OPTIMAL DESIGN WHEN OUTCOME VALUES ARE NOT MISSING AT RANDOM

Kim May Lee, Robin Mitra and Stefanie Biedermann

University of Southampton, UK

Abstract: The presence of missing values complicates statistical analyses. In design of experiments, missing values are particularly problematic when constructing optimal designs, as it is not known which values are missing at the design stage. When data are missing at random it is possible to incorporate this information into the optimality criterion that is used to find designs; Imhof, Song, and Wong (2002) develop such a framework. However, when data are not missing at random this framework can lead to inefficient designs. We investigate and address the specific challenges that not missing at random values present when finding optimal designs for linear regression models. We show that the optimality criteria depend on model parameters that traditionally do not affect the design, such as regression coefficients and the residual variance. We also develop a framework that improves efficiency of designs over those found when values are missing at random.

Key words and phrases: Covariance matrix, information matrix, linear regression model, missing observations, not missing at random, optimal design.

1. Introduction

Missing values are a common problem in many fields. Their presence complicates statistical analysis, and appropriate methods are required to handle the missing data to ensure valid inferences. There is a wide variety of techniques to handle missing values once the data are observed, but the objective in this paper is to focus on handling the missing data problem at the design stage of an experiment. By incorporating information about the missing data mechanism we may be able to design a more efficient experiment that allows more information to be obtained from the data collected.

There has been work on finding optimal designs for experiments with potentially missing. The majority of the contributions is concerned with robustness of designs to missing values; see for example Hedayat and John (1974), Ghosh (1979), Ortega-Azurduy, Tan, and Berger (2008), and Ahmad and Gilmour (2010). Herzberg and Andrews (1976) and Hackl (1995) introduce design criteria that account for the presence of missing responses for some special cases. Imhof, Song, and Wong (2002) develop a framework that finds optimal designs by taking the expectation of the information matrix with respect to the missing data mechanism; this has been extended by Lee, Biedermann, and Mitra (2017) to improve the approximation of the

covariance matrix.

These contributions to optimal design implicitly assume that the data are missing at random, that the missing data mechanism depends on only observed variables. This is referred to as a missing at random (MAR), Rubin (1976). If it is assumed that the missing data mechanism depends on unobserved variables, such as the missing values themselves, Rubin (1976) referred to this as not missing at random (NMAR). Typically NMAR problems are much more challenging to handle, as learning about the exact form of the NMAR mechanism is not typically possible, and thus often leads to biased inferences.

To our knowledge there has not been any explicit consideration of dealing with NMAR when finding optimal designs. This article intends to address the specific problems that NMAR causes in optimal design. We mean to extend the framework of Imhof, Song, and Wong (2002) to incorporate the possibility of NMAR, using an approximation to the bias. By doing so we can mitigate the problems caused by NMAR and find more efficient designs.

We assume that inferences stem from a linear regression model once the experiment has been performed, and we deal with the missing data using the complete cases. Complete case analysis is a widely used strategy. In

the context of regression analysis, complete case analysis can be appropriate when the missing mechanism is MAR (Little (1992)). Under NMAR there are obvious problems that can occur and these will be noted and mitigated is our optimal design framework.

The remainder of the article is organised as follows. Section 2 presents some background for the key elements of missing data and optimal design. Section 3 motivates the problems NMAR causes in optimal design. Section 4 presents an optimal design framework that takes NMAR into account, and compares how it relates to the traditional MAR framework. Section 5 empirically evaluates the proposed framework to determine the benefits of using this approach. Section 6 evaluates our methodology in a data scenario. Section 7 ends with some concluding remarks.

2. Background

We review the relevant background for dealing with missing data, then present the key concepts in constructing optimal designs when a linear regression analysis model is used, and we review how the potential for missing data can be taken into account when finding optimal designs.

2.1 Missing data

Let $\mathbf{x}_i,\ i = 1, \ldots, n$, represent a set of explanatory variables for unit i in the experiment, and let y_i be the outcome for unit i once the experiment

is performed. We assume that inferences are drawn by fitting a linear regression model to the data of the form,

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \tag{1}$$

where $\mathbf{y} = (y_1, \ldots . y_n)$, $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of regression coefficients and $\epsilon \sim N(\mathbf{0}, \sigma^2 I)$ is the error vector with residual variance σ^2. We define a missing indicator, m_i, for each unit i; $m_i = 1$ if y_i is missing and $m_i = 0$ otherwise. We write $y_{mis} = \{y_i : m_i = 1\}$ and $y_{obs} = \{y_i : m_i = 0\}$ as the missing and observed outcomes, respectively. Typically, inference on $\boldsymbol{\beta}$ is made using the joint likelihood for (y_i, m_i). This can be expressed as

$$p(m_i|\mathbf{x}_i, y_i, \boldsymbol{\gamma})p(y_i|\mathbf{x}_i, \boldsymbol{\beta}), \tag{2}$$

known as the selection model framework (Little and Rubin (2002)), where the vector $\boldsymbol{\gamma}$ represents parameters characterising the model for m_i, also known as the missing data mechanism. We implicitly assume in this model that the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are distinct. Under MAR, $p(m_i|\mathbf{x}_i, y_i, \boldsymbol{\gamma}) = p(m_i|\mathbf{x}_i, \boldsymbol{\gamma})$, and one sees that (2) factorises, so that inferences concerning $\boldsymbol{\beta}$ can be made using only $p(y_i|x_i, \boldsymbol{\beta})$. Here we assume the analyst will base inferences on the complete cases, those units where $m_i = 0$. Under MAR, estimates for $\boldsymbol{\beta}$ are unbiased (Little (1992)). In this paper, we assume that the missing mechanism can be modelled using a logit link function.

Specifically, under MAR,

$$p(m_i = 1|\mathbf{x}_i, \boldsymbol{\gamma}) = \frac{exp(\mathbf{x}_i'\boldsymbol{\gamma})}{1 + exp(\mathbf{x}_i'\boldsymbol{\gamma})}. \tag{3}$$

We denote the expression in (3) by $P(\mathbf{x}_i)$ for short, indicating that it is explicitly dependent on values of $\mathbf{x}_i$. A corresponding NMAR mechanism, which incorporates the (potentially missing) values of the response variable and includes (3) as a special case, is proposed in Section 3.

If the missing mechanism is NMAR, then estimates for $\boldsymbol{\beta}$, based only on $p(y_i|\mathbf{x}_i, \boldsymbol{\beta})$, are biased (including those obtained using a complete case analysis). The presence of NMAR is an untestable assumption, and if it exists there is currently little that can be done to adjust for this, beyond assessing sensitivity of the results to different NMAR mechanisms (Little and Rubin (2002)). In Section 4 we propose a strategy that mitigates the effect NMAR has in finding designs and estimating regression coefficients.

2.2 Optimal design

In experimental design the goal is to choose values of $\mathbf{x}_i$ that optimise a relevant criterion to obtain maximum information from the experiment. Typically, the optimality criterion minimises a function of the covariance matrix of the estimators. We take the estimate of $\boldsymbol{\beta}$ to be

$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y}$, with covariance matrix

$$\boldsymbol{var}\ (\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

We consider designs of the form

$$\xi = \begin{Bmatrix} \mathbf{x}_1^* & \cdots & \mathbf{x}_m^* \\ w_1 & \cdots & w_m \end{Bmatrix}, \ 0 < w_i \leq 1, \ \sum_{i=1}^{m} w_i = 1,$$

where $\mathbf{x}_1^*, \ldots, \mathbf{x}_m^*$ $(m \leq n)$ are the distinct values of the explanatory variables, referred to as the support points of the design, and the weights $w_1, \ldots, w_m$ are the relative proportions of observations taken at the corresponding support points $\mathbf{x}_i^*$, $i = 1, \ldots, m$.

This approach avoids the problem of discrete optimisation and is thus widely used in finding optimal designs for experiments. Since nw_i, $i = 1, \ldots, m$, are not necessarily integer valued, a rounding procedure is applied; see, for example, Pukelsheim, and Rieder (1992).

For an approximate design ξ, the Fisher information matrix for model (1) is

$$\boldsymbol{M}(\xi) = \sum_{i=1}^{m} \boldsymbol{f}(\mathbf{x}_i^*)\boldsymbol{f}^T(\mathbf{x}_i^*)\ w_i$$

where the vector $\boldsymbol{f}^T(\mathbf{x}_i^*)$ is a row in the design matrix $\boldsymbol{X}$ corresponding to $\mathbf{x}_i^*$, and its inverse, $\boldsymbol{M}^{-1}(\xi)$, is proportional to $\boldsymbol{var}\ (\hat{\boldsymbol{\beta}})$.

We consider two optimality criteria: D-optimality: Minimise $|\boldsymbol{M}^{-1}(\xi)|$; A-optimality: Minimise trace$(\boldsymbol{M}^{-1}(\xi))$.

2.3 Optimal design for missing values

When certain values y_i may be missing we can take account of this through the missing data mechanism. Assuming MAR, the Fisher information matrix containing the missing data indicators $\boldsymbol{\mathcal{M}} = \{m_1, m_2, \ldots, m_n\}$ is given, say, by $\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})$ and we have

$$\begin{aligned} \boldsymbol{E}[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})] &= \boldsymbol{E}[\sum_{i=1}^{n} \boldsymbol{f}(\mathbf{x}_i)\boldsymbol{f}^T(\mathbf{x}_i)\ (1 - m_i)] \\ &= \sum_{i=1}^{n} \boldsymbol{f}(\mathbf{x}_i)\boldsymbol{f}^T(\mathbf{x}_i)[1 - P(\mathbf{x}_i)] \\ &= n\sum_{i=1}^{m} \boldsymbol{f}(\mathbf{x}_i^*)\boldsymbol{f}^T(\mathbf{x}_i^*)\ w_i[1 - P(\mathbf{x}_i^*)] \qquad (4) \end{aligned}$$

which is equivalent to $\boldsymbol{M}(\xi)$ if the responses are fully observed. Imhof, Song, and Wong (2002) proposed a general framework where $\boldsymbol{M}(\xi)$ is replaced by (4) in the respective optimality criterion. This assumes that $E\{[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]^{-1}\}$ is proportional to $E[\boldsymbol{var}\ (\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})]$, and may result in a crude approximation to the covariance matrix, in particular for small to moderate sample sizes. Lee, Biedermann, and Mitra (2017) develop an improved approximation by considering the expectation of a 2nd order Taylor expansion of $[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]^{-1}$ which also results in better designs. For large sample sizes, however, the two approaches generate similar designs.

These approaches are implicitly based on assuming MAR. If the potential for NMAR exists then this framework may lead to inefficient designs,

with biased estimates. We first look to determine what effect NMAR might have on the performance of designs found assuming MAR holds, then consider how to best address the problem of NMAR in Section 4. In Sections 5 and 6 we present results that incorporate our findings from Section 4 to find designs and evaluate performance.

3. Effect of NMAR on optimal designs

If we have NMAR when constructing optimal designs then our missing data mechanism implicitly depends on the outcome variable. We consider one such situation and modify the missing data mechanism in (3) to

$$p(m_i = 1|\mathbf{x}_i, y_i, \boldsymbol{\gamma}) = \frac{exp(\mathbf{x}_i'\gamma + \delta y_i)}{1 + exp(\mathbf{x}_i'\gamma + \delta y_i)} \tag{1}$$

for $i = 1, \ldots, n$.

We now illustrate what effect, if any, NMAR might have in the construction of optimal designs and their resulting performance. We focus on the simple linear regression model for the design region $\mathfrak{X} = [0, u]$ for some value $0 < u < \infty$. As such we treat the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \tag{2}$$

for $i = 1, \ldots, n$. Without loss of generality we assume $\delta = 1$ which gives us the missing data mechanism as

$$p(m_i = 1|x_i, y_i, \gamma_0, \gamma_1) = \frac{exp(\gamma_0 + \gamma_1 x_i + y_i)}{1 + exp(\gamma_0 + \gamma_1 x_i + y_i)}. \tag{3}$$

From (2) and (3) it is clear that our design depends on the regression coefficients β_0 and β_1. This can be seen by re-expressing (3) as

$$p(m_i = 1|x_i, y_i, \gamma_0, \gamma_1) = \frac{exp(\gamma_0^* + \gamma_1^* x_i + \epsilon_i)}{1 + exp(\gamma_0^* + \gamma_1^* x_i + \epsilon_i)} \tag{4}$$

where $\gamma_0^* = \gamma_0 + \beta_0$ and $\gamma_1^* = \gamma_1 + \beta_1$. We assume that designs are constructed under some known fixed values of β_0 and β_1. Knowing these values is unrealistic, finding their values is the goal of the experiment. It may be possible that the analyst has some prior information about likely values of β_0 and β_1 that can be used. The resultant designs will be locally optimal. This is a specific complication that arises due to NMAR.

It is not clear what effect, if any, σ^2 has on the efficiency of the design. As ϵ_i has zero mean, it may be the case that this term does not influence the design, but the larger the value of σ^2 the greater the uncertainty about the expected amount of missing data at any given point x_i within the design region $\mathfrak{X}$. This might influence what design we choose.

Let $u = 2$, so the design space is $\mathfrak{X} = [0, 2]$ in what follows. We first find the optimal two-point designs, under D- and A- optimality, assuming $(\gamma_0, \gamma_1, \beta_0, \beta_1) = (-5.572, 2.191, 1, 1)$ and $\sigma^2 = 0$. This is equivalent to setting $\epsilon_i = 0$ in (4) and assumes a MAR mechanism with parameters $(\gamma_0^*, \gamma_1^*) = (-4.572, 3.191)$. With these values we find the probability of missing at the end points of the design space, 0 and 2, are 0.01 and 0.859

respectively and is monotone increasing over the space. Thus the potential for missing data is not too extreme at any point in the design space, but allowing the potential for missing data to have an impact on the performance of any given design. When the missing mechanism is monotone increasing, Lee, Biedermann, and Mitra (2017) show that the lower bound of the design space is always one of the support points in an optimal design. Thus in a two-point design it suffices to find the second support point, x_2^* and its weight w_2, as $w_1 = 1 - w_2$. Using the *fmincon* function in *Matlab*, we find an optimal design of $\{x_1^*, x_2^*; w_1, w_2\} = \{0, 1.3766; 0.5, 0.5\}$ under the D-optimality criterion and $\{x_1^*, x_2^*; w_1, w_2\} = \{0, 1.5147; 0.546, 0.454\}$ under the A-optimality criterion.

For each optimal design, we then simulated $n = 60$ (where $n_1 = nw_1$ and $n_2 = nw_2$ with integer rounding if necessary) observations from (2) using the support points, the values of β_0, β_1 above, and under different σ^2. Some outcome were missing using (3) with the values of these γ_0, γ_1, as well as the simulated y_i values. Estimates of β_0, β_1 were obtained using the complete case data. This process was repeated 100,000 times to obtain measures of bias and mean squared error for β_0 and β_1. We also found the determinant and trace of the variance-covariance and mean squared error matrix that correspond to the objective functions we sought to minimise

under D- and A- optimality.

Table 1 presents the performance of the two optimal designs under different missing data mechanisms and different values of σ^2. The outputs under NMAR correspond to the situations where ϵ_i in (4) has the corresponding σ^2 whereas those under MAR correspond to the situations where ϵ_i in (4) has $\sigma^2 = 0$. In all cases, the responses were simulated with the corresponding values of σ^2. We see that the bias and the mean squared error increase as σ^2 increases. Comparing the two scenarios for the same σ^2, the estimates obtained in the presence of a NMAR mechanism have more bias and larger mean squared errors than those obtained in the presence of the MAR mechanism. We also find a similar profile for the determinant and trace of the covariance and the mean squared error matrix.

Focusing on the bias and the mean squared error of the estimates in the presence of a NMAR mechanism, in Figure 1 we plot how this varies with different values of σ^2 under the D- and A- optimal designs found above. The mean squared error of each estimate increases with the values of σ^2 and the estimates are biased downward when σ^2 is large. Thus σ^2 plays a role in affecting the performance of any design under NMAR. In the next section we investigate how we can take account of the effect of σ^2 in constructing optimal designs.

Table 1: Simulation outputs of A- and D-optimal designs across 100,000 simulated data sets under different missing data mechanisms.

	under NMAR			under MAR		
	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
A-optimal design, $\{x_1^*, x_2^*; n_1, n_2\} = \{0, 1.5147; 33, 27\}$						
bias of $\hat{\beta}_0$	-0.00303	-0.0163	-0.0555	-7.82 $\times 10^{-5}$	-2.34 $\times 10^{-4}$	-3.91 $\times 10^{-4}$
bias of $\hat{\beta}_1$	-0.0855	-0.292	-0.538	2.56 $\times 10^{-4}$	7.69 $\times 10^{-4}$	0.00128
mse of $\hat{\beta}_0$	0.00765	0.0306	0.0701	0.00191	0.0172	0.0478
mse of $\hat{\beta}_1$	0.0198	0.130	0.379	0.00327	0.0294	0.0817
tr(mse)	0.0275	0.161	0.449	0.00518	0.0466	0.130
|mse|	1.29 $\times 10^{-4}$	0.00374	0.0263	4.65 $\times 10^{-6}$	3.77 $\times 10^{-4}$	0.00291
$var(\hat{\beta}_0)$	0.00764	0.0303	0.0670	0.00191	0.0172	0.0478
$var(\hat{\beta}_1)$	0.0125	0.0447	0.0891	0.00327	0.0294	0.0817
$tr(\boldsymbol{var}(\hat{\boldsymbol{\beta}}))$	0.0201	0.0750	0.156	0.00518	0.0466	0.130
$\lvert\boldsymbol{var}(\hat{\boldsymbol{\beta}})\rvert$	7.01 $\times 10^{-5}$	9.53 $\times 10^{-4}$	0.00401	4.65 $\times 10^{-6}$	3.77 $\times 10^{-4}$	0.00291
D-optimal design, $\{x_1^*, x_2^*; n_1, n_2\} = \{0, 1.3766; 30, 30\}$						
bias of $\hat{\beta}_0$	-0.00312	-0.0165	-0.0559	-1.26 $\times 10^{-4}$	-3.79 $\times 10^{-4}$	-6.31 $\times 10^{-4}$
bias of $\hat{\beta}_1$	-0.0761	-0.266	-0.501	2.43 $\times 10^{-4}$	7.29 $\times 10^{-4}$	0.00121
mse of $\hat{\beta}_0$	0.00840	0.0335	0.0766	0.00210	0.0189	0.0525
mse of $\hat{\beta}_1$	0.0180	0.116	0.345	0.00317	0.0285	0.0793
tr(mse)	0.0264	0.150	0.422	0.00527	0.0475	0.132
|mse|	1.17 $\times 10^{-4}$	0.00351	0.0258	4.37 $\times 10^{-6}$	3.54 $\times 10^{-4}$	0.00273
$var(\hat{\beta}_0)$	0.00839	0.0333	0.0735	0.00210	0.0189	0.0525
$var(\hat{\beta}_1)$	0.0123	0.0456	0.0945	0.00317	0.0285	0.0793
$tr(\boldsymbol{var}(\hat{\boldsymbol{\beta}}))$	0.0206	0.0789	0.168	0.00527	0.0475	0.132
$\lvert\boldsymbol{var}(\hat{\boldsymbol{\beta}})\rvert$	6.59 $\times 10^{-5}$	9.36 $\times 10^{-4}$	0.00410	4.37 $\times 10^{-6}$	3.54 $\times 10^{-4}$	0.00273

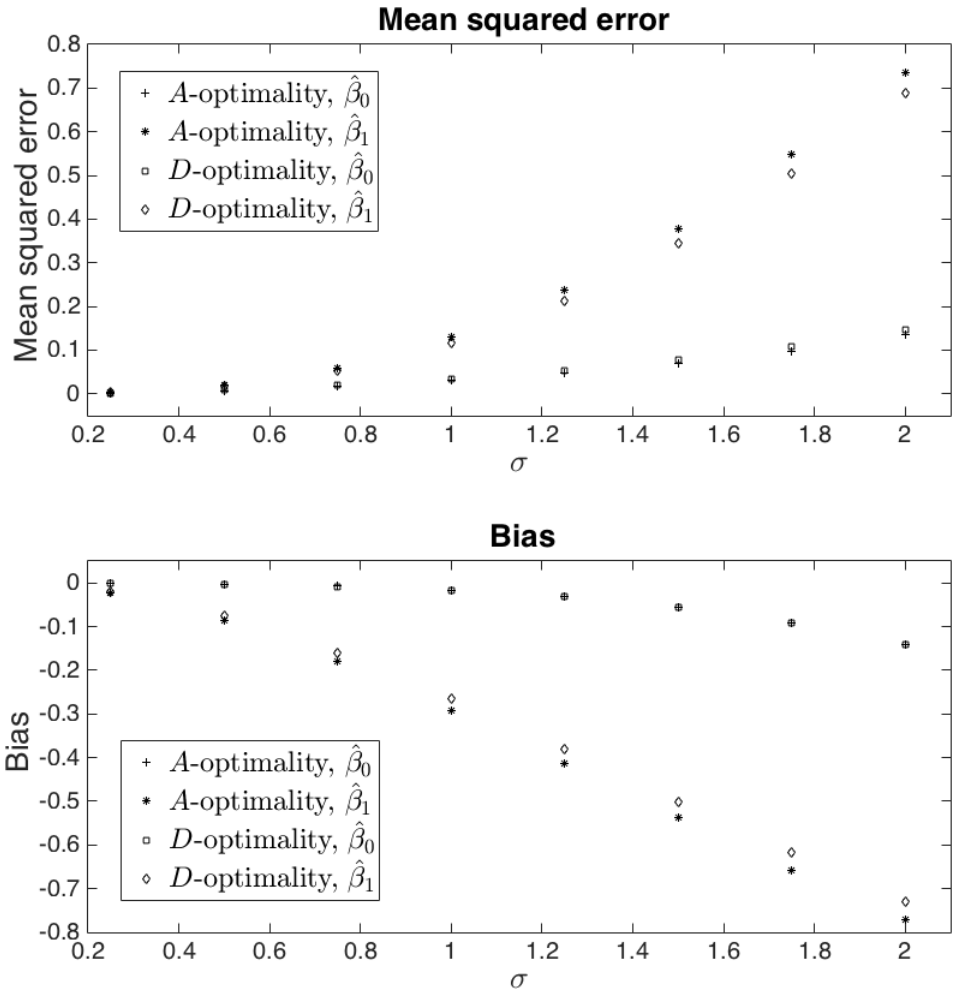

Figure 1: Mean squared error and bias of the estimates that were computed using the A- and the D-optimal design in the presence of NMAR mechanisms.

4. Optimal design under NMAR

We first provide intuition behind why new theory needs to be developed in constructing optimal designs when NMAR is present, then present details concerning our investigation into approximating the missing indicator probability, and consider broadening the framework to include bias into the optimality criterion.

4.1 Incorporating NMAR into the design framework

When missing data are present, we seek to minimise a function of $E\{[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]^{-1}\}$ as this can be viewed as a surrogate for minimising the corresponding function of $E[\boldsymbol{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})]$. Evaluating this expectation is not straightforward and must be approximated. Imhof, Song, and Wong (2002) approximate it by $\{E[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]\}^{-1}$, while Lee, Biedermann, and Mitra (2017) first take a 2nd order Taylor expansion of $[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]^{-1}$ and then take the expectation.

Regardless, both approaches assume MAR, and the expectations involve taking expectations of the missing data indicators $E(m_i) = P(x_i)$ that are then components of the resulting optimality criterion. To account for NMAR when finding optimal designs, we use $P(x_i, y_i)$, where $P(x_i, y_i) = E(m_i|x_i, y_i)$ is now random.

To proceed, we replace $P(x_i, y_i)$ with its expected value $E[P(x_i, y_i)]$

where the expectation is taken with respect to y_i. This expectation is not typically available in closed form and we investigate ways to approximate it in Section 4.2.

A key consideration is the potential for bias. When NMAR is present, estimates are likely to be biased as is evident from the results in Section 3. Optimal design criteria then must incorporate bias, or some approximation to it, to find designs with small MSE. This is discussed in more detail in Section 4.3.

4.2 Evaluating the expectation of $P(x_i, y_i)$

To evaluate the expectation of $P(x_i, y_i)$ we consider the specific example of the NMAR mechanism (3) introduced in Section 3. In principle, the approach would work with any appropriate NMAR missing data mechanism. We can write

$$\begin{aligned} P(x_i, y_i) &= \frac{exp(\gamma_0 + \gamma_1 x_i + y_i)}{1 + exp(\gamma_0 + \gamma_1 x_i + y_i)} \\ &= \frac{exp(z_i)}{1 + exp(z_i)} \end{aligned} \tag{5}$$

where $z_i \sim N(\gamma_0 + \beta_0 + (\gamma_1 + \beta_1)x_i, \sigma^2)$. Thus $exp(z_i)$ has a Log-normal distribution with parameters given by the mean and variance of z_i, and $P(x_i, y_i) = \frac{exp(z_i)}{1+exp(z_i)}$ has a logit-normal distribution with parameters given by the mean and variance of z_i. As the mean of the logit normal distribution

is not available in closed form, we consider approximating the expected value of $P(x_i, y_i)$.

The simplest approach replaces z_i with its expected value in (5),

$$E[P(x_i, y_i)] \approx \frac{exp[E(z_i)]}{1 + exp[E(z_i)]}. \tag{6}$$

This is equivalent to the naive approach of finding an optimal design in Section 3 which assumes MAR, and we see that it does not perform well.

An improved approximation uses the fact that $E[exp(z_i)] = exp[\gamma_0 + \beta_0 + (\gamma_1 + \beta_1)x_i + \sigma^2/2]$, and taking a first order Taylor expansion of $P(x_i, y_i)$ as a function of $exp(z_i)$ about the mean of $exp(z_i)$,

$$\begin{aligned} E[P(x_i, y_i)] &\approx \frac{E[exp(z_i)]}{1 + E[exp(z_i)]} \\ &= \frac{exp[\gamma_0 + \beta_0 + (\gamma_1 + \beta_1)x_i + \sigma^2/2]}{1 + exp[\gamma_0 + \beta_0 + (\gamma_1 + \beta_1)x_i + \sigma^2/2]}, \end{aligned} \tag{7}$$

We also consider approximating the expectation of $P(x_i, y_i)$ using numerical methods. Write $P(x_i, y_i) = t_i$ for simplicity, we use the function *integral* in *Matlab* to evaluate

$$E\left[\frac{exp(z_i)}{1 + exp(z_i)}\right] = E(t_i) = \int_0^1 t_i \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{t_i(1 - t_i)} e^{-\frac{[logit(t_i) - \mu_i]^2}{2\sigma^2}} dt_i \tag{8}$$

We conducted simulation studies to empirically evaluate the performance of these methods for approximating $E[P(x_i, y_i)]$. We generated data that followed a specific logit normal distribution, with parameters μ and

σ, and computed the estimated mean of this distribution using the different approximations. This was repeated many times and estimates from the different methods were averaged over the replications and compared to the "true mean" obtained empirically by averaging the sample mean of observations over the replications. This process was then repeated for a range of different values of μ and σ. Our simulation studies showed that approximations from (6) and (7) performed poorly compared to (8). The approximation given by (8) gives us very small magnitude absolute differences for $-30 \leq \frac{\mu_i}{\sigma} \leq 30$. We also considered approximating the expected value using a second order Taylor expansion about $exp(z_i)$ as well as first and second order Taylor expansions about z_i. We tried using the median of the logit normal distribution implied by $P(x_i, y_i)$, available in closed form, as a surrogate for the expected value. None of them performed as well as the numerical approximation considered, and we use (8) in our design framework going forward.

4.3 Incorporating bias into the design criterion

When responses are not missing at random, estimates will be biased. Hence, instead of simply considering $\boldsymbol{var(\hat{\beta})}$, we consider broadening the framework to incorporate bias. We focus on optimising a function of the mean squared error. Returning to the example of obtained regression coef-

ficient estimates $\hat{\boldsymbol{\beta}}$, the mean squared error incorporates both variance and bias,

$$m.s.e.\ (\hat{\boldsymbol{\beta}}) = \boldsymbol{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\boldsymbol{T}}] = \boldsymbol{var}\ (\hat{\boldsymbol{\beta}}) + \left[\boldsymbol{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}\right]\left[\boldsymbol{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}\right]^{T}. \tag{9}$$

We denote $\boldsymbol{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}$ by $\Delta(\sigma, \xi)$, assuming the bias depends on σ as well as the design. Other more complex bias functions that depend on more parameters could be considered.

To find optimal designs in the presence of NMAR with good MSE properties, we numerically approximated the bias function by simulating it over a range of different pairs of values (σ, ξ). Each simulation step involved fitting the model and evaluating the bias for the given pair. We then fit a smooth function, e.g. a second order response surface or a LOESS function, to these simulated 'bias data', and used this function, $B(\sigma, \xi)$ say, as an approximation to the true bias.

In the next section we evaluate how the approach of finding optimal designs based on the approximation given in (8) and the inclusion of a bias term performs in the presence of NMAR, and whether it offers any improvements over the optimal designs that assume MAR.

5. Simulation study

We set the design region $\mathfrak{X} = [0, 2]$ and sample size $n = 60$. For a given

design we simulated a response variable as

$$y_i = 1 + x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

for a given σ^2. We then introduced missing values into the observed y_i, $i = 1, \ldots, n$, through the logistic model

$$P(x_i, y_i) = \frac{exp(\gamma_0 + \gamma_1 x_i + y_i)}{1 + exp(\gamma_0 + \gamma_1 x_i + y_i)}$$

with $\gamma_0 = -5.572$ and $\gamma_1 = 2.191$. We fit a simple linear regression model to the complete case data, obtaining estimates of the coefficients, $(\hat{\beta}_0, \hat{\beta}_1)$, and their variances, from the cases for which y_i was observed.

We restricted our optimal designs to the class of designs with two support points. From Lee, Biedermann, and Mitra (2017), the lower bound of $\mathfrak{X}$, 0, was chosen as one of the support points, x_1^*. To find the second support point, x_2^*, we substituted the approximation to $E[P(x_i^*, y_i)]$ given by (8) with mean $-5.572 + 1 + (2.191 + 1)x_i^*$ and a known value of σ^2, the value of x_1^* and $w_1 = 1 - w_2$ into the mean squared error given in (9). The expected bias term in (9) was treated as being a function of x_2^* and σ^2, and was approximated numerically. An optimal design was then found by minimising a function of this matrix with respect to x_2^* and w_2 in *Matlab* with the *fmincon* function.

Table 2 presents the values of x_2^* under the D- and A- optimality criteria

Table 2: The first column from the left shows the optimal designs that assume a MAR mechanism (6); the other columns show the optimal designs for NMAR mechanisms (5) with different σ^2. In all designs, $x_1^* = 0$, $n = 60$ and $w_1 = 1 - w_2$.

		MAR	σ =1	σ =1.5	σ =2
D-optimal	x_2^*	1.3766	0.9793	1.0202	1.1210
design	$w_2(n_2)$	0.5000(30)	0.3811 (23)	0.3194 (19)	0.2879 (17)
A-optimal	x_2^*	1.5147	1.0871	1.0617	1.0671
design	$w_2(n_2)$	0.4539(27)	0.4462 (27)	0.4508 (27)	0.4534 (27)

for various different values of σ^2, the corresponding weight w_2, and the (rounded) number of replicates, n_2, of x_2^*. The optimal designs that account for the impact of NMAR have smaller x_2^* for both design criteria than those that assume the presence of a MAR mechanism. The optimal weights of A-optimal designs remain constant in the considered cases whereas w_2 of the D-optimal design decreases with σ^2 when responses are assumed to be NMAR. Figure 2 further illustrates the optimal designs that account for the impact of NMAR.

To illustrate the performance of these designs we repeatedly simulated an incomplete data set 200,000 times, using each of the designs given in Table 2 and the models for the response and the missing data mechanism.

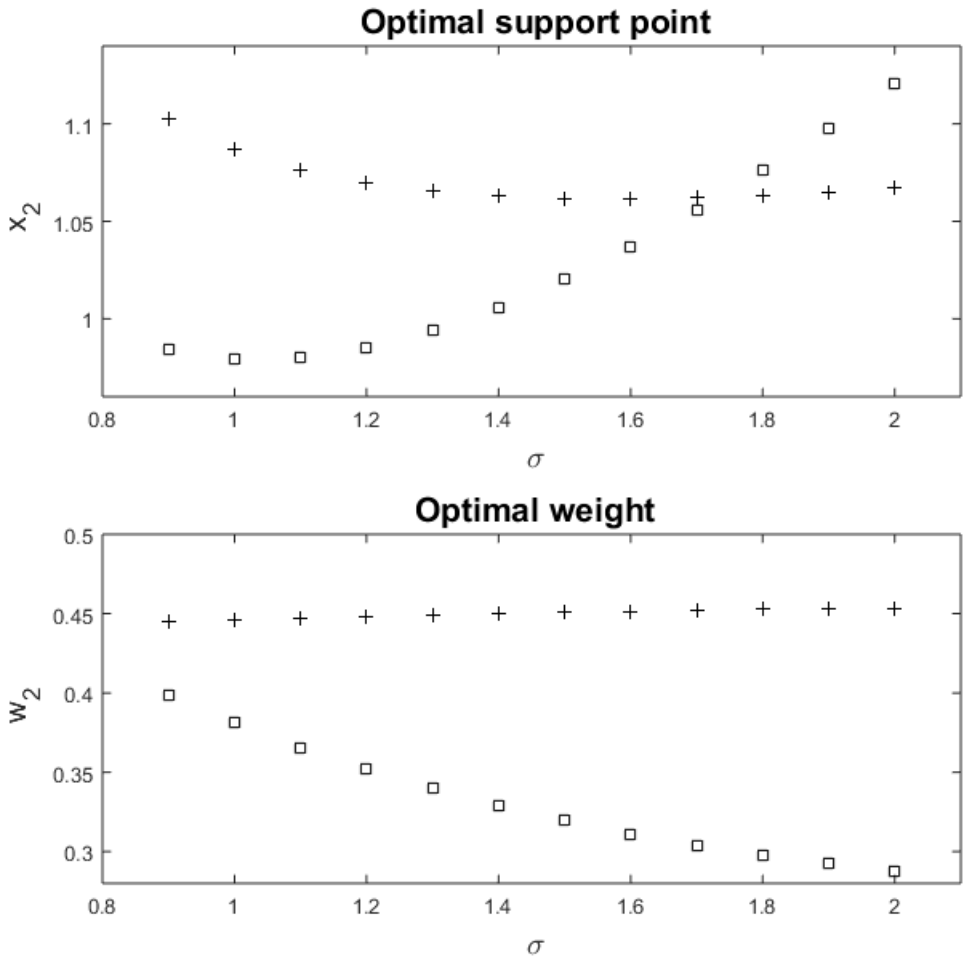

Figure 2: "+" correspond to A-optimal designs, "□" correspond to D-optimal designs in the presence of different NMAR mechanisms with $x_1 = 0$ and $w_1 = 1 - w_2$.

For each design, we calculated the empirical bias and the mean squared error for β_0 and β_1, as well as the determinant and trace of the empirical mean squared error matrix for (β_0, β_1). Table 3 presents these results for various different values of σ.

The designs that assume the presence of MAR have the largest biases and $m.s.e.$ $(\hat{\boldsymbol{\beta}})$ across the board. By taking NMAR into account at the design stage, we can mitigate some of its effects. For example, the A-optimal design for $\sigma = 1.5$ reduces the bias of $\hat{\beta}_1$ by more than 23% from -0.53864 to -0.41095, and a similar reduction applies to the trace of $m.s.e.$ $(\hat{\boldsymbol{\beta}})$. The NMAR design with the conjectured value of σ performs best with respect to the relevant optimality criterion, and the NMAR designs with different conjectured values of σ also perform well, far better than the designs that assume MAR.

We consider the problem of assuming the presence of NMAR when in fact a MAR assumption is reasonable. We evaluated the performance of the designs given in Table 2 when the missing mechanism was in fact MAR, with

$$P(x_i) = \frac{exp(\gamma_0 + \gamma_1 x_i)}{1 + exp(\gamma_0 + \gamma_1 x_i)},$$

where $\gamma_0 = -4.572$ and $\gamma_1 = 3.191$. The performance metrics considered were empirical bias and mean squared error for β_0 and β_1, as well as the de-

Table 3: Performance of various designs in the presence of NMAR mechanism over 200,000 simulated data sets.

$\sigma^2 = 1$ in generating y_i and in the NMAR mechanism								
	D-optimal design that assumes				*A*-optimal design that assumes			
	MAR	σ =1	σ =1.5	σ =2	MAR	σ =1	σ =1.5	σ =2
bias of $\hat{\beta}_0$	-0.015710	-0.015657	-0.015559	-0.015525	-0.015717	-0.015717	-0.015717	-0.015717
bias of $\hat{\beta}_1$	-0.26664	-0.18472	-0.19344	-0.21511	-0.29240	-0.20739	-0.20208	-0.20313
$m.s.e.$ $(\hat{\beta}_0)$	0.033581	0.027279	0.024665	0.023522	0.030604	0.030604	0.030604	0.030604
$m.s.e.$ $(\hat{\beta}_1)$	0.11689	0.11449	0.12077	0.12403	0.13022	0.10697	0.10728	0.10713
$tr(m.s.e.\ (\hat{\boldsymbol{\beta}}))$	0.15047	0.14176	0.14544	0.14756	0.16083	**0.13758**	0.13788	0.13774
$\lvert m.s.e.\ (\hat{\boldsymbol{\beta}})\rvert$	0.0035232	**0.0025149**	0.0025445	0.0026165	0.0037448	0.0026704	0.0026408	0.0026451
$\sigma^2 = 1.5^2$ in generating y_i and in the NMAR mechanism								
	D-optimal design that assumes				*A*-optimal design that assumes			
	MAR	σ =1	σ =1.5	σ =2	MAR	σ =1	σ =1.5	σ =2
bias of $\hat{\beta}_0$	-0.054443	-0.054393	-0.054202	-0.054178	-0.054465	-0.054465	-0.054465	-0.054465
bias of $\hat{\beta}_1$	-0.50182	-0.38675	-0.39934	-0.42936	-0.53864	-0.41838	-0.41095	-0.41264
$m.s.e.$ $(\hat{\beta}_0)$	0.076555	0.062639	0.056827	0.054331	0.070012	0.070012	0.070012	0.070012
$m.s.e.$ $(\hat{\beta}_1)$	0.34630	0.32185	0.33703	0.34929	0.37910	0.31198	0.31145	0.31162
$tr(m.s.e.\ (\hat{\boldsymbol{\beta}}))$	0.42285	0.38449	0.39386	0.40362	0.44912	0.38199	**0.38146**	0.38163
$\lvert m.s.e.\ (\hat{\boldsymbol{\beta}})\rvert$	0.025828	0.018580	**0.018181**	0.018456	0.026319	0.020325	0.020139	0.020183

terminant and trace of the empirical mean squared error matrix for (β_0, β_1).

Table 4 presents these results for MAR optimal designs and different NMAR optimal designs constructed assuming various values of σ^2. In this simulation, we used a residual variance of $\sigma^2 = 1.5^2$ in generating the responses under each different design. The empirical biases are negligible, as expected. We thus focus on the mean squared errors. The designs generated assuming MAR perform best but there is evidence to suggest that the loss in assuming a positive value of σ is less severe than the one incurred when using the MAR design for NMAR data.

6. Case study: Two-group A-optimal design for Alzheimer's Disease Trial

As an application, we used data from an Alzheimer's disease study that investigated the benefits of administering donepezil, memantine, and the combination of the two, to patients over a period of 52 weeks, on various quality of life measures. See Howard et al. (2012) for full details of the study. We only considered the experimental units in the placebo group and the donepezil-memantine treatment group that were included in the primary intention-to-treat sample. The sample size in each group (n_1, n_2) is 72, resulting in a total sample size of 144. Here we treat the rate of change of the primary outcome measure, SMMSE score (higher score indicates better

Table 4: Performance of various designs in the presence of a MAR mechanism, i.e. NMAR with $\sigma^2 = 0$. Responses y_i are generated with $\sigma^2 = 1.5^2$, and over 200,000 simulated data sets.

	D-optimal design that assumes			
	MAR	σ =1	σ =1.5	σ =2
bias of $\hat{\beta}_0$ ($10^{-4}\times$)	3.7083	4.0648	5.8203	6.2709
bias of $\hat{\beta}_1$ ($10^{-4}\times$)	4.4560	2.9727	-5.9871	-8.3833
$m.s.e.$ ($\hat{\beta}_0$)	0.075687	0.061415	0.055479	0.052892
$m.s.e.$ ($\hat{\beta}_1$)	0.11455	0.19076	0.19898	0.18913
$tr(m.s.e.$ $(\hat{\boldsymbol{\beta}}))$	0.19024	0.25218	0.25446	0.24202
$\lvert m.s.e.$ $(\hat{\boldsymbol{\beta}})\rvert$	0.0056653	0.0078116	0.0081087	0.0077921
	A-optimal design that assumes			
	MAR	σ =1	σ =1.5	σ =2
bias of $\hat{\beta}_0$ ($10^{-4}\times$)	3.3924	3.3924	3.3924	3.3924
bias of $\hat{\beta}_1$ ($10^{-4}\times$)	2.4240	2.4140	3.2951	3.3618
$m.s.e.$ ($\hat{\beta}_0$)	0.068937	0.068937	0.068937	0.068937
$m.s.e.$ ($\hat{\beta}_1$)	0.11793	0.15299	0.15843	0.15727
$tr(m.s.e.$ $(\hat{\boldsymbol{\beta}}))$	0.18687	0.22192	0.22736	0.22621
$\lvert m.s.e.$ $(\hat{\boldsymbol{\beta}})\rvert$	0.0060607	0.0065411	0.0067209	0.0066843

cognitive function), as the response variable in a simple linear model,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

where $x_i = 0$ if subject i is in the placebo group and $x_i = 1$ if subject i is in the treatment group, $i = 1, \ldots, 144$.

From the data set for the per-protocol analysis, we found 46 patients in the placebo group and 23 patients in the treatment group who had missing responses by the end of the study. Assuming that these responses were not missing at random, a logistic regression model was fit to the missing data indicator, obtaining

$$\frac{exp(\hat{\gamma}_0 + \hat{\gamma}_1 x_i)}{1 + exp(\hat{\gamma}_0 + \hat{\gamma}_1 x_i)}$$

where $\hat{\gamma}_0 = 0.5705$ and $\hat{\gamma}_1 = -1.3269$. Using the observed responses, we fit a linear model to the data, obtaining $\hat{\beta}_0 = -0.10503$, $\hat{\beta}_1 = 0.04302$ and $\sigma^2 = 0.06143^2$. We then used these estimates to construct a NMAR mechanism, (5), where the logit-normal variable, t_i, had mean $\gamma_0 + \beta_0 + (\gamma_1 + \beta_1)x_i = 0.5705 - 0.10503 - (1.3269 - 0.04302)x_i$ and variance $\sigma^2 = 0.06143^2$. We used this information in (6) to approximate the expected NMAR mechanism, present in the elements of the approximation to $E[\boldsymbol{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})]$ when finding optimal designs. In practice NMAR is an untestable assumption and there is no guarantee that such a conjectured mechanism corresponds to the true missing mechanism.

Table 5: Fitted coefficients for the approximation function $B(\sigma, \xi)$ of $\Delta(\sigma, \xi)$ for $\hat{\beta}_0$ (first row) and $\hat{\beta}_1$ (second row), respectively.

$\hat{\lambda}_0$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$
-2.5282×10^{-5}	1.8727×10^{-6}	-1.2511×10^{-3}	-1.4028×10^{-8}	8.7490×10^{-7}	-0.6023
-2.9306×10^{-5}	-4.1954×10^{-7}	1.6884×10^{-3}	2.9213×10^{-9}	3.1693×10^{-6}	0.2919

The support points of an optimal design are given as $x_1^* = 0$ (placebo) and $x_2^* = 1$ (active treatment) since we are comparing two groups. We considered A-optimality with $m.s.e.\ (\hat{\beta}_0) + m.s.e.\ (\hat{\beta}_1)$. The optimisation problem is now in one variable, w_2, with the condition $w_1 + w_2 = 1$.

We conducted simulation studies on designs that had $n_2 = 37, 38, ..., 107$ in each design, with $\sigma = 0.04,\ 0.05, ...,\ 0.09$ in each case, to obtain empirical biases for $\hat{\beta}_0$ and $\hat{\beta}_1$. Fitting a second order response surface to these observed biases and values of n_2 and σ, we approximated bias as a function of n_2 and σ, as

$$B(\sigma, \xi) = \hat{\lambda}_0 + \hat{\lambda}_1 n_2 + \hat{\lambda}_2 \sigma + \hat{\lambda}_3 {n_2}^2 + \hat{\lambda}_4 n_2 \sigma + \hat{\lambda}_5 \sigma^2$$

for each estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ (see Table 5).

Using this information, we found the A-optimal design by using the *fmincon* numerical method in *Matlab*. The optimal design resulted in $w_2 = 0.34365$, or $n_2 = 144 \times w_2 = 49.486 = 49$ subjects in the treatment

Table 6: Performance of various designs where n_2 is the sample size of the treatment group and $n_1 = 144 - n_2$ for each design.

n_2	52	51	50	49	72
$tr(m.s.e.\ (\hat{\boldsymbol{\beta}}))(\times 10^{-4})$	3.2950	3.2927	3.2934	**3.2919**	3.6155

group, with $n_1 = 95$ subjects in the placebo group. We then conducted a simulation study comparing this design with other design candidates using the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}_0$, $\hat{\gamma}_1$ and $\hat{\sigma}^2$ in generating responses (both observed and missing). Table 6 shows the performance of these designs in the simulation. We repeatedly simulated incomplete data under the various designs and computed the trace of the mean squared error matrix obtained from each design. The simulation study shows that the A-optimal design that accounts for NMAR and bias in the experiment performs better than all other designs considered, and in particular is better than the original design that assumes equal sample size for both groups. There is about a 9% $(1 - 3.2919/3.6155) \times 100\%$ efficiency loss if we use the equal sample size design instead of the optimal design. This indicates that there is the potential for obtaining estimates with smaller mean squared error if the proposed design is used rather than conventional designs.

7. Discussion and remarks

There are many open problems left to investigate. A similar approximation to (7) can be found for nonlinear models with normally distributed errors, and extensions to generalised linear models are also possible in our framework.

The designs we find are locally optimal in the sense that they depend on the unknown model parameters. Our numerical investigations show that, even when the value of σ^2 is misspecified at the design stage, the designs assuming NMAR with an incorrect σ^2 perform still better than the MAR design when the missing data mechanism is NMAR. For the other parameters, we assume that good information can be elicited from the experimenter. If this is not the case, parameter robust design criteria, such as Bayesian or standardised maximin criteria (see, e.g., Chaloner and Verdinelli (1995), and Dette (1997)), need to be developed for our approach.

There is a plethora of possible methods to handle the problem of missing values, in addition to complete case analysis considered here. Other common approaches include multiple imputation, methods based on the EM algorithm, Hot Deck methods, and more. We do not investigate these here, as our approach focuses on the design aspect of the problem, rather than the specific method for dealing with the missing data. It would be

interesting to investigate whether the benefits seen here could be similarly observed when other methods are used to handle the missing data.

Acknowledgement

The first author's research was funded by the Institute for Life Sciences at the University of Southampton. We would like to acknowledge Clive Holmes; Robert Howard and Patrick Philips for supplying us with the data from the Domino study RCTN49545035 which was funded by the MRC and Alzheimer's Society UK. The second author would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Data Linkage and Anonymisation where work on this paper was undertaken. This work was supported by EPSRC grant no EP/K032208/1.

References

Ahmad, T., and Gilmour, S. G. (2010). Robustness of subset response surface designs to missing observations. *Journal of Statistical Planning and Inference* **140**, 92-103.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science* **10**, 273-304

Dette, H. (1997). Designing experiments with respect to standardized

optimality criteria. *J. Roy. Statist. Soc. Ser. B* **59**, 97-110.

Ghosh, S. (1979). On robustness of designs against incomplete data. Sankhyā: *The Indian Journal of Statistics, Series B* 204-208.

Hackl, P. (1995). Optimal design for experiments with potentially failing trials. In *Proc. of MODA4: Advances in Model-Oriented Data Analysis* (Edited by C. P. Kitsos and W. G. Müller) 117-124. Physica Verlag, Heidelberg.

Hedayat, A. and John, P. W. M. (1974). Resistant and susceptible BIB designs. *Ann. Statist. 2* **1**, 148–158.

Herzberg, A. M. and Andrews, D. F. (1976). Some considerations in the optimal design of experiments in non-optimal situations. *J. Roy. Statist. Soc. Ser. B* **38**, 284-289.

Howard, R., McShane, R., Lindesay, J., Ritchie, C., Baldwin, A., Barber, R., ... and Phillips, P. (2012). Donepezil and memantine for moderate-to-severe Alzheimer's disease. *New England Journal of Medicine* **366**, 893-903.

Imhof, L. A and Song, D. and Wong, W. K. (2002). Optimal design of experiments with possibly failing trials. *Statistica Sinica* **12**, 1145-

1155.

Lee, K.M., Biedermann, S. and Mitra, R. (2017). Optimal design for experiments with possibly incomplete observations. *Statistica Sinica*.

Little, R. J. A. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association* **87**, 1227-1237.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data (Second Edition)*. Wiley-Interscience.

Ortega-Azurduy, S. A., Tan, F. E. S. and Berger, M. P. F. (2008). The effect of dropout on the efficiency of D-optimal designs of linear mixed models. *Statist. Med.* **27**, 2601-2617.

Pukelsheim, F. and Rieder, S. (1992). Efficient rounding of approximate designs. *Biometrika* **79**, 763-770.

Rubin, D.B (1976). Inference and missing data. *Biometrika* **63**, 581-592.