu_1,mu_2),'-'))
```

α_1 is estimated by using the function optim that minimized the function objfn. In R, we can call the method BFGS (after Broyden, Fletcher, Goldfarb and Shanno) to make the gradient function gradfn for minimization. For the initial value in the par argument we specify as c(0,0), and X=X.EM.0 is passed to objfn and gradfn as an additional argument in here.

```
# Estimate weights BFGS" a quasi-Newton method
print(opt1 <- optim(par = c(0,0), fn = objfn,
      gr =gradfn, X = X.EM.0, method = "BFGS"))
a1 <- opt1$par
```

Each individual patient weight data is estimated by $\hat{w}_{it}$

```
# Estimated weights for each individual
wt <- exp(X.EM.0 %*% a1)
```

After estimating weights for each individual $\hat{w}_{it}$, we calculate the MAIC weight as:

$$w_i = \frac{\hat{w}_{it}}{\sum_{i,t} \hat{w}_{it}}$$

```
# MAIC wight
w <- (wt / sum(wt))
```

Calculate MAIC weighted average characteristics

After calculating MAIC weight w_i, we calculate the weighted characteristics:

$$\sum_{j=1}^{200} w_i x_{1j} = 31.5$$
$$\sum_{j=1}^{200} w_i x_{2j} = 0.3$$

Calculate ESS of MAIC

We calculate the ESS for the MAIC method:

$$ESS_{MAIC} = \frac{(\sum_i^n w_i)^2}{\sum_i^n w_i^2} = 103.8$$

Lagrangian Method

We now need to find weighted average of characteristics and ESS using Lagrangian method. In Chapter 3, the optimal design weight p^* is calculated by the Lagrangian method. We consider Age $\sim N(\mu = 30, \sigma = 2)$, sample size n = 200. We choose $f(z\delta) = exp(\delta z)$, $\delta = 0.005$. We take $n = 6$ in the directional derivatives $F_j \leq 10^{-n}$ at p^*. The initial weights start at $p_j^0 = 1/20$. After 1327225 iterations, the direction

derivatives $F_j's$ of U, T and z are given by

$$T = -0.000000007528671$$

$$U = -0.03964046$$

$$-0.000002232949 \leq F_j^T \leq 0.0000009999991$$

$$F_j^U = 0$$

$$-0.000002232949 \leq F_j^z \leq 0.0000009999991$$

The p^* leads to

$$\sum_{j=1}^{200} p_j^* x_{1j} = 31.5$$
$$\sum_{j=1}^{200} p_j^* x_{2j} = 0.3$$

Then, we calculate the ESS for the Lagrangian method:

$$ESS = \frac{(\sum_i^n p_i^*)^2}{\sum_i^n p_i^{*2}} = 105.1$$

We will further consider more examples in this model, and combine all the results together.

Further examples

1. $E(y|x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$,

   (a) Age $\sim N(\mu = 30, \sigma = 2)$, Number of patients n= 400

(b) Age $\sim N(\mu = 50, \sigma = 5)$, Number of patients n= 200

(c) Age $\sim N(\mu = 50, \sigma = 5)$, Number of patients n= 400

4.2.2 Summary of Results

We already gave an example of the calculation procedure of Lagrangian method and MAIC method. We will now focus on interpretation. The summary result is shown in table 4.1. The values are rounded to 1 decimal place. The Age column listed the distribution of age selection, the AgD column listed the pre-defined baseline characteristics, mean of age μ_1 and the proportion of gender μ_2 in AgD, the Lagrangian/-MAIC column shows the weighted average of characteristics match to μ_1 and μ_2 in AgD by using Lagrangian method and MAIC method. The ESS_L/ESS_{MAIC} listed the value of ESS calculated by optimal design weights (using Lagrangian method) and ESS calculated by MAIC weight. The sample size column is the number of patients assigned in design.

Table 4.1: ESS in Model: $E(y|x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

Age	AgD	Lagrangian/MAIC	ESS_L/ESS_{MAIC}	Sample size
$N(\mu = 30, \sigma = 2)$	$\mu_1 = 31.5$ $\mu_2 = 0.3$	31.5/31.5 0.3/0.3	105.14/103.8	200
$N(\mu = 30, \sigma = 2)$	$\mu_1 = 31.5$ $\mu_2 = 0.3$	31.5/31.5 0.3/0.3	204.2/199.2	400
$N(\mu = 50, \sigma = 5)$	$\mu_1 = 51.5$ $\mu_2 = 0.3$	51.5/51.5 0.3/0.3	160.2/158.1	200
$N(\mu = 50, \sigma = 5)$	$\mu_1 = 51.5$ $\mu_2 = 0.3$	51.5/51.5 0.3/0.3	321.9/314.4	400

Look at the age generated from $N(\mu = 30, \sigma = 2)$, AgD $\mu_1 = 31.5, \mu_2 = 0.3$. By using Lagrangian method, we obtained optimal design p^* that matches to baseline characteristics of AgD, which are 31.5 and 0.3. The MAIC method obtained same weighted average of characteristics that matches the characteristics of AgD, which are 31.5 and 0.3. The ESS of Lagrangian method is greater than the ESS of MAIC $ESS_L = 105.1 > ESS_{MAIC} = 103.8$. It is a reduction from the sample size of 200, but it is still reasonably large. If we increase the sample size to 400, when applied Lagrangian method we choose the same function $f(z\delta) = exp(\delta z)$, and parameter $\delta = 0.005$, n=6 in the directional derivatives $F_j \leq 10^{-n}$ at p^*. The Lagrangian method and MAIC method obtained the same values that matched the baseline characteristics in AgD. The $ESS_L = 204.2 > ESS_{MAIC} = 199.2$. For age generated from $N(\mu =$

$50, \sigma = 5$), $\mu_1 = 51.5$ and $\mu_2 = 0.3$ in AgD. Both the Lagrangian method and MAIC method obtain same matched values according to AgD and the ESS_L is greater than ESS_{MAIC}. In sample size 200, $ESS_L = 160.2 > ESS_{MAIC} = 158.1$; in sample size 400, $ESS_L = 321.9 > ESS_{MAIC} = 314.4$. A high ESS related to the sample size in the IPD would make the application of the Lagrangian method lose less information than the MAIC method. In model $E(y|x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$, we can conclude that the Lagrangian method and MAIC can obtain the same matched value, whereas applying the Langrangian method the sample size reduced from n to ESS less than the MAIC Method, the sample size in the IPD would make the application of the Lagrangian method lose less information than the MAIC method.

4.3 Illustrative Examples for the Model with Interaction Term

We now consider some examples with the model $E(y|x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2$. We will consider the following examples. When we apply Lagrangian method, we choose $f(z\delta) = exp(\delta z)$, $\delta = 0.005$. We take n=5 in the directional derivatives $F_j \leq 10^{-n}$ at p^* in the following examples.

1. $E(y|x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2$

   (a) Age $\sim N(\mu = 20, \sigma = 0.8)$, Number of patients n= 40

   (b) Age $\sim N(\mu = 20, \sigma = 0.8)$, Number of patients n=100

   (c) Age $\sim N(\mu = 30, \sigma = 1)$, Number of patients n= 80

(d) Age $\sim N(\mu = 30, \sigma = 1)$, Number of patients n= 100

4.3.1 Summary of Results

Table 4.2: ESS in Model: $E(y|x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2$

Age	AgD	Lagrangian/MAIC	ESS_L/ESS_{MAIC}	Sample size
$N(\mu = 20, \sigma = 0.8)$	$\mu_1 = 20.2$ $\mu_2 = 0.3$	20.2/20.2 0.3/0.3	22.2/21.9	40
$N(\mu = 20, \sigma = 0.8)$	$\mu_1 = 20.2$ $\mu_2 =0.3$	20.2/20.2 0.3/0.3	75.1/74.8	100
$N(\mu = 30, \sigma = 1)$	$\mu_1 = 30.8$ $\mu_2 = 0.7$	30.8/30.8 0.7/0.7	28.4/26.9	80
$N(\mu = 30, \sigma = 1)$	$\mu_1 = 30.8$ $\mu_2 = 0.7$	30.8/30.8 0.7/0.7	46.3/44.0	100

Look at Table 4.2, the results are rounded to one decimal. For Age $\sim N(\mu = 20, \sigma = 0.8)$, sample size n=40, the Lagrangian method and the MAIC method gave the same matched values according to characteristics of AgD, which are 20.2 and 0.3. The $ESS_L = 22.2 > ESS_{MAIC} = 21.9$. For the sample size 100, the $ESS_L = 75.1 > ESS_{MAIC} = 74.8$.

For age generated from $N(\mu = 30, \sigma = 1)$, baseline characteristics of AgD is $\mu_1 = 30.8$ and $\mu_2 = 0.7$. The Lagrangian and MAIC methods obtained the same values that matched the AgD. The $ESS_L = 28.4 > ESS_{MAIC} = 26.9$ in sample size

80. The $ESS_L = 46.3 > ESS_{MAIC} = 44.0$ in sample size 100. We can summarize that the ESS of Lagrangian is greater than ESS of MAIC. The sample size in the IPD would make the application of the Lagrangian method lose less information than the MAIC method in the model of $E(y|x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2$.

In conclusion, the Lagrangian method performs better than the MAIC method. Also, the Lagrangian method is quite flexible and can be implemented to different types of constraints (for example, other than the means in baseline characteristics).

Chapter 5

Conclusions and Future Work

5.1 Discussions

This thesis aimed to develop a method for optimizing design subject to matching the pre-defined baseline characteristics. This research was inspired by applying the optimal design theory to matching-adjusted indirect comparisons in clinical trials. We know that before a pharmaceutical company launches a new drug, an important step is to compare the new drug's effectiveness with existing similar drugs. Ideally, the two drugs can be compared directly. However, in practice, a direct comparison is normally impossible to achieve. Indirect comparisons with MAIC are increasingly being used in comparative validity studies. This method is to re-weight the existing individual new drug data to match the aggregated data of the existing drugs. In this way, comparisons can be made effectively. The critical point of this method is to match the summary data of the drugs that need to be compared. However, the experiments designed by this method only focus on the matching results and

cannot take into account the optimization problem of the experiment. Based on the above inspiration, we convert this indirect matching problem into an experiment with optimization constraints through our optimal design theory to satisfy the matching conditions.

In Chapter 2, we first reviewed methods and concepts in optimal design and determined the required optimality conditions of our problem along with a class of multiplication algorithms, indexed by functions that satisfy certain conditions. We also gave an overview of the application of matching adjusted indirect comparison methodology and its concept.

In Chapter 3, we formulated the problem as a constrained optimization problem using Lagrangian methods. We focused on constrained optimal design under the two variable models with and without the interaction term. We defined the available individual patient variables in the model as baseline characteristics of age and gender. We focused on the D-optimality criterion as we are interested in all of the parameters in the models. We successfully constructed optimal designs subject to matching the pre-defined baseline characteristics using Lagrangian method implemented by the multiplicative algorithm. This method worked out very well.

In Chapter 4, we compared the Lagrangian and MAIC methods in different models by calculating each method's effective sample size (ESS). We also summarized the weighted characteristics. We first compared their summarized weighted characteristics values. The two methods obtained the same matched value. Then, we compared their ESS values. The Lagrangian method outperforms the MAIC method in the two-variable model either with or without interaction term. We lose less information

in the Lagrangian method compared to the MAIC method.

5.2 Limitation and Future Work

This method also has certain limitations. First, not all the values can be matched. Second, the choice of $f(.)$ and the free parameter δ plays an important role here; not all $f(.)$ can be used. In this thesis, we only chose three functions, namely the logistic c.d.f $f(z) = exp(\delta z)/(1 + exp(\delta z))$, the normal cdf $f(z) = \Phi(\delta z)$ and the exponential function $f(z) = exp(\delta z)$. Since the derivatives of the Lagrangian function are very large and logistic c.d.f has a bounded value, so it behaves better than unbounded $exp(\delta z)$. Third, the directional derivatives satisfy the optimality conditions when taking more n. If we take more n, the result more closely matches pre-defined values. Thus, it will take more interactions, and so it requires high performance computing.

We only discussed the two baseline characteristics of age and gender and the D-optimal criteria. As for future work, we can extend to more variables, such as weight, region, and treatment, as well as using other optimality criteria such as the A-optimality and c-optimality criterion. When we deal with the constrained optimization problem, we only considered the one case by choosing the transformed problem that maximized the sum of two functions with common maximum of zero. Where maximized $\phi_L(\boldsymbol{p}) = U(\boldsymbol{p}) + T(\boldsymbol{p})$. We can extend to the *Maximin* problem in which we maximize $min\{U(\boldsymbol{p}), T(\boldsymbol{p})\}$. For data generation, we only considered the normal distribution. We can extend to uniform and other distributions as well.

Bibliography

Alsop, J. C. and L. O. Pont (2022). Matching-adjusted indirect comparison via a polynomial-based non-linear optimization method. *Journal of Comparative Effectiveness Research 11*(8), 551–561.

Atkinson, A., A. Donev, and R. Tobias (2007). *Optimum experimental designs, with SAS*, Volume 34. OUP Oxford.

Atwood, C. L. (1969). Optimal and efficient designs of experiments. *The Annals of Mathematical Statistics 40*, 1570–1602.

Berger, M. P. and W.-K. Wong (2009). *An introduction to optimal designs for social and biomedical research*. John Wiley & Sons.

Bucher, H. C., G. H. Guyatt, L. E. Griffith, and S. D. Walter (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of clinical epidemiology 50*(6), 683–691.

Deodhar, A. (2018). Mirror, mirror, on the wall, which is the most effective biologic of all? *The Journal of Rheumatology 45*(4), 449–450.

Dias, S., A. J. Sutton, A. Ades, and N. J. Welton (2013). Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making 33*(5), 607–617.

Elfving, G. (1952). Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics 23*, 255–262.

Fedorov, V. V. (2013). *Theory of optimal experiments*. Elsevier.

Karlin, S. and W. J. Studden (1966). Optimal experimental designs. *The Annals of Mathematical Statistics 37*(4), 783–815.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics 2*, 849–879.

Kiefer, J. and J. Wolfowitz (1959). Optimum designs in regression problems. *The Annals of Mathematical Statistics 30*(2), 271–294.

Kiefer, J. and J. Wolfowitz (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics 12*, 363–366.

Kish, L. (1965). Survey sampling. New York: John Wiley & sons.

Mandal, S. and B. Torsney (2006). Construction of optimal designs using a clustering approach. *Journal of statistical planning and inference 136*(3), 1120–1134.

Mandal, S., B. Torsney, and K. Carriere (2005). Constructing optimal designs with constraints. *Journal of statistical planning and inference 128*(2), 609–621.

Mandal, S., B. Torsney, and M. Chowdhury (2017). Optimal designs for minimizing covariances among parameter estimates in a linear model. *Australian and New Zealand Journal of Statistics 59*(3), 255–273.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John wiley & sons.

Nietert, P. J., A. E. Wahlquist, and T. L. Herbert (2013). Characteristics of recent biostatistical methods adopted by researchers publishing in general/internal medicine journals. *Statistics in Medicine 32*(1), 1–10.

Phillippo, D., T. Ades, S. Dias, S. Palmer, K. R. Abrams, and N. Welton (2016). Nice dsu technical support document 18: methods for population-adjusted indirect comparisons in submissions to nice.

Phillippo, D. M., A. E. Ades, S. Dias, S. Palmer, K. R. Abrams, and N. J. Welton (2018). Methods for population-adjusted indirect comparisons in health technology appraisal. *Medical Decision Making 38*(2), 200–211.

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

Rohde, C. A. (1965). Generalized inverses of partitioned matrices. *Journal of the Society for Industrial and Applied Mathematics 13*(4), 1033–1035.

Rugo, H. S., A. Haltner, L. Zhan, A. Tran, E. Bananis, B. Hooper, D. Mitra, and C. Cameron (2021). Matching-adjusted indirect comparison of palbociclib versus ribociclib and abemaciclib in hormone receptor-positive/her2-negative advanced breast cancer. *Journal of Comparative Effectiveness Research 10*(6), 457–467.

Sibson, R. (1972). Da-optimality and duality. In *Progress in Statistics, Colloquia Mathematica Societatis János Bolyai*, Volume 9, pp. 677–692.

Signorovitch, J. E., E. Q. Wu, A. P. Yu, C. M. Gerrits, E. Kantor, Y. Bao, S. R. Gupta, and P. M. Mulani (2010). Comparative effectiveness without head-to-head trials. *Pharmacoeconomics 28*(10), 935–945.

Silvey, S. (2013). *Optimal design: an introduction to the theory for parameter estimation*, Volume 1. Springer Science & Business Media.

Sverdlov, O., Y. Ryeznik, and W. K. Wong (2020). On optimal designs for clinical trials: an updated review. *Journal of Statistical Theory and Practice 14*(1), 1–29.

Torsney, B. (1977). Contribution to discussion of 'maximum likelihood estimation via the em algorithm' by Dempster et al. *Journal of the Royal Statistical Society B 39*, 26–27.

Torsney, B. and A. Alahmadi (1992). Further development of algorithms for constructing optimizing distributions. *Model oriented data analysis 215*, 121–129.

Torsney, B. and S. Mandal (2001). Construction of constrained optimal designs. In *Optimum design 2000*, pp. 141–152. Springer.

Torsney, B. and S. Mandal (2006). Two classes of multiplicative algorithms for constructing optimizing distributions. *Computational statistics & data analysis 51*(3), 1591–1601.