Supplementary Material: Principles of Experimental Design for Big Data Analysis

Christopher C Drovandi, Christopher Holmes, James M McGree,
Kerrie Mengersen, Sylvia Richardson and Elizabeth G Ryan

November 16, 2016

Appendix A: Experimental Design Overview

Statistical experimental design, which here also includes sampling design, provides rules for the allocation of resources in a data collection exercise. In a decision theoretic framework, the design should take into consideration the aim of the experiment and the corresponding cost function based on the defined experimental units, allowable errors, resource and implementation constraints, potential sources of heterogeneity and possible biases and other quality concerns. Optimal designs aim to meet the experimental aim with the least resources, which are typically measured in terms of sample size and cost. The optimality of an experimental design is assessed via a utility function or a design criterion, which incorporates the experimental aims and is specific to the design problem.

In the classical framework, optimal experimental designs are commonly derived using optimality criteria that are based on the expected Fisher information matrix (e.g., Fedorov (1972); Pukelsheim and Torsney (1991); Atkinson and Donev (1992)). Thus a design $\mathbf{d}$ may be optimal if it maximises the utility function $U(\cdot)$ over the design space D:

$$\mathbf{d}^* = \arg\max_{\mathbf{d}\in\mathsf{D}} \, U(\mathbf{d}, \boldsymbol{\theta}), \tag{1}$$

where $\mathbf{d}^*$ is the optimal design and $\boldsymbol{\theta} \in \Theta$ is the parameter of interest and is fixed. The design $\mathbf{d}$ often represents the chosen values of some covariates of interest, which is what we consider in the main paper. Analytical solutions to equation (1) can only be found in a small number of design problems, and in most cases, numerical or stochastic search algorithms are required to find the optimal design.

Pseudo-Bayesian designs are also typically derived by averaging $U(\mathbf{d}, \boldsymbol{\theta})$ over a "prior" $p(\boldsymbol{\theta})$ to account for parameter uncertainty (e.g., Pronzato and Walter (1985)). The word "pseudo" refers to the fact that the utility may be a function of the likelihood rather than a full Bayesian posterior. Such approaches are particularly valuable for nonlinear models in that the prior can be used to overcome the dependence of the design on the parameter values. The optimal pseudo-Bayesian design $\mathbf{d}^*$ thus maximises the expected utility function $U(\mathbf{d})$ over the design space D with respect to the model parameter $\boldsymbol{\theta}$:

$$\begin{aligned}\mathbf{d}^* &= \arg\max_{\mathbf{d}\in\mathsf{D}} \; E_{\Theta}\{U(\mathbf{d},\boldsymbol{\theta})\} \\ &= \arg\max_{\mathbf{d}\in\mathsf{D}} \; \int_{\Theta} U(\mathbf{d},\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \qquad (2)\end{aligned}$$

Numerical techniques are often required to solve the expectation and maximisation problem in equation (2), which are generally computationally intensive since the integral needs to be solved at each iteration of the search (see Pronzato and Walter (1985); Pronzato and Zhigljavsky (2012)). Laplace approximations have been used (e.g., Dodds et al. (2005)) to reduce the computational burden by forming a tractable approximation to the expectation.

Fully Bayesian methods for optimal experimental design (Chaloner and Verdinelli, 1995) have become more prominent in the literature in recent years (e.g., Amzal et al. (2006), Cook et al. (2008), Han and Chaloner (2004), Huan and Marzouk (2013), Müller et al. (2006)). Also see a recent review of computational algorithms for Bayesian design given by Ryan et al. (2015). Bayesian optimal design involves defining a utility function $U(\mathbf{d},\boldsymbol{\theta},\mathbf{y})$ that describes the worth (based on the experimental aims) of choosing the design $\mathbf{d}$ from the design space D yielding data $\mathbf{y}$, with model parameter value $\boldsymbol{\theta}$. Bayesian design criteria involve functionals of the posterior distribution, and are often based upon the expected gain in Shannon information from the prior to the posterior distribution (also known as the 'mutual information' or the expected 'Kullback-Leibler distance') (e.g., Chaloner and Verdinelli (1995)). Bayesian design criteria are also commonly based on the spread of the posterior distribution, which may be measured, for example, by the precision or entropy (e.g., Stroud et al. (2001)). A probabilistic model, $p(\boldsymbol{\theta},\mathbf{y}|\mathbf{d})$, is also required. This consists of a likelihood $p(\mathbf{y}|\mathbf{d},\boldsymbol{\theta})$ for observing a new set of measurements $\mathbf{y}$ at the design points $\mathbf{d}$, given parameter value $\boldsymbol{\theta}$, and a prior distribution $p(\boldsymbol{\theta})$ for the parameter $\boldsymbol{\theta}$.

The Bayesian optimal design, $\mathbf{d}^*$, maximises the expected utility function $U(\mathbf{d})$ over the design space D with respect to the future data $\mathbf{y}$ and model parameter $\boldsymbol{\theta}$:

$$\begin{aligned}\mathbf{d}^* &= \arg\max_{\mathbf{d}\in\mathsf{D}} \; E\{U(\mathbf{d},\boldsymbol{\theta},\mathbf{y})\} \\ &= \arg\max_{\mathbf{d}\in\mathsf{D}} \; \int_{\mathsf{Y}}\int_{\Theta} U(\mathbf{d},\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta},\mathbf{y}|\mathbf{d})d\boldsymbol{\theta}d\mathbf{y}. \qquad (3)\end{aligned}$$

The integration is performed over the sample space Y of the data, and the parameter space Θ. Unless the utility function, likelihood and prior are specifically chosen to enable analytic evaluation of the integration problem, equation (3) does not usually have a closed form solution. Therefore, numerical approximations or stochastic solution methods are required to solve the maximisation and integration problem.

Experimental designs are often divided into two groups: static and adaptive (or sequential). For the former, the same design is used throughout the experimental process, regardless of the incoming information that is collected from the experiment. These static designs are useful for experiments in which data are collected in a batch, according to a fixed protocol, or when it is not time-feasible to collect data sequentially. Adaptive or sequential experimental design problems are those that involve an alternating sequence of decisions and observations. Instead

	Coef	95% CI	SE	p
β_0	-1.57	$(-1.58, -1.57)$	0.0014	< 0.0001
β_1	0.49	$(0.48, 0.49)$	0.0014	< 0.0001
β_2	0.15	$(0.145, 0.149)$	0.0011	< 0.0001
β_3	0.09	$(0.09, 0.10)$	0.0038	< 0.0001
β_4	-0.02	$(-0.02, -0.01)$	0.0026	< 0.0001

Table 1: A summary of results obtained from analysing the full airlines dataset.

of using the same design throughout the experimental process, as in static design problems, the design which maximises the expected utility is chosen at each stage of experimentation, based on the outcomes of previous experiments. The Bayesian paradigm is useful for adaptive design problems since the posterior can be used as the prior distribution for the next experiment.

In some situations one may not be able to sample at specific design points, so “design windows” or “sampling windows” may instead be required. These consist of a range of near optimal designs and represent regions of planned sub-optimality. Sampling windows have been used for the design of population pharmacokinetic studies (e.g., Ogungbenro and Aarons (2007); Duffull et al. (2012)), which consisted of specific sampling time intervals. These were found to be more clinically relevant and were generally preferred since unplanned sub-optimality inevitably occurs when an attempt is made to implement the optimal design at its fixed time points. The sampling window designs provided flexibility in the collection of the samples and also provided satisfactory parameter estimation.

Appendix B: Case Study - On-Time Airline Arrivals

Ensuring on-time arrivals of flights is a key performance indicator for almost all airlines and airports (Wu and Mengersen, 2013). In addition to incurring a substantial financial penalty, failure to arrive and depart within a specific time window induces a range of operational, social and economic costs. As such, this study focuses on modelling on-time performance of flights in the USA, based on a dataset extracted from the public site http://stat-computing.org/dataexpo/2009/the-data.html, which comprises over 160 million records for the years 1987 to present. These data have been used by a range of authors to illustrate methods for analysing Big Data; see for example, Wang et al. (2015). We follow these authors by creating a binary response variable denoting late arrival, defined as a plane arriving more than 15 minutes after the scheduled arrival time ($y = 1$) or not ($y = 0$). Two binary covariates were considered, namely departure time ($x_3 = 1$ if departure occurred between 8pm and 5am and 0 otherwise) and weekend ($x_4 = 1$ if departure occurred on weekends and 0 otherwise), and two continuous covariates departure hour (x_1, DepHour, range 0 to 24) and distance from origin to destination (x_2 in 1000 miles). For the purposes of illustration, we focus on the year 1995 which has a total of 5,229,619 observations. For interest, the results which were obtained from analysing the full dataset based on the model given by Wang et al. (2015) are shown in Table 1. These results were obtained by fitting a full main effects logistic regression model using maximum likelihood.

Covariate	Levels
x_1	$-2.5, -2, -1, -0.5, 0, 0.5, 1, 2$
x_2	$-1, -0.5, 0, 0.5, 1, 2, 3, 4$
x_3	$0, 1$
x_4	$0, 1$

Table 2: The levels of each covariate available for selection in the sequential design process.

From Table 1, the probability of a late departure increases with departure hour and miles from the destination. Further, departures between 8pm and 5am have a greater chance of a late departure, while weekends generally have fewer late departures. We note that x_1 and x_2 were scaled to have a mean of 0 and a variance of 1.

In order to undertake Bayesian design in this context, prior information needs to be constructed. As we assume no prior knowledge about the influence of covariates on late arrivals was available, we employ an initial learning phase where data are extracted from the Big Data so that relatively informative priors can be considered. The design proposed for this initial learning phase is a random selection of 10,000 data points. In terms of building a model, Wang et al. (2015) proposed that the full main effects model was appropriate for analysis. Here, we assume it is of interest to determine whether a more complex model is preferred (for the data collected in 1995). This more complex model can take a variety of forms, however, we chose the full main effects model plus a quadratic term for x_1 so that we might be able to determine if curvature exists between x_1 and the log odds of late departure. This covariate was specifically chosen as it represents departure hour, and seems doubtful that a linear relationship with the log odds would be maintained for all departure hours (particularly as departure hour is across all 24 hours in a day). Thus, in this work, two models will be considered. To construct prior distributions about the parameter values, multivariate normal distributions were constructed based on maximum likelihood fits to the extracted data. That is, for each model, the maximum likelihood estimate (MLE) was chosen as the mean and the inverse of the observed Fisher information matrix was used as the variance-covariance matrix. The main reason for forming parametric priors here is to avoid keeping a (potentially large) record of the extracted data throughout the sequential design process.

Following the initial learning phase, our sequential design process was run for the extraction of an additional 2,000 data points to explore what our methods might reveal. For this study, we consider the two models outlined above within the sequential design. To determine if a more complex model is needed to appropriately describe the data, the mutual information utility for model discrimination of Drovandi et al. (2014) was used. The designs available for selection at each iteration of our sequential design process are all combinations of the covariate levels shown in Table 2 resulting in a total of 256 choices. To determine which design was optimal, we implemented an exhaustive search of all potential designs.

Once an optimal design was located, the airline dataset was subsetted based on the combination of x_3 and x_4. Then, the Euclidean distance from the optimal design and each design point in this reduced dataset was calculated. The design which yielded the smallest Euclidean distance from the optimal was extracted. In this way, optimal values for x_3 and x_4 were always equal to the corresponding values in the extracted design. The only discrepancies that may appear will therefore be for x_1 and x_2.

Figure 1 shows the posterior model probabilities for the model derived by Wang et al. (2015) and our more complex model throughout the design process. Once all 2,000 observations have been extracted, we are quite certain that the more complex model is preferred for analysis. Given this, it may be difficult to properly interpret the results from fitting only the main effects model.

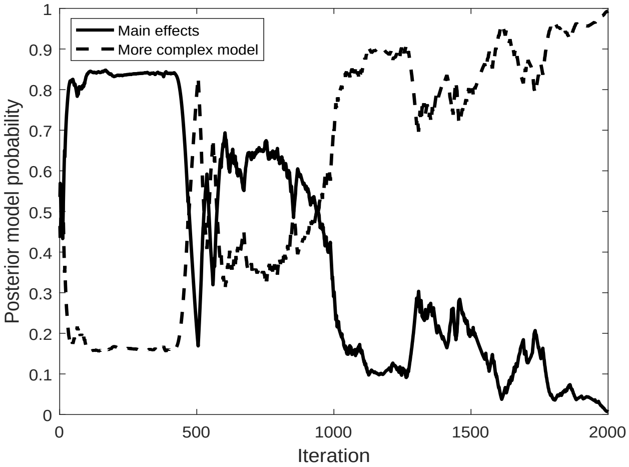


Figure 1: Posterior model probabilities for the model given by Wang et al. (2015) ('-') and a more complex model ('- -') for the airline example.

Figure 2 shows the optimal designs against the actual designs extracted from the Big Data. Ideally, there would be a one-to-one relationship in each plot. This ideal situation certainly appears reasonable for x_2. However, there are noticeable variations for x_1. For example, when a departure hour of 18 was selected as optimal, departure hours near 22 were selected. This suggests that, despite there being over 5 million data records in this Big Data, there is a potential lack of data at certain departure hours. Fortunately, given the results in Figure 1, this potential lack of data did not appear to greatly reduce our ability to determine that a more complex model is needed.

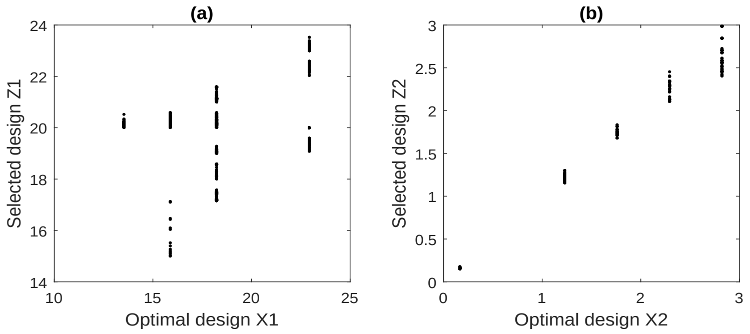


Figure 2: Optimal design versus selected design points for (a) departure hour and (b) miles from destination.

In summary, by considering a small fraction of the Big Data collected on late arrivals of airline flights, we were able to determine that a more complex model than what has been proposed in the literature is required for data analysis. Further we were able to identify that additional data may be required in specific areas of the covariate space in order to efficiently answer model choice problems like the one considered here.

The importance of obtaining an appropriate model or set of models for consideration in our sequential design process cannot be understated. In fact, we do not advocate that our more complex model is the most appropriate for analysis as our aim was just to determine if a more complex model was required. It is worth noting that we could use our designed approach in order to identify inadequacies in terms of model fit with a small amount of data. As shown here, an approach that considers efficient design for models that include more complex terms should help reveal the poor fit of the model with only, say, linear terms. Moreover, we could similarly rule out additionally more complex models through iteratively considering such models.

Appendix C: Colorectal Cancer Case Study

Here we will demonstrate our approach on a data set of colorectal cancer patients in Queensland, Australia obtained from a population-based cancer registry administered by the Cancer Council Queensland. The data set used for this study contains 26182 observations that consists of a 10-year 'at-risk' period of data (cases could be diagnosed in the 2 years preceding the start of the at-risk period), and censoring occurs after 5 years (or at the end of the at-risk period, whichever comes first). The data set contains information relating to the patients' age, sex, whether they were censored or died, their survival time, and their hazard or risk of death (given their age, sex, and the time period). The risk of death is calculated from life tables using population data and mortality data.

For this example, we are interested in determining whether there are differences in the risk of death between the different age groups and genders. A two-way ANOVA model (with both main effects and interactions) will be used to model the data and answer our question of interest. If one was interested in modelling the survival time or the relative survival of the cancer patients, then more complex models, such as flexible parametric survival models (e.g., Nelson et al. (2007); Royston and Lambert (2011)) could be used.

A full factorial ANOVA design (balanced design) is proposed to determine whether any differences exist in the risk of death between the different age groups and genders. As a proof-of-concept of our design approach for Big Data, we investigate analysing a small fraction, say less than 5%, of the full data set to determine our analysis aims. Specifically, for this case study, 167 data points were allocated to each of the six groups (see Table 3) yielding a total of 1002 observations in the subset of the Big Data.

Using the design displayed in Table 3, we randomly sample 167 observations from the full data set for each of the six combinations of the factor levels and add them to the data subset.

We are interested in testing the following hypotheses for the data set:

- H_I: There are no interaction effects, i.e., the effects of age (group) and gender are

		Gender	
		Male	Female
Age	≤ 60	167	167
	61-70	167	167
	Over 70	167	167

Table 3: Balanced design for ANOVA displaying the number of observations that are allocated to each combination of the factor levels.

additive.

- H_A: There are no differences in the mean risk of death amongst the different age groups, i.e., $\mu_{\leq 60} = \mu_{61-70} = \mu_{>70}$.
- H_S: There are no differences in the mean risk of death between males and females, i.e., $\mu_{male} = \mu_{female}$.

We consider the following model specification to test our hypotheses of interest for the log risk of death, Y, for the ith observation:

$$E(Y_{i,j,k}) = \mu + \text{agegroup}_j + \text{sex}_k + (\text{agegroup} \times \text{sex})_{j,k};\ j = 1, 2, k = 1.$$

Figure 3(a) demonstrates that the hazard of death appears to increase with age and Figure 3(b) demonstrates that males appear to have a higher risk of death than females. From Figure 3(c), it appears that a slight interaction effect may be present.

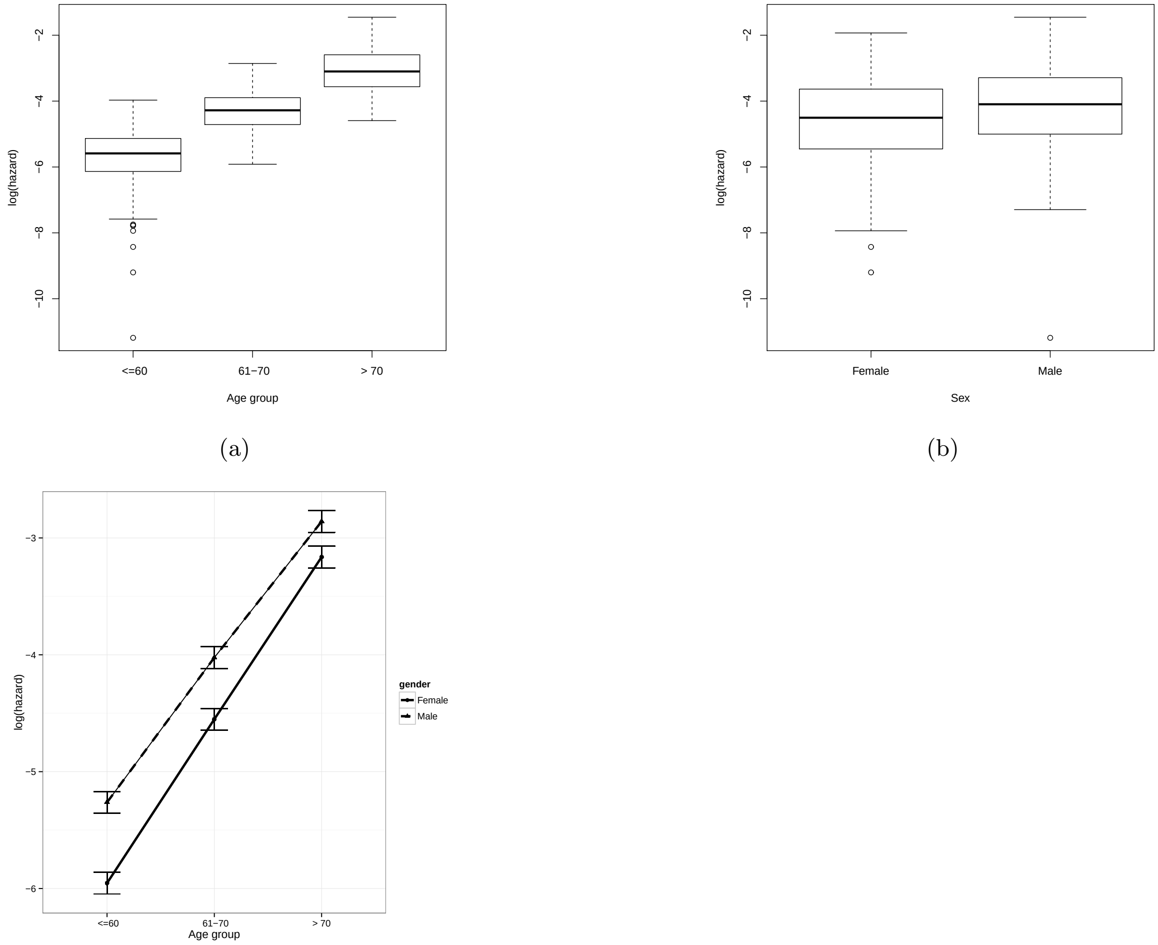
The mean risk of death was found to be statistically significantly different between the age groups ($p < 0.001$), as well as between the sexes ($p < 0.001$). An interaction effect between age group and sex was also found to be statistically significant ($p = 0.0344$). Analysis of residual plots indicate satisfactory model fit.

The same ANOVA model is fitted to the full data set and we again find that the mean risk of death was significantly different between the age groups ($p < 0.001$), as well as between the sexes ($p < 0.001$), and an interaction effect between age group and sex was present ($p < 0.001$).

Appendix D: Relation to other approaches

The experimental design methodology could also be incorporated in other Big Data algorithms, in particular the divide-and-conquer, divide-and-recombine and consensus Monte Carlo techniques mentioned in Section 1 of the main paper. A shared feature of these algorithms is the practice of subsampling the Big Data. A designed approach to choosing these sub-samples has the potential to substantially increase the accuracy and precision of the resultant estimates, with less computational and analytic cost.

Similar ideas of designed subsampling of Big Data have been developed in other contexts. For example, experimental design has been performed for large datasets arising in computer experiments (e.g., Gittins (1979); Scott (2010)). Matching cases and controls via propensity scoring can also be seen as a form of designed sampling of Big Data (Austin, 2011). However,



(a)

(b)

(c)

Figure 3: (a) Boxplot of the log hazard of death for each age group, (b) Boxplot of the log hazard of death for each sex, (c) Interaction plot (with 95% CI) comparing the sexes across the age groups for the log hazard of death.

to our knowledge, the experimental design and decision theoretic approach has not previously been used as a means of sampling large data sets for statistical analysis.

Although their commentary is from the perspective of social scientists and media researchers, Wang et al. (2015) articulate a number of concerns with Big Data analysis that are shared by many research and practice communities. Their concerns include the following: a model based purely on Big Data ignores other theories and disciplines; Big Data can give a false impression of accuracy and objectivity; Big Data can lose its meaning when it is taken out of context; and unequal access to Big Data can create new digital divides. The principled design approach suggested in this paper might provide at least partial relief from these concerns. For example, the focus of the model can remain on the other theories and disciplines, since the theories can inform the models which inform the design which then inform the sampling and hence the analyses. The second and third points can be addressed by smaller, more targetted and design-induced samples. Finally, it may be that access might be more willingly granted to subsets of data rather than the dataset as a whole. Thus a principled experimental design approach holds promise in assuaging these and other similar concerns about Big Data analysis across the span of disciplines.

We reiterate that this approach currently only applies to ‘tall data’. The focus on sampling according to an experimental design implies that there is a specific aim and utility underlying the analysis. A wide class of other problems exist that are more focused on exploratory analyses, with the aim of ‘mining’ the data. These more unstructured analyses are valuable, but are less amenable to design. Having said this, the principles of sampling design might still provide guidance in selecting high quality, informative data to best answer the question at hand. Ongoing research is required to establish optimal procedures in this setup.

References

Amzal, B., Bois, F., Parent, E., and Robert, C. P. (2006). Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association*, 101(474):773–785.

Atkinson, A. C. and Donev, A. N. (1992). *Optimum Experimental Designs*. Oxford University Press, New York.

Austin, P. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46:399–424.

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

Cook, A., Gibson, G., and Gilligan, C. (2008). Optimal observation times in experimental epidemic processes. *Biometrics*, 64(3):860–868.

Dodds, M., Hooker, A., and Vicini, P. (2005). Robust population pharmacokinetic experiment design. *Journal of Pharmacokinetics and Pharmacodynamics*, 32(1):33–64.

Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2014). A sequential Monte Carlo algorithm to incorporate model uncertainty in Bayesian sequential design. *Journal of Computational and Graphical Statistics*, 23(1):3–24.

Duffull, S. B., Graham, G., Mengersen, K., and Eccleston, J. (2012). Evaluation of the pre-posterior distribution of optimized sampling times for the design of pharmacokinetic studies. *Journal of Biopharmaceutical Statistics*, 22(1):16–29.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. Academic Press, New York.

Gittins, J. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 41:148–177.

Han, C. and Chaloner, K. (2004). Bayesian experimental design for nonlinear mixed-effects models with application to HIV dynamics. *Biometrics*, 60:25–33.

Huan, X. and Marzouk, Y. M. (2013). Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288–317.

Müller, P., Berry, D. A., Grieve, A. P., and Krams, M. (2006). A Bayesian decision-theoretic dose-finding trial. *Decision Analysis*, 3(4):197–207.

Nelson, C., Lambert, P., Squire, I., and Jones, D. (2007). Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine*, 26:5486–5498.

Ogungbenro, K. and Aarons, L. (2007). Design of population pharmacokinetic experiments using prior information. *Xenobiotica*, 37:1311–1330.

Pronzato, L. and Walter, E. (1985). Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75(1):103–120.

Pronzato, L. and Zhigljavsky, A. (2012). *Optimal Design and Related Areas in Optimization and Statistics*. Springer, U.S.A.

Pukelsheim, F. and Torsney, B. (1991). Optimal weights for experimental designs on linearly independent support points. *The Annals of Statistics*, 19(3):1614–1625.

Royston, P. and Lambert, P. (2011). *Flexible parametric survival analysis using Stata: Beyond the Cox model*. StataCorp LP, College Station, Texas.

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2015). A review of modern computational algorithms for Bayesian optimal design. *To appear in International Statistical Review*.

Scott, S. (2010). A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639–658.

Stroud, J., Müller, P., and Rosner, G. (2001). Optimal sampling times in population pharmacokinetic studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):345–359.

Wang, C., Chen, M.-H., Schifano, E., Wu, J., and Yan, J. (2015). A survey of statistical methods and computing for big data. *arXiv:1502.07989v1*. arXiv:1502.07989 [stat.CO].

Wu, P.-Y. and Mengersen, K. (2013). A review of models and model usage scenarios for an airport complex system. *Transportation Research Part A-Policy and Practice*, 47:124–140.