

**Optimizing
Pharmacokinetic Studies
Utilizing Microsampling**

by

Helen Yvette Barnett

B.Sc. (Hons), Lancaster University, 2013

M.Res., Lancaster University, 2014

Submitted for the degree of

Doctor of Philosophy

at Lancaster University

October 2017

Optimizing Pharmacokinetic Studies Utilizing Microsampling

by Helen Yvette Barnett, B.Sc (Hons), M.Res.

Submitted for the degree of Doctor of Philosophy at Lancaster University,

October 2017

Abstract

In Pharmacokinetic (PK) studies, inference is made on the absorption, distribution, metabolism and excretion (ADME) of an externally administered compound within the body. This is done by measuring the concentration of the compound in some form of bodily tissue (such as whole blood or plasma) at a number of time points after administration. There are two approaches to PK analysis, modelling and non-compartmental (NCA). The modelling approach uses assumptions of the behaviour of the compound in the body to fit models to the data in order to approximate the concentration versus time curve. Whereas in NCA, no such assumptions are made, and numerical methods are used to approximate this curve. The PK behaviour is summarised by PK parameters that are derived from this approximation, such as the area under the curve (AUC), the maximum concentration (C_{max}) and the time at which this maximum occurs (t_{max}).

In this thesis, three separate topics in the area of PK studies are explored. The first two are motivated by the new blood sampling method of microsampling, which requires a smaller sample volume than traditionally used. Firstly, a methodology is introduced for comparing microsampling to traditional sampling using the derived PK parameters from PK modelling, to find evidence of equivalence of the two sampling methods. The next topic establishes an algorithm for choosing an optimal sparse sampling scheme for PK studies that use microsampling using NCA, developing a two-stage procedure that minimizes bias and variance of the PK parameter estimates. The final topic concerns how

PK analysis can be conducted when some measurements are too low to be reliably detected, again using NCA. Seven methods are explored, with the introduced method of using kernel density estimation to impute values onto censored responses using an iterative procedure showing

Dedicated to Lou, my Ohana.

"Ohana means family. Family means nobody gets left behind, or forgotten."

Acknowledgements

My first acknowledgement is to Lancaster University, both the Mathematics & Statistics Department and the STOR-i DTC. When I first came to Lancaster seven years ago, I never imagined that I would find a home here, with the opportunities that I have been offered. The enjoyment I found in my undergraduate studies motivated me to follow this path to a PhD, and the advantages (including funding) that STOR-i have given me have been indispensable. I have been lucky enough to have an amazing supervisory team during this PhD, both at the University and at Janssen. I'd like to thank Jack, Helena and Tom for all of their help, it has been a real pleasure working with them.

I would also like to acknowledge the friends that have shared the PhD experience with me. Doing a PhD is not an easy three years, but supporting each other has been so important in both our successes and setbacks. Katie and Lucy have been blessings, and I know for certain that my PhD experience has been brighter because they have been a part of it.

I thank Emma, my best friend of many years, for the continuing understanding, joy and laughter that a true friendship can bring. Weekends together have many a time been the calm in the storm and have given me the perspective so often needed in research, which has been a lifesaver in the PhD.

My fiancé Matt, with whom I can't wait to spend the rest of my life, never fails to surprise me in the ways in which he makes me smile. We have grown so much over the past seven years; even though the past three have been stressful for us both, we make the perfect team and have supported each other to make us shine both individually and as a pair. I'm so lucky to have found someone so amazing, and his love has been instrumental to get me where I find myself today. For this I am truly grateful.

Mum and Dad, who have always encouraged me to be the best I can be,

have been a constant source of support in every way. Their ongoing faith has made me continue to work hard and I thank them unreservedly for this. I'm so appreciative of their loving help and guidance; I can only hope that my achievements, including this PhD, make them proud of the daughter that they have inspired me to be.

My sister Lou, is the Anna to my Elsa. Her love, support and most importantly belief in me has been integral not only in the past three years, but also the twenty-two before that. She has always helped me through the tough times and celebrated with me through the good times. I can never thank her enough for what she has done for me, but I hope that she realises how much she means to me and how much her help has contributed to the completion of this PhD.

My final acknowledgement is to the little bundle of crazy black and white fluff, Mrs Moo. She has been with me for the entirety of my PhD and coming home to her beautiful little face every night has brought me happiness, magic and love that I thought were only possible in Disney films.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Chapter 3 has been accepted for publication as Barnett, H.Y., Geys, H., Jacobs, T. and Jaki, T. (2017) Comparing sampling methods for pharmacokinetic studies using model averaged derived parameters. *Statistics in Medicine*.

Chapter 4 has been submitted for publication as Barnett, H.Y., Geys, H., Jacobs, T. and Jaki, T. (2017) Optimal Designs for Non-Compartmental Analysis of Pharmacokinetic Studies. *Statistics in Biopharmaceutical Research*.

The word count for this thesis is 42052 words.

Helen Yvette Barnett

Contents

Chapter	
1	Background 1
1.1	Introduction 2
1.2	Microsampling 4
1.3	Pharmacokinetic (PK) Studies 7
1.3.1	Modelling Approach 10
1.3.2	Non-compartmental Analysis (NCA) 17
1.3.3	Sparse Sampling Schemes 18
1.4	Model Selection, Model Averaging and Simultaneous Inference for Multiple Parameters 21
1.4.1	Model Selection 21
1.4.2	Model Averaging 22
1.4.3	Simultaneous Inference for Multiple Parameters 23
1.5	Optimal Design Theory 29
1.5.1	Model Based Optimality 30
1.5.2	Cost-based Designs 34
1.5.3	Application to PK/PD Studies 34
1.5.4	D-Optimality for multiple response non-linear mixed ef- fect models 41
1.6	Measurements that Cannot be Reliably Detected 46
1.6.1	Definitions 46
1.6.2	Methods 48
2	Thesis Summary 51
3	Paper A: Comparing sampling methods for pharmacokinetic studies us- ing model averaged derived parameters 55
3.1	Introduction 56
3.2	Superiority Testing 57

	x
3.2.1	Baseline Method 57
3.2.2	Extension 63
3.3	Equivalence Testing 65
3.3.1	Example revisited 67
3.3.2	Simulation Studies 68
3.3.3	Extension to Longitudinal Data 73
3.4	Discussion 75
4	Paper B: Optimal Designs for Non-Compartmental Analysis of Pharmacokinetic Studies 78
4.1	Introduction 79
4.2	Method 83
4.3	Results 86
4.3.1	Set Up 86
4.3.2	Initial Results 87
4.3.3	Comparison to Model Based Optimal Designs 91
4.3.4	Application of Minimax Criterion 93
4.4	Choice of Time Points 95
4.5	Discussion 103
5	Paper C: Methods for Non-Compartmental Pharmacokinetic Analysis with Observations below the Limit of Quantification 105
5.1	Introduction 106
5.2	Methods 108
5.2.1	Method 1: Replace BLOQ values with 0 110
5.2.2	Method 2: Replace BLOQ values with $\frac{LOQ}{2}$ 110
5.2.3	Method 3: Regression on Order Statistics (ROS) Imputation 111
5.2.4	Method 4: Maximum Likelihood per timepoint (Summary) 112
5.2.5	Method 5: Maximum Likelihood per timepoint (Imputation) 113

5.2.6	Method 6: Full Likelihood	115
5.2.7	Method 7: Kernel Density Imputation	116
5.2.8	Example Application	118
5.3	Results	119
5.4	Discussion	127
6	Thesis Conclusions, Limitations and Further Work	130
6.1	Overview	131
6.2	Conclusions	131
6.3	Limitations	133
6.4	Further Work	134
	Bibliography	137
	Appendix	
A	PK Parameters as Functions of Model Parameters	143
A.1	For Candidate Model 3.1	144
A.2	For Candidate Model 3.2	144
A.3	For Candidate Models 3.6 and 3.7	144
B	Sampling Time Points for Simulations	145
C	Derivation of Second Order Approximation	147
D	Average Bias of Parameter Estimates	150
E	Additional Type I Error Rate Results	152
F	Minimax Scenarios	154
G	Results Tables I	156

List of Tables

Table

4.1	Sparse Sampling Scheme as suggested by Chapman et al. ¹⁶ All main study animals are sampled, with 10 animals per sex per group. A total of 30 samples per sex per group are taken, with 3 per each of the 10 animals and 5 per each of the 6 timepoints.	81
4.2	The top 10 and bottom 10 ranking designs according to the minimization of variance of the AUC estimate	88
4.3	The top 10 ranking designs according to the minimization of variance of $\Psi = w_1 \text{var}(\widehat{AUC})^* + w_2 \text{var}(\widehat{C}_{max})^*$ with $w_1 = w_2 = 0.5$. Efficiency Measure for a given scheme is the ratio between the variance for that scheme and the best scheme, the larger the value, the less efficient the scheme.	92
4.4	Top 10 rankings of sparse schemes with choice of time points	97
4.5	Top 10 rankings of sparse schemes with choice of time points by $\Psi = w_1 \text{var}(\widehat{AUC})^* + w_2 \text{var}(\widehat{C}_{max})^*$ with $w_1 = w_2 = 0.5$	97
4.6	Optimal Sparse Sampling Scheme with sampling time points (0.5, 1.0, 3.5, 4.0, 7.5, 12.0). \times indicates that the individual subject scheme is shared by the scheme from Chapman et al ¹⁶ and \circ indicates that it is not.	102
5.1	The application of the seven methods to the motivating example illustrated in Figure 5.1.	119
B.1	Sampling Time Points Used in Simulation Studies	146
D.1	The average bias of the estimate of the PK parameters. The true values are $t_{\frac{1}{2}} = 42.80264$ and $C_{max} = 110.9412$	151

E.1	Type I error rate for varying numbers of time points and combinations of PK parameters for an oral administration of a compound. Error bounds for 10,000 simulations are 2.214 and 2.806 for equivalence testing.	153
E.2	Type I error rate for varying numbers of time points and combinations of PK parameters for an oral administration of a compound. Error bounds for 10,000 simulations are 2.214 and 2.806 for equivalence testing.	153
G.1	Top 5 overall schemes using time points (0.5, 1.0, 2.0, 4.0, 9.0, 12.0) according to minimax criteria applied to equally weighted scaled sum of AUC and C_{max} variance: Ranks in the 8 scenarios, maximum rank, and total sum rank. (* indicates the maximum rank for that scheme)	157
G.2	Top 5 overall schemes using time points (0.5, 1.0, 2.0, 4.0, 9.0, 12.0) according to minimax criteria applied to equally weighted scaled sum of AUC and C_{max} variance: Efficiency measure in the 8 scenarios, maximum efficiency measure, and total sum efficiency measure. (* indicates the maximum efficiency measure for that scheme)	157
G.3	Optimal Time Point Choices Top 5 overall time point choices according to minimax criteria: Ranks in the 8 scenarios, maximum rank, and total sum rank. (* indicates the maximum rank for that time point choice)	158
G.4	Top 5 overall schemes using optimal time points (0.5, 1.0, 3.5, 4.0, 7.5, 12.0) according to minimax criteria applied to equally weighted scaled sum of AUC and C_{max} variance: Ranks in the 8 scenarios, maximum rank, and total sum rank. (* indicates the maximum rank for that scheme)	158

- H.1 Results showing average deviation from the non-compartmental \widehat{AUC} and its variance with data generated from the fixed effects model with higher dose and clearance. Results over 1000 simulations. (10 subjects, 6 timepoints) (A) = Analysed using arithmetic means. (G) = Analysed using geometric means. (N) = Confidence interval calculated using normal distribution on the \widehat{AUC} . (L) = Confidence interval calculated using log-normal distribution on the \widehat{AUC} . M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation. * indicates not all analyses were successful. 160
- H.2 Results showing average deviation from the non-compartmental \widehat{AUC} and its variance with data generated from the mixed effects model with higher dose and clearance. Results over 1000 simulations. (10 subjects, 6 timepoints) (A) = Analysed using arithmetic means. (G) = Analysed using geometric means. (N) = Confidence interval calculated using normal distribution on the \widehat{AUC} . (L) = Confidence interval calculated using log-normal distribution on the \widehat{AUC} . M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation. * indicates not all analyses were successful. 161

- H.3 Results showing average deviation from the non-compartmental \widehat{AUC} and its variance with data generated from the fixed effects model with lower dose and clearance. Results over 1000 simulations. (10 subjects, 6 timepoints) (A) = Analysed using arithmetic means. (G) = Analysed using geometric means. (N) = Confidence interval calculated using normal distribution on the \widehat{AUC} . (L) = Confidence interval calculated using log-normal distribution on the \widehat{AUC} . M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation. * indicates not all analyses were successful. 162
- H.4 Results showing average deviation from the non-compartmental \widehat{AUC} and its variance with data generated from the mixed effects model with lower dose and clearance. Results over 1000 simulations. (10 subjects, 6 timepoints) ((A) = Analysed using arithmetic means. (G) = Analysed using geometric means. (N) = Confidence interval calculated using normal distribution on the \widehat{AUC} . (L) = Confidence interval calculated using log-normal distribution on the \widehat{AUC} . M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation. * indicates not all analyses were successful. 163

List of Figures

Figure

1.1	An illustration of some common PK parameters	9
1.2	Diagram of intravenous bolus administration, one compartmental model	11
1.3	Illustration of intravenous bolus administration, one compartmental model on a semi-logarithmic scale	12
1.4	Diagram of intravenous bolus administration, two compartmental model	12
1.5	Illustration of intravenous bolus administration, two compartmental model on a semi-logarithmic scale	13
1.6	Diagram of oral administration, one compartmental model	14
1.7	Illustration of oral administration, one compartmental model on a semi-logarithmic scale	14
1.8	Diagram of oral administration, two compartmental model	15
1.9	Illustration of oral administration, two compartmental model on a semi-logarithmic scale	15
1.10	Illustration of linear interpolation between responses used in NCA.	18
1.11	Examples of different types of sparse sampling scheme	19
1.12	Two different sampling grids: Uniform with respect to response (left) and uniform with respect to AUC (right). ²⁶	37
1.13	MSE as a function of N and k for $u = 2.4$, $\sigma = 9$ and $25 \leq N \leq 40$ ²⁶	40
1.14	An illustration of the relationship between LOB, LOD, and LOQ, indicating relative frequencies of raw analytical signals at varying concentrations. The black line represents results from a blank sample. The red line represents the imprecision of results from a sample with a prespecified low concentration. The blue line represents the distribution of results for a specimen of low concentration meeting the target for total error (bias and imprecision).	48

3.1	Example dataset with individual concentrations (left) and spaghetti plot (right).	57
3.2	Comparison of observed type I error rate for varying number of time points and subjects for use of z and t quantile. Horizontal dotted lines show error bounds for 10,000 simulations	63
3.3	Comparison of observed type I error for varying number of time points and subjects for use of 1st and 2nd order approximation. Horizontal dotted lines show error bounds for 10,000 simulations.	65
3.4	Comparison of observed type I error for varying number of time points and subjects for equivalence testing. Horizontal dotted lines show error bounds for 10,000 simulations.	69
3.5	Power of procedure for equivalence testing with 5, 20 and 100 subjects at 3 time points.	70
3.6	Power of procedure for equivalence testing with 5, 20 and 100 subjects at 7 time points for an oral administration of a compound.	72
3.7	Type I error rate for varying numbers of time points for 5, 10 and 100 total subjects considering AUC and C_{max} as PK parameters for an oral administration of a compound. Horizontal dotted lines show error bounds for 1000 simulations. (Equivalence testing)	74
3.8	Power of procedure for equivalence testing for 5 time points for 5 and 10 total subjects considering AUC and C_{max} as PK parameters for an oral administration of a compound.	75
4.1	An illustration of the population PK model described and the sampling time points.	88
4.2	The relationship between ranks given to schemes using MSE vs variance of the AUC estimate as optimality criterion.	89

4.3	The relationship between ranks given to all schemes using the variances of \widehat{AUC} and \widehat{C}_{max} (top). The relationship between ranks given to schemes using the variances of \widehat{AUC} and \widehat{C}_{max} for the top 50 schemes according to the AUC ranks. Horizontal dashed line represents the top half of the ranks, dotted line represents rank 50 (bottom).	91
4.4	The relationship between ranks given to schemes using the weighted sum of scaled variances of \widehat{AUC} and \widehat{C}_{max} , and D -optimality.	93
4.5	Top 5 overall schemes according to minimax criterion applied to equally weighted scaled sum. Ranks for each of the eight scenarios are plotted.	94
4.6	Measuring the difference between the true population curve and the simulated data at chosen time points.	99
4.7	Optimal Sampling Time Points	100
4.8	Top 5 overall schemes according to minimax criterion applied to equally weighted scaled sum for optimal time points. Ranks for each of the eight scenarios are plotted.	102
5.1	A motivating example, the red numbers indicate the number of observations that are BLOQ for that time point.	107
5.2	Regression on Order Statistics example illustrates how imputed values are calculated.	113
5.3	A graphical illustration of the KD algorithm. Blue crosses indicate a new k_i calculated based on the current \hat{f}_i . Green crosses indicate previous k_i values. The red line indicates the LOQ.	117
5.4	Illustration of Example from Beal, ¹⁰ the red numbers indicate the number of observations that are BLOQ for that time point.	120

- 5.5 Illustration of Example extended from Beal,¹⁰ with smaller clearance and dose. The red numbers indicate the number of observations that are BLOQ for that time point. 121
- 5.6 Results showing deviation from the non-compartmental \widehat{AUC} and its variance with data generated from models with higher dose and clearance. Results over 1000 simulations. (10 subjects, 6 time-points) (F)=Data generated using fixed effects model. (M)=Data generated using mixed effects model. (A)=Analysed using arithmetic means. (G)=Analysed using geometric means. M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation. 123
- 5.7 Results showing deviation from the non-compartmental \widehat{AUC} and its variance with data generated from models with lower dose and clearance. Results over 1000 simulations. (10 subjects, 6 time-points) (F)=Data generated using fixed effects model. (M)=Data generated using mixed effects model. (A)=Analysed using arithmetic means. (G)=Analysed using geometric means. M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation. 124

5.8	Results showing deviation from the non-compartmental \widehat{AUC} and its variance with data generated from both models. Results over 1000 simulations. (10 subjects, 6 timepoints) (F)=Data generated using fixed effects model. (M)=Data generated using mixed effects model. (A)=Analysed using arithmetic means. (G)=Analysed using geometric means. M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M5: ML per timepoint Imputation, M7: Kernel Density Imputation.	125
-----	---	-----

CHAPTER 1

Background

1.1 Introduction

In order to ensure the protection of human subjects in clinical trials and future human patients, it is necessary to use laboratory animals in pre-clinical research. However, a balance must be established between the desire to cure disease in humans and the ethical considerations of the use of animals to achieve this.

These ethical considerations of using animals in drug development are encapsulated by the 3 R principles: replacement, reduction and refinement.⁶⁵ Replacement involves finding different means of collecting the data that does not include the use of conscious living vertebrates. One option is absolute replacement, where no animals are used at all. The other option is relative replacement, where animals that are not conscious living vertebrates are used. Relative replacement can be performed using, for example, *in vitro* methodologies or invertebrates such as nematode worms or fruit flies.

The second R of reduction is quite straightforwardly reducing the number of animals that are needed to take part in the study in order to still obtain the same quality and validity of results. Alternatively, by using the same number of animals to obtain additional information, one can reduce the number of future animals needed for such studies. This can be achieved by developing new experimental design and statistical analysis, in addition to better sharing of resources and data.

The third R of refinement refers to the improvement of laboratory procedures and care of the laboratory animals in order to minimize the exposure of the animals to any potential distress, pain or lasting harm. This ensures that where there is no alternative but to use laboratory animals in a study, the welfare of the animals is considered of the highest priority. Refinement can be implemented by using non invasive or less invasive techniques to test the animals, and the en-

richment of the animals' living environments in order to provide best for both their physical and behavioural needs.

A fourth R, responsibility,⁸ is also considered by some. This is the responsibility of those working with animals in pre-clinical research. The nature of this slightly overlaps with the third R of refinement, as the responsibility of the researchers to treat the laboratory animals with respect and to care for the welfare of the animals is partly encompassed by the criteria of refinement.

One of the most recent developments in laboratory techniques to improve the outlook for laboratory animals is the novel blood sampling method of microsampling (discussed in Section 1.2). This new sampling technique, which requires a reduced volume per sample, is the main motivation behind this research, covering considerations in the issues of reduction and refinement, and of course responsibility. The area of application of this technique that we focus on in this research is for pharmacokinetic (PK) studies in pre-clinical research (discussed in Section 1.3).

Microsampling is an important step forward for pre-clinical research, hence for it to be more widely used, there must be clear evidence that it gives equivalent results to previous sampling methods used, the subject of the first paper in this thesis (Chapter 3). It must also be ensured that the PK studies collect the most accurate information they can, this is done by controlling the sampling times to construct an optimal design, the subject of the second paper in this thesis (Chapter 4). The subject of the third paper is considering how to deal with concentration values that are too low to be reliably detected in such PK studies (Chapter 5).

To provide background on this area, the following sections give an overview of microsampling, PK studies and their analysis, multiple comparison tests, the theory of optimal trial designs and the current methods for dealing with concentration measurements that are too low to be reliably detected.

1.2 Microsampling

The term microsampling covers many laboratory techniques such as dried blood spots, that involve taking samples of considerably less blood than traditional methods. Generally traditional sampling involves taking samples with volume around 500 μL per sample, whereas microsampling requires up to only 10% of this. To extract the large volume of blood required for traditional sampling, the animals must first be warmed in a hot box to increase their blood flow. This puts the animals under stress, which is undesirable not only for the welfare of the animals, but it may also alter some of the characteristics under study. The large volume of blood taken also puts restrictions on the frequency of samples so that animals do not suffer significant blood loss and are not put under unnecessary repeated stress. Microsampling was developed to address these issues, as well as improving the practicability and economic factors related to taking these samples. This relates to the reduction and responsibility principles of the 4 R's of research discussed earlier, as more frequent sampling may lead to reduced animal numbers but also if this stress on animals is avoidable or can be reduced without loss of information from the study, then it is the responsibility of the researchers to take this action. The main two methods of microsampling are dried blood spot (DBS) microsampling and capillary microsampling (CMS).

DBS involves a blood sample being spotted onto pretreated filter paper (known as DBS cards) before being dried, shipped to the laboratory and analyzed using Liquid Chromatography - Tandem Mass Spectrometry (LC-MS/MS). This procedure can be time consuming and costly,²¹ as the process of spotting and drying involves manual intervention and organization. For the appropriateness of using DBS in various settings, it is noted that it is particularly useful in large clinical trials when samples that previously had to be shipped on dry ice could now be kept at room temperature and hence costs were reduced by

tens of thousands of euros.²⁴ The blood : plasma ratio and whether the hematocrit, blood cell partitioning and unbound fraction in plasma are constant were all found to affect the appropriateness of the use of DBS.

CMS uses a predefined low volume of sample collected in a glass capillary micropipette from, for example, the tail vein of a rat. This volume ranges from 8- 25 μ L for whole blood and 32 or 64 μ L for plasma collection.²¹ The process involves filling the glass capillary end to end with capillary force, then placing the capillary in a sample tube before washing out the sample by mixing with water or internal standard (IS) solution and leaving the capillary in the tube.⁴⁸ This diluted sample can then be handled in the same way, with the same laboratory apparatus, as standard plasma samples are dealt with currently. The appropriateness of CMS is more vast than DBS and hence its potential usefulness is more widespread, so it may be used in a wider range of studies than DBS. This is because it offers handling of samples of blood, plasma and other biofluids in the liquid state.⁵⁶

In May 2013 the NC3Rs held a workshop in central London titled 'Overcoming the barriers for uptake of microsampling techniques in regulatory toxicology'. This comprised of representatives from pharmaceutical companies and regulatory bodies to share knowledge and information on microsampling and what the barriers were for further implementation. It was found that there were two main aspects contributing to the barrier: (i) functional and clinical pathology evaluation and (ii) approaches to bioanalysis and toxicokinetics (TK). This illustrates the reluctance of companies to embrace the use of microsampling, and suggests further evidence is needed to support its usage.

Despite this, many advantages to using microsampling over traditional sampling techniques have been established. The reduction in volume of the samples taken addresses the refinement area of the 4 R's. The smaller volumes of sample reduce the suffering of the animal, and do not require the animals to be

warmed prior to sampling to increase the blood flow, as they did for traditional methods.⁴⁸ As well as this, the time that the animals have to be restrained is reduced. This all reduces the stress that the animals are under and hence aids the refinement of the study procedure. The use of capillary microsampling makes use of current laboratory apparatus and procedure for analysis and hence only the sampling procedure changes which reduces retraining needed.

However, one of the main potential benefits of microsampling is the possible reduction of number of animals needed in such studies. Currently, two separate groups of animals are often needed for pre-clinical and toxicokinetic studies, the main study animals in which the pharmacodynamics (therapeutic or adverse effects) are measured, and the satellite animals in which the pharmacokinetic (PK) or toxicokinetic effects are measured. This is because the large blood volume of the sample taken for PK or TK analysis can cause anemia or other secondary effects such as bone marrow or haematological changes that could potentially confound the interpretation of the primary endpoints of the study.¹⁶ For example, in a typical repeated oral dose 4 week rat study with 10 study animals per dose group of each sex, an additional 3 to 9 satellite animals per sex are required.

The use of satellite animals also means that there is no way to correlate the drug action directly with the level of drug in the blood because they are measured in different groups of animals. It is reported that there has been a lack of specific data to demonstrate the magnitude of effect of different blood sampling regimes on critical toxicology endpoints, which has hampered the adoption of toxicokinetic monitoring of main test rats on toxicology studies.⁵⁸ When 1.2 mL of blood was collected at the beginning and end of a 14 day study in adult rats, all clinical pathology parameters remained within background range. It was also found that $6 \times 32 \mu\text{L}$ sampling regimes are toxicologically benign in groups of adult rats, hence unlikely to interfere with the interpretation of toxicologi-

cal endpoints in a regulatory repeated dose toxicity study of at least two weeks length.⁵⁸ These results suggest that the use of microsampling may facilitate the elimination of the use of satellite animals.

Microsampling has already started to be implemented by some companies in the drug development process. Anecdotal evidence exists for its use in non good laboratory practice (non-GLP), it has not been widely practiced and extended to good laboratory practice (GLP), which must adhere to a strict set of principles regarding such factors as consistency and reproducibility.⁵⁸ It is hoped that more research and development in the area of microsampling will provide evidence that will help extend its usage and contribute to reduction and refinement in pre-clinical studies.

1.3 Pharmacokinetic (PK) Studies

This section provides some background to the basic concepts of pharmacokinetics. Since the contents of this section are largely well known and established concepts in pharmacokinetics, many of the individual ideas and formulations are not referenced. The background information is largely based on books by Källén,⁴⁹ Gibaldi³² and Jambhekar & Breen.⁴⁵

Pharmacokinetics is the study of the movement of drugs over time throughout the body. That is, it is concerned with the effect the body has on the drug. This is split into four main categories:

- Absorption- how drugs move from the site of administration (oral, inhalation, etc) to the blood. This does not apply to drugs given by intravenous pathways, as the drug is directly administered to the blood in that case, and there is no need for it to be absorbed;
- Distribution- how drugs move from the blood to other parts of the body (tissues, organs, etc);

- Metabolism- how drugs are transformed or broken down by the body into smaller molecules, which may or may not be pharmacologically active or toxic;
- Excretion- how drugs are removed from the body.

These four aspects are often referred to as the ADME process.

Typically one can only measure the concentration of the drug in some compartment such as whole blood serum or plasma, and do so at specific predefined time points. The aim of pharmacokinetic studies is to then derive as much information as possible about how the body handles the drug from only these measurements.

Some common PK parameters of interest that are estimated from this are:

- C_{max} - the maximum plasma concentration of the drug;
- t_{max} - the time the maximum plasma concentration of the drug is reached;
- AUC - the area under the concentration versus time curve, a measure of exposure to the drug;
- $t_{\frac{1}{2}}$ - elimination half life, the time taken for the plasma concentration to fall to half its maximum value.

These are illustrated in Figure 1.1 with the bottom graph on a semi-logarithmic scale for the elimination phase. That is, using a linear scale on the x -axis and logarithmic scale on the y -axis.

There are two approaches to measuring PK: (i) modelling and (ii) non-compartmental analysis. In the modelling approach, the body is pictured as being made up of compartments and the flow of the compound is modelled using kinetic models involving differential equations. These models are fitted to the data in order to estimate the PK parameters. In non-compartmental analysis, no

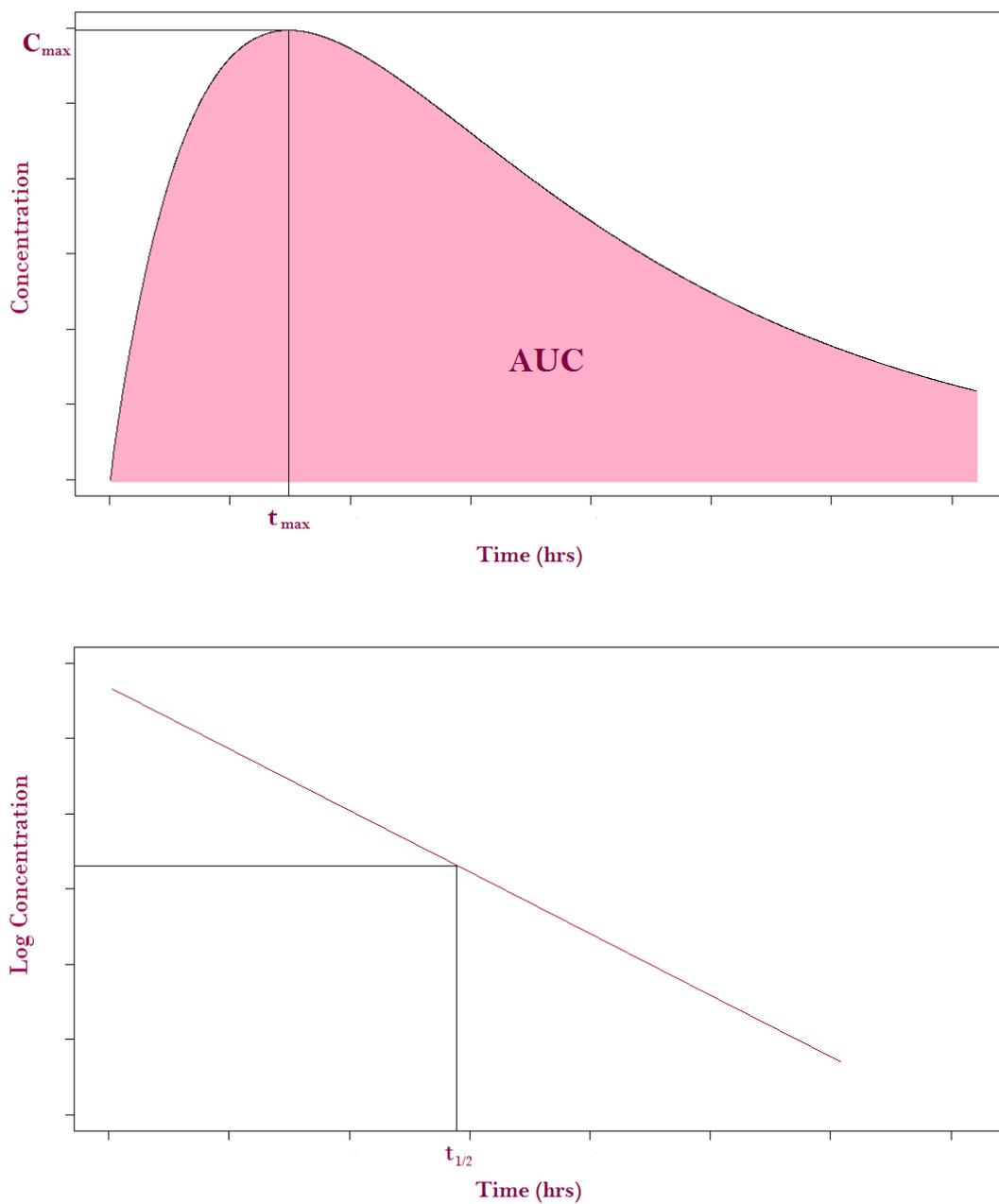


Figure 1.1: An illustration of some common PK parameters

models are fitted to the data, but PK parameters are estimated using numerical methods. The most popular of which is a linear interpolation between responses at consecutive time points.

1.3.1 Modelling Approach

In the modelling approach, the concentration of the drug in the blood plasma is described by a mathematical model. Often these models are derived from assumptions that involve a simplification of the body by breaking it down into compartments, and modelling the diffusion of the drug between these compartments. The more compartments considered, the more complex the PK model.

The choice of model depends on the distribution characteristics of the drug following its administration. The general rule is that the slower the distribution of the drug in the body, the more compartments are required to characterize the concentration versus time curve. Therefore for drugs that are rapidly distributed, a one-compartmental model will adequately describe the plasma concentration over time.

The model is also dependent on the route of administration of the drug. Drugs can be administered in two main ways: oral administration and intravenous (IV) administration. IV administration involves delivering the drug directly into the bloodstream. This can be as an infusion, with a slow increase in concentration, or as a bolus dose, with a very rapid increase in concentration. Oral administration involves ingesting the drug through the digestive system, and hence the absorption of the drug into the bloodstream must be modelled.

When fitting a PK model $C(t)$ to data $y(t)$, it is assumed that the data is related to the PK model by an error model $e(t)$, either additively (1.1) or multiplicatively (1.2). Realizations from the error model $e(t)$ are assumed to be normally distributed with 0 mean.

Additive Error Model:

$$y(t) = C(t) + e(t) \tag{1.1}$$

Multiplicative Error Model:

$$y(t) = C(t) \exp(e(t)) \tag{1.2}$$

What follows are four of the main PK models generally used (although many others are used in practice):

1.3.1.1 Intravenous bolus administration, one compartmental model

In this case the body is modelled as one compartment, with the bolus dose administered directly to this compartment. Therefore the model is solely described by the elimination of the drug from this compartment, as there is no absorption phase. Figure 1.2 shows a diagram of this system.

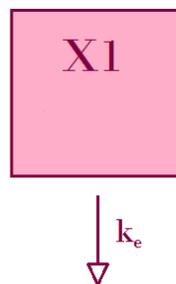


Figure 1.2: Diagram of intravenous bolus administration, one compartmental model

The differential equation used to describe the rate of change of plasma concentration is given in equation 1.3. The integrated equation for this model is given in equation 1.4 where $C(t)$ is the plasma drug concentration at time t , C_0 is the plasma concentration at time 0 and k_e is the elimination rate constant.

$$\frac{dX1}{dt} = -k_e \cdot X1 \quad (1.3)$$

$$C(t) = C_0 e^{-k_e t} \quad (1.4)$$

Figure 1.3 shows an example of data from a one-compartmental IV bolus dose model, plotted on a semi-logarithmic scale. This plot shows a linear relationship, evidence that the distribution of the drug is instantaneous, so the drug is very rapidly distributed in the body, and there is a single phase of elimination.

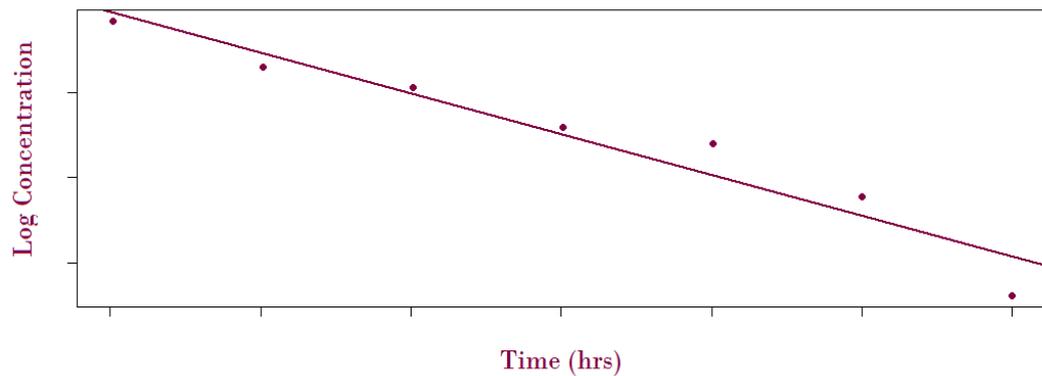


Figure 1.3: Illustration of intravenous bolus administration, one compartmental model on a semi-logarithmic scale

1.3.1.2 Intravenous bolus administration, two compartmental model

Here the body is modelled as two compartments, a central compartment (blood, liver, kidneys) and a peripheral compartment (fat, bone, skin). Again, the bolus dose is administered directly to the central compartment, so there is no absorption phase. However since there are two compartments, the rates of diffusion between the compartments must be taken into account. Figure 1.4 shows a diagram of these compartments.

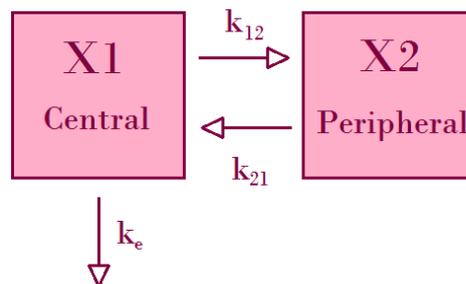


Figure 1.4: Diagram of intravenous bolus administration, two compartmental model

The differential equation to describe the rate of change of plasma concentration in the central compartment is given in equation 1.5, where k_{12} is the rate of diffusion from the central compartment to the peripheral compartment, and k_{21} is the rate of diffusion from the peripheral compartment to the central com-

partment. The integrated equation for this model is given by equation 1.6.

$$\frac{dX_1}{dt} = -k_e \cdot X_1 - k_{12} \cdot X_1 + k_{21} \cdot X_2 \quad (1.5)$$

$$C(t) = Ae^{-\alpha t} + Be^{-\beta t} \quad (1.6)$$

Figure 1.5 shows a graphical illustration of such a model, with the two phases: distribution and post-distribution.

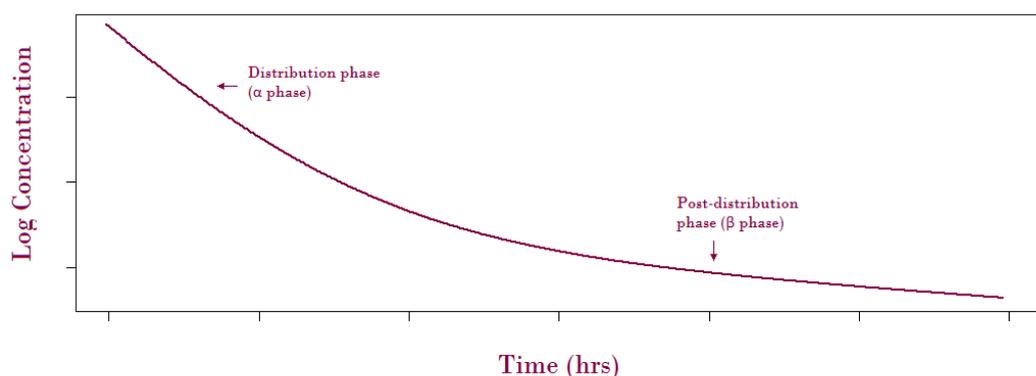


Figure 1.5: Illustration of intravenous bolus administration, two compartmental model on a semi-logarithmic scale

1.3.1.3 Oral administration, one compartmental model

When the drug is administered orally, the plasma concentration depends on the absorption from the gastro-intestinal (GI) tract to the compartment as well as the elimination from the compartment. Figure 1.6 shows a diagram of this process.

The differential equation to describe the rate of change of plasma concentration is given in equation 1.7, where k_a is the absorption rate from the GI tract, and k_e is the elimination rate. The integrated equation is given in equation 1.8, with F the absorbable fraction of the drug, D the dose and V the volume of distribution. Figure 1.7 shows an illustration of this model on a semi-logarithmic

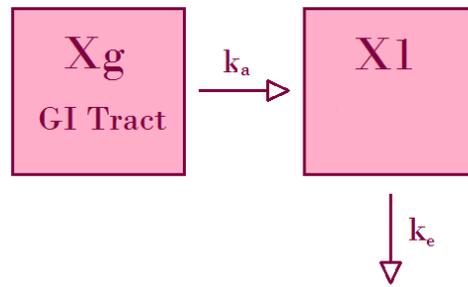


Figure 1.6: Diagram of oral administration, one compartmental model

scale.

$$\frac{dX1}{dt} = k_a \cdot Xg - k_e \cdot X1 \quad (1.7)$$

$$C(t) = \frac{k_a F D}{V(k_a - k_e)} (e^{-k_e t} - e^{-k_a t}) \quad (1.8)$$

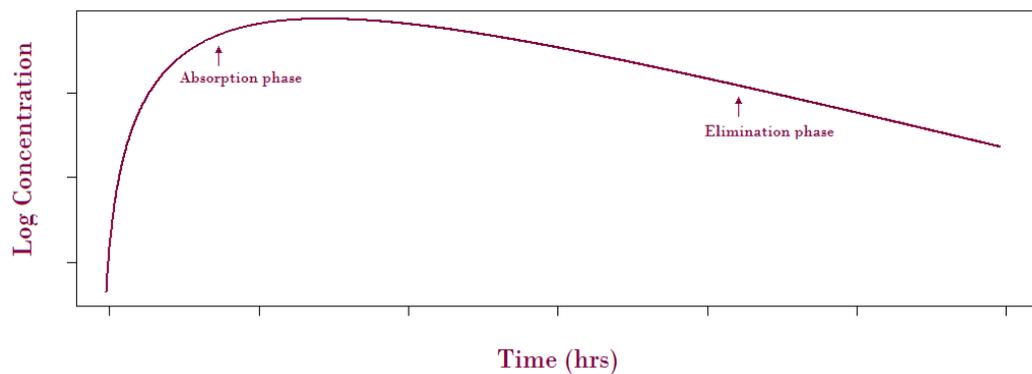


Figure 1.7: Illustration of oral administration, one compartmental model on a semi-logarithmic scale

1.3.1.4 Oral administration, two compartmental model

Here, the concentration in the central compartment depends on four rates of diffusion, the rate from the GI tract to the central compartment, the rate from the central compartment to the peripheral, the rate from the peripheral to the

central, and the rate of elimination from the central compartment. Figure 1.9 shows a diagram of this flow between compartments.

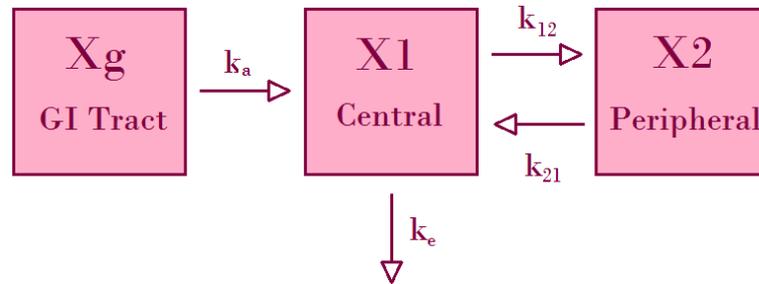


Figure 1.8: Diagram of oral administration, two compartmental model

The differential equation to describe the rate of change of concentration in the central compartment is given in 1.9, with the integrated equation given in equation 1.10.

$$\frac{dX_1}{dt} = k_a \cdot X_g + k_{21} \cdot X_2 - (k_{12} + k_e) \cdot X_1 \quad (1.9)$$

$$C(t) = Ae^{-\alpha t} + Be^{-\beta t} + Ce^{-k_a t} \quad (1.10)$$

Figure 1.9 shows the graphical representation of how the log plasma concentration changes over time, with the α and β phases labelled.

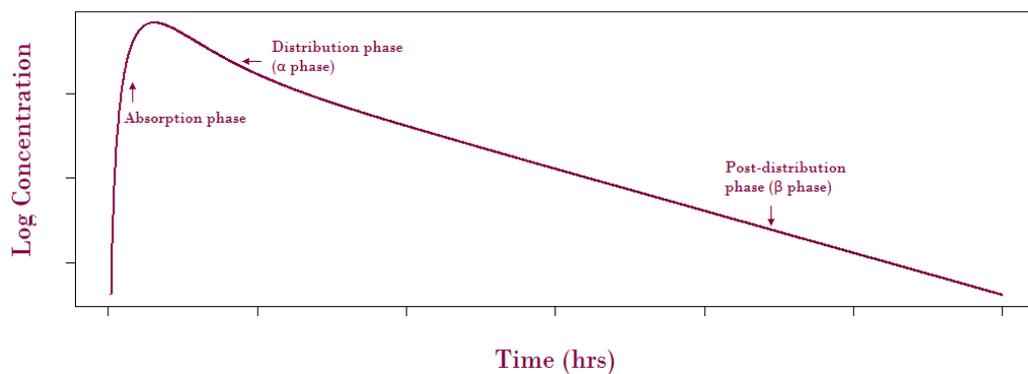


Figure 1.9: Illustration of oral administration, two compartmental model on a semi-logarithmic scale

1.3.1.5 Calculation of PK Parameters

Once the corresponding model has been fitted to the data and model parameters calculated, the PK parameters can be estimated as functions of these parameters. For example, for the one-compartmental oral dose model, t_{max} can be calculated by differentiating the concentration versus time model with respect to t , setting equal to 0 and solving. This gives:

$$t_{max} = \frac{\log k_a - \log k_e}{k_a - k_e}.$$

The maximum concentration, C_{max} , is then given by substituting in $t = t_{max}$ to the model:

$$C_{max} = \frac{k_a F D}{V(k_a - k_e)} (e^{-k_e t_{max}} - e^{-k_a t_{max}}).$$

The AUC_T (Area under the concentration versus time curve until time T) can be calculated by integrating the concentration versus time model over t between $t = 0$ and $t = T$, giving:

$$AUC_T = \frac{k_a F D}{V(k_a - k_e)} \left(\left(\frac{\exp(-k_a T) - 1}{k_a} \right) - \left(\frac{\exp(-k_e T) - 1}{k_e} \right) \right).$$

Then in the limit as $T \rightarrow \infty$, we obtain:

$$AUC_\infty = \frac{k_a F D}{V(k_a - k_e)} \left(\left(\frac{1}{k_a} \right) - \left(\frac{1}{k_e} \right) \right).$$

The variances of these estimates can be approximated using the variance-covariance matrix of the model parameters from the model fitting, and the delta method (covered in Chapter 3).

1.3.1.6 Non-Linear Mixed Effects Models

So far, the models considered have only been concerned with fixed effects. That is, given a dataset and a model, the point estimate and variance are calculated for each of the model parameters. However, when considering a population, as often is the case in PK studies, it is generally more appropriate to

consider each subject in the population having their own individual model parameters.

Take for example the volume of distribution parameter in the one compartmental oral dose PK model. It may be assumed that the parameter has additive random effects $V_i = \tilde{V} + \eta_i$ or exponential random effects $V_i = \tilde{V} \exp \eta_i$, where \tilde{V} is the population parameter, V_i is the individual parameter and $\eta_i \sim N(0, \sigma^2)$ for some variance σ^2 . If more than one model parameter has mixed effects then there may also be a correlation between the random effects. In this framework, individual PK parameters can be estimated as well as population PK parameters.

1.3.2 Non-compartmental Analysis (NCA)

In this approach, no assumptions are made on the processes within the body that control the ADME process and hence no models are fitted to the data in the analysis. This offers the obvious advantage that with fewer assumptions, there is less room for error due to mis-specification of the model, and also avoids any complication that may arise in model fitting if the data is not harmonious with the model.

Since no model is fitted in the analysis, an approximation to the concentrations versus time curve must be made by some other means. The most prevalent method uses a linear approximation between measurements, as illustrated in Figure 1.10.

The non-compartmental estimate to the AUC for an individual subject can then be calculated using the trapezium rule as follows:

$$\widehat{AUC} = \sum_{j=1}^J \omega_j C_j, \quad (1.11)$$

where sampling times are labelled t_j for $j = 1, \dots, J$, $C_j = C(t_j)$ equals the

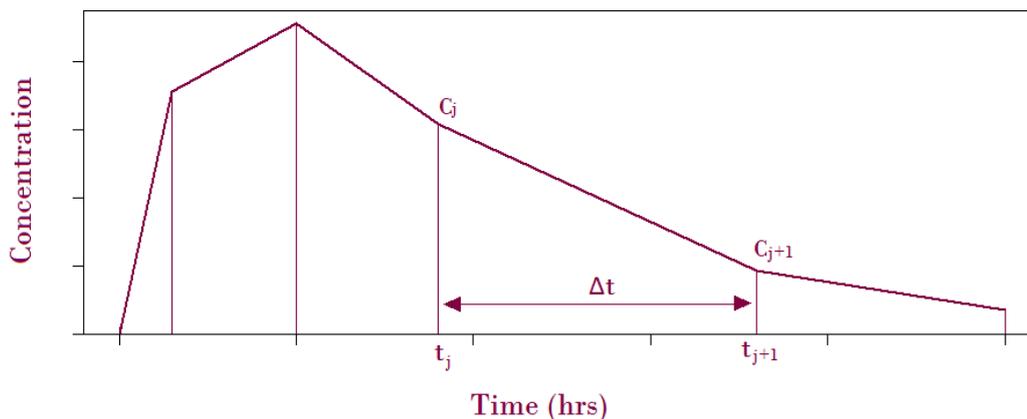


Figure 1.10: Illustration of linear interpolation between responses used in NCA.

concentration at time t_j and ω_j are weights defined as:

$$\begin{aligned}\omega_j &= \frac{t_{j+1} - t_{j-1}}{2} \quad \text{for } j = 1, 2, \dots, (J - 1), \\ &= \frac{t_J - t_{j-1}}{2} \quad \text{for } j = J.\end{aligned}$$

For a population estimate, one may instead use the mean concentration observed at each time t_j .

1.3.3 Sparse Sampling Schemes

In many scenarios, it is not possible to sample from every subject at every time point, and hence a sparse sampling scheme must be used. There are a number of different types of sparse sampling scheme, illustrated in Figure 1.11. The most simple is the serial sacrifice design, where each subject is sampled only once. In the illustration, only one subject is sampled per timepoint, but in larger studies this is extended to allow multiple subjects to be sampled per timepoint. The second type of sparse sampling scheme considered is the batch sampling design. The timepoints are split into disjoint batches so that these batches partition the full set of timepoints and each subject is assigned to a batch. This means that at each time point, only one batch of subjects is sampled and no other batch may also be sampled at that time point. Flexible sampling designs also allow

for subjects to be sampled at multiple timepoints but do not place the restriction on the timepoints that they must be split into disjoint batches. The designs are restricted however by the rule that for each set of timepoints that a subject is sampled at, at least two subjects must be sampled at these timepoints. This can result in unbalanced designs, i.e. different numbers of subjects sampled at each timepoint. The final type of sparse sampling design considered is the alternative sampling design. Here there is no restriction on when each subject can be sampled, although in the example in Figure 1.11 the number of subjects sampled at each timepoint has been restricted in order to make the design balanced.

Serial Sacrifice Sampling Design						
Subject number	Sampling timepoint					
	#1	#2	#3	#4	#5	#6
1	X					
2		X				
3			X			
4				X		
5					X	
6						X
	n=1	n=1	n=1	n=1	n=1	n=1

Batch Sampling Design						
Subject number	Sampling timepoint					
	#1	#2	#3	#4	#5	#6
1	X			X		X
2	X			X		X
3	X			X		X
4		X	X		X	
5		X	X		X	
6		X	X		X	
	n=3	n=3	n=3	n=3	n=3	n=3

Flexible Sampling Design						
Subject number	Sampling timepoint					
	#1	#2	#3	#4	#5	#6
1	X		X		X	
2	X		X		X	
3		X	X			X
4		X	X			X
5	X			X	X	
6	X			X	X	
	n=4	n=2	n=4	n=2	n=4	n=2

Alternative Sampling Design						
Subject number	Sampling timepoint					
	#1	#2	#3	#4	#5	#6
1	X	X				X
2		X		X		X
3	X		X	X		
4		X	X		X	
5			X		X	X
6	X			X	X	
	n=3	n=3	n=3	n=3	n=3	n=3

Figure 1.11: Examples of different types of sparse sampling scheme

From a modelling perspective, the estimation of the PK parameter estimates and their variance does not rely on the sampling scheme. However, for NCA, the point estimate of the PK parameter is not affected but the variance of this estimate is. Methods for approximating the variance of the \widehat{AUC} have been explored for serial sacrifice designs with multiple subjects sampled at each time

point^{6,44,55,71} and for batch sampling designs.^{37,41,75} The basis of these methods is that the variance of the total AUC can be estimated based on the sample variance of the individual partial AUC s for each batch. The variance of the \widehat{AUC} until the last observed timepoint t_J is calculated by the following.

Letting $J_b \subseteq \{1, \dots, J\}$ be the indices of time points investigated in batch $b = 1, \dots, B$, n_b the number of subjects in batch b and X_{it_j} the observed response of subject i at time t_j . The approximation of the variance is given by:

$$\widehat{V}(\widehat{AUC}) = \sum_{b=1}^B \frac{s_b^2}{n_b}, \quad (1.12)$$

where

$$s_b^2 = \frac{1}{n_b - 1} \sum_{i=1}^{n_b} \left(\sum_{j \in J_b} \omega_j X_{it_j} - \frac{1}{n_b} \sum_{k=1}^{n_b} \sum_{j \in J_b} \omega_j X_{kt_j} \right)^2.$$

Full sampling designs and serial sampling designs are special cases of the batch design. Full sampling considers one batch of subjects, with all timepoints investigated for that batch. Serial sampling considers J batches, each batch with only one timepoint investigated.

For the flexible sampling designs, the method for approximating the variance of the \widehat{AUC} is more complex.⁴³ Again, the variance of the total AUC is estimated based on the sample variance of the individual partial AUC s for each schedule. Let N_{jk} be the number of subjects sampled at both time points t_j and t_k , N_j the number of subjects sampled at time t_j , $J_s \subseteq \{1, \dots, J\}$ be the indices of time points investigated in schedule s , n_s the number of subjects on schedule s , \bar{X}_j the mean of observed concentrations at time t_j and $\delta_j^s = \frac{n_s}{N_j}$ (that is the proportion of samples at time t_j that come from subjects on schedule s). The variance is then approximated as follows:

$$\widehat{V}(\widehat{AUC}) = \sum_{s=1}^S \frac{1}{n_s} \sum_{j \in J_s} \sum_{k \in J_s} \delta_j^s \delta_k^s \omega_j \omega_k \hat{\sigma}_{j,k}, \quad (1.13)$$

where

$$\hat{\sigma}_{j,k} = \sum_{i=1}^{N_{jk}} \frac{(X_{it_j} - \bar{X}_j)(X_{it_k} - \bar{X}_k)}{(N_{jk} - 1) + \left(1 - \frac{N_{jk}}{N_j}\right) \left(1 - \frac{N_{jk}}{N_k}\right)}.$$

For all such types of design, the R package PK⁴² can be used to approximate the variance of the *AUC* estimate in the NCA framework.

For the alternative sampling design, there is no analytic form of the variance of the *AUC*. This is due to no requirements for repeated schedules. However in practice, these designs are used and hence must be considered.

1.4 Model Selection, Model Averaging and Simultaneous Inference for Multiple Parameters

This section provides background to the issues of model selection, model averaging, and simultaneous inference for multiple parameters, as these are relevant for the work completed in Chapter 3.

1.4.1 Model Selection

In statistical inferences, including PK studies, the process of model selection is an important one. The aim of the model selection process is to find the best model out of a set of proposed models. There are various established methods to do such.

For nested models, one may construct pairwise comparisons of models using the Likelihood Ratio Test. This involves, say for model A with likelihood L_A and p_A parameters nested inside model B with likelihood L_B and p_B parameters, constructing a null hypothesis that model A is adequate, and an alternative hypothesis that the more complex model B is required. Then the test statistic $Q = 2 \log\left(\frac{L_B}{L_A}\right)$ is calculated and the null hypothesis is rejected if Q is greater than the critical value of the Chi-squared distribution with $p_B - p_A$ degrees of freedom.

However more versatile and easier to implement methods are often preferred, which make use of an information criterion in the form of a penalized likelihood function:

$$I = -2 \log(L) + q, \quad (1.14)$$

where L is the likelihood and q is a penalty function. Akaike's Information criterion (AIC) uses the penalty function $q = 2p$, where p is the number of parameters in the model,² and Bayes information criterion (BIC) uses the penalty function $q = 2p \log n$ where n is the number of observations. AIC_c is also often used, which has penalty function $q = 2p + \frac{2k(k+1)}{n-k-1}$. This gives more of a penalty for extra parameters than the AIC, and as n gets very large, approaches the value of the AIC. The model with the smallest value of I is deemed to be the best approximating model.

Using one of these criteria, it is often the case that once the best model is chosen, then all further inference is conditional on the chosen model being the truth, which of course may not be the case. This provides the motivation for the use of model averaging, discussed in the following sections.

1.4.2 Model Averaging

Model averaging deviates from choosing one best model, instead including the variability in the model selection process in the estimation of parameter uncertainty.^{12,17} The main idea is to give weights w_k to each model M_k that are then incorporated into the value and variance of the estimator of the parameter of interest. These weights are scaled such that $\sum_{k=1}^K w_k = 1$. The parameter common to all models, θ can then be estimated by:

$$\hat{\theta} = \sum_k w_k \hat{\theta}_k, \quad (1.15)$$

where $\hat{\theta}_k$ is the estimate of θ under model M_k .

This poses two obvious questions, the first being how to estimate the weights and the second how to incorporate them in the value and variance of the estimator.

To consider how to estimate the weights w_k , suppose there are K models with information criterion $I_k = -2 \log(L_k) + q_k$ for model k , Buckland et al.¹² consider the ratio:

$$\frac{L_i \exp(-q_i/2)}{L_j \exp(-q_j/2)} = \frac{\exp(-I_i/2)}{\exp(-I_j/2)}, \quad (1.16)$$

to compare model i with model j . If $q_i = q_j$, i.e. the penalties are the same for each model then this is simply the ratio of likelihoods, which is the Bayes Factor for comparing simple models. If the prior odds ratio is 1, then this represents the value of the posterior odds ratio. Therefore a sensible choice of w_k ¹² is:

$$w_k = \frac{\exp(I_k/2)}{\sum_{i=1}^K \exp(-I_k/2)}, \quad k = 1, \dots, K. \quad (1.17)$$

This ensures that two models with the same value of I are given the same weight.

The method of bootstrapping²³ can also be used to estimate the weights. Resampling with replacement gives the bootstrap resamples that are then used to calculate the weights. The weight w_k is the proportion of times in the bootstrap resamples that model M_k is chosen to be the best approximating model. This bootstrap method has the feature that if two models have identical likelihoods, the method will always choose one model over the other, even if that choice is random. Then the sum of the weights of the two models is the same as the weight of one of the models would have been, were the other model to have been omitted.

1.4.3 Simultaneous Inference for Multiple Parameters

In order to make the comparison between the previous method and the new method of microsampling, simultaneous inference on multiple derived parameters must be made. The simultaneous inference of multiple parameters is

a topic that has been of much discussion, especially in the context of clinical trials where there are multiple endpoints. This is due to the complications caused by attempting to compute the simultaneous comparisons, especially in terms of controlling the type I error rate.³ Attempting to carry out each test as if it were the only one, at the same type I error rate, will lead to misleading results and in fact invalidate the hypothesis tests.

One method of carrying out the comparison is adjusting p -values for these simultaneous hypothesis tests, building on the basic and widely used Bonferroni adjustment. The adjusted p -value is defined as, for a particular hypothesis within the many being tested, *'the smallest overall (i.e., "experimentwise") significance level at which the particular hypothesis would be rejected'*⁷³. Therefore this adjusted p -value may be directly compared to a given significance level α , the null hypothesis is rejected if the adjusted p -value is less than or equal to α .

The Bonferroni adjustment alters the significance level used for each test. Say k null hypotheses H_1, H_2, \dots, H_k are being tested at an experimentwise significance level of α . Then hypothesis H_i is tested individually at significance level α_i such that $\sum_{i=1}^k \alpha_i = \alpha$. It is standard for each $\alpha_i = \frac{\alpha}{k}$, although it is possible for the allocations to be uneven. Therefore in terms of p -values, if p_i is the unadjusted p -value for testing H_i , then using the equal allocation of individual significance levels as described above, H_i is rejected when $kp_i \leq \alpha$. The Bonferroni adjusted p -value is therefore $p_{\text{Bonf}} = kp_i$.

Holm³⁸ suggests an improvement on this adjustment by proposing a sequential procedure for rejecting the hypotheses, a more powerful adjustment called Holm's procedure. Consider again the situation with k null hypotheses as above. This method still has the experimentwise error rate of α but is less conservative. The unadjusted p -values are ordered so that $p_1 \leq p_2 \leq \dots \leq p_k$. For each hypothesis test, p_i is compared to $\frac{\alpha}{n-i+1}$ as opposed to $\frac{\alpha}{n}$. The adjusted p -value to be compared to α is not necessarily $(n-i+1)p_i$ however, due to the

sequential nature of the procedure. The hypothesis associated with the smallest p -value is tested first, followed by the next largest and so on. The procedure is stopped when a non-significant result is found, and all other tests are considered non-significant. That is, H_i is rejected if $(n - i + 1)p_i \leq \alpha$ provided that $(n - j + 1)p_j \leq \alpha \quad \forall \quad j < i$.

A similar approach to Holm's procedure is Hochberg's procedure,³⁶ which is essentially the reverse of Holm's, in that the hypothesis associated with the largest p -value is tested first. The same level of significance $\frac{\alpha}{n-i+1}$ is used for each test of hypothesis H_i , and testing continues until a significant result is obtained, then all further untested hypotheses are considered significant. H_i is therefore rejected if $(n - j + 1)p_j \leq \alpha \quad \forall \quad j \geq i$.

For all hypotheses in a set of hypotheses, Simes⁶⁸ introduced the following global test: Reject $H_0 = \{H_1, H_2, \dots, H_n\}$ if $p_i \leq (\frac{i\alpha}{n})$ for at least one p_i , where the p_i are the ordered unadjusted p -values. It was proven by Simes that when the tests are independent, this global test has level α . For dependent tests, Simes provided simulations to show that the test also has level α apart from in exceptional circumstances. The p -value for this global Simes test is simply the smallest of the $\frac{np_i}{i}$ values, since H_0 is rejected if any $\frac{np_i}{i} \leq \alpha$. This test is used as part of Hommel's procedure.^{39,40}

Both Holm and Hommel's test are based around the *closed test procedure* principle⁵² that is as follows. Suppose we have a collection of n hypotheses: H_1, H_2, \dots, H_n . Let all possible combinations of subsets of these hypotheses be defined as $H_I = \bigcap \{H_i : i \in I\}$ for all $I \in K$, where K is the set of all non-empty subsets of $\{1, 2, \dots, n\}$. For each H_I , let there exist a test based on statistic T_I . Define $J \in K$ and $J \supseteq I$, so that subset I is included in subset J . For any given α , H_I is rejected if every H_J is rejected at level α by the corresponding T_J . The probability of wrongly rejecting at least one hypothesis when testing all of H_I is at most α .

To carry out the procedure, one starts with the global test $H_I = \bigcap \{H_i : i = 1, \dots, n\}$. Then if this test is rejected at level α then continue to test at level α each subset of $n - 1$ hypotheses. This process continues until either a test is not rejected at the level α , or eventually all tests, including those of subset size 1 are rejected. If the test statistic is from a Bonferroni test, then the resulting closed test procedure is Holm's procedure. If the test statistic is from Sime's test, then the closed test procedure is Hommel's procedure. In terms of adjusted p -values, let p_I be the unadjusted p -values for hypothesis H_I and test T_I . If $p_J \leq \alpha \quad \forall \quad H_J : J \supseteq I$ then reject H_I . The adjusted p -value for H_I is therefore the largest of the p_J 's. This of course unfortunately means conducting tests for each of the possible combinations of subsets, which is $\sum_{i=1}^n \binom{n}{i} = 2^n - 1$, which will obviously get very large very quickly. In some cases, however, shortcuts exist so that all of the tests need not be calculated.

These methods all stand alone for the simultaneous inference of multiple parameters, which provides a benchmark for future methods. However for the purpose of this research, this inference must also incorporate the model averaging discussed previously. Jensen & Ritz⁴⁷ discuss simultaneous inference for model averaging of derived parameters, specifically for the use of finding the derived parameters Bench Mark Dose (BMD) and the lower limit of the confidence interval for this (BMDL) in non-linear dose response modelling. This is a special kind of effective dose estimation in toxicology that includes additional prior information available. This procedure involves calculating simultaneous confidence intervals for the multiple comparison procedure; the adjusted quantiles for these intervals (corresponding to the theory of adjusting p -values discussed previously) are based on correlation estimates between the multiple parameter estimates.

The setup of the procedure is as follows. n independent observations Y_1, \dots, Y_n

are assumed to follow a semi-parametric model, for example :

$$Y_i \sim \mathcal{N}(\mu_i(\tau_0), \sigma^2), \quad (1.18)$$

for some known linear or non-linear mean function μ_i that depends on an unknown parameter vector τ_0 .

The model-averaged estimate used is similar to that considered by Buckland et al.,¹² that is the weighted average of parameter estimates. In this context, consider K candidate models that may or may not be mutually nested are parametrized by parameters (τ_1, \dots, τ_K) of dimensions (p_1, \dots, p_K) . Then for model $k = 1, \dots, K$, there are L derived parameters of interest : $\theta_1, \dots, \theta_L$. These are differentiable functions of the model parameters, so that $\theta_{1k} = g_{1k}(\tau_k), \dots, \theta_{Lk} = g_{Lk}(\tau_k)$. Therefore the model averaged estimate of the parameter θ_l ($l = 1, \dots, L$) is the weighted mean of the estimates $\hat{\theta}_{l1}, \dots, \hat{\theta}_{lK}$ from all K models:

$$\hat{\theta}_{l,MA} = \sum_{k=1}^K w_k \hat{\theta}_{lk} = \sum_{k=1}^K w_k g_{lk}(\hat{\tau}_k), \quad (1.19)$$

where the w_k 's are the model specific weights such that $\sum_{k=1}^K w_k = 1$. It is noted that for these methods, the specification of the particular weights is not required, although one may use the methods suggested by Buckland et al.¹²

The simultaneous inference is based on the methods of Pipper et al.,⁵⁷ and depends on the p_k -dimensional asymptotic expansion of the maximum likelihood estimator $\hat{\tau}_k$ for each of the K models :

$$n^{\frac{1}{2}}(\hat{\tau}_k - \tau_k) = n^{-\frac{1}{2}} \sum_{i=1}^n (\mathbf{I}_k^{-1}) \tilde{\Psi}_{ki} + o_P(1) \quad (1.20)$$

$$= n^{-\frac{1}{2}} \sum_{i=1}^n \Psi_{ki} + o_P(1), \quad (1.21)$$

where \mathbf{I}_k^{-1} is the inverse Fisher information matrix for model k and $\tilde{\Psi}_{ki}$ is the score function for observation i of model k , both evaluated at the parameter value τ_k .

The key idea exploited by Pippenger et al.⁵⁷ is that the asymptotic representation is retained for stacked parameter estimates. That is, the following $\sum_{k=1}^K p_K$ -dimensional asymptotic expansion:

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) = n^{-\frac{1}{2}} \sum_{i=1}^n \Psi_i + o_P(1), \quad (1.22)$$

where $\hat{\boldsymbol{\tau}} = (\hat{\boldsymbol{\tau}}_1, \dots, \hat{\boldsymbol{\tau}}_K)$, $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_K)$ and $\Psi_i = \Psi_{1i}, \dots, \Psi_{Ki}$. Since the standardized score vectors, the Ψ_i 's, are independent and identically distributed random variables with mean zero and finite variance, application of the central limit theorem gives asymptotic normality, that is:

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \xrightarrow{D} \text{MVN}_K(0, \boldsymbol{\Sigma}) \quad \text{as } n \rightarrow \infty, \quad (1.23)$$

where the variance-covariance matrix $\boldsymbol{\Sigma}$ is defined as the limit in probability of $n^{-1} \sum_{i=1}^n \Psi_i^T \Psi_i$ as a consequence of the law of large numbers. By substituting in parameter estimates, a consistent estimator of the variance-covariance matrix is obtained: $\hat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^n \hat{\Psi}_i^T \hat{\Psi}_i$. Hence one can obtain estimates of the correlations between the different parameter estimates from different model fits to the same data.

An asymptotic approximation for a single model-averaged estimate may be obtained by use of the delta method,⁷⁰ a method which uses a Taylor expansion to approximate a vector. This approximation is given by⁴⁷ to be:

$$n^{\frac{1}{2}}(\hat{\theta}_{l,MA} - \theta_0) = \mathbf{w}^T \left(\frac{dg_{lk}}{d\boldsymbol{\tau}} \right)^T n^{\frac{1}{2}}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) + o_P(1), \quad (1.24)$$

where $\mathbf{w} = (w_1, \dots, w_k)$ and $\frac{dg_{lk}}{d\boldsymbol{\tau}}$ is the $K \times p$ matrix $(\frac{d}{d\boldsymbol{\tau}} g_{lk}(\hat{\boldsymbol{\tau}}_1)^T, \dots, (\hat{\boldsymbol{\tau}}_K)^T)$ for $l = 1, \dots, L$. The variance of the model-averaged estimate, $\hat{\theta}_{l,MA}$, may be approximated using the previous asymptotic results to be:

$$\text{var}(\hat{\theta}_{l,MA}) \approx n^{-1} \left\{ \left(\frac{dg_{lk}}{d\boldsymbol{\tau}} \right) \mathbf{w} \right\}^T \hat{\boldsymbol{\Sigma}} \left(\frac{dg_{lk}}{d\boldsymbol{\tau}} \right) \mathbf{w}, \quad (1.25)$$

where both within and between model variability is captured in terms of the diagonal and off-diagonal elements of $\hat{\boldsymbol{\Sigma}}$ respectively.

The combined asymptotic representation of the vector of the L model-averaged estimates is then written by Jensen & Ritz⁴⁷ as:

$$n^{\frac{1}{2}}(\hat{\theta}_{lMA} - \theta_0) = (\mathbf{I}_L \otimes w) \cdot \frac{dg}{d\boldsymbol{\tau}} n^{\frac{1}{2}}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) + o_P(1), \quad (1.26)$$

where \mathbf{I}_L is the $L \times L$ identity matrix and $\frac{dg}{d\boldsymbol{\tau}}$ is the $KL \times p$ matrix obtained by stacking the matrices $\frac{dg_{1k}}{d\boldsymbol{\tau}}, \dots, \frac{dg_{Lk}}{d\boldsymbol{\tau}}$.

A simulation study was carried out by Jensen & Ritz⁴⁷ alongside application to example datasets from the literature in order to explore the coverage properties of this asymptotic approach. Performance is compared to various other methods, for example those of Buckland et al,¹² Wu et al,⁷⁴ the unadjusted Bonferroni and methods using a single model. Results from these studies show that using a single model can result in confidence intervals too narrow, as they do not include model uncertainty. Other methods show conservative coverage, especially in cases with high correlations.

The method introduced by Jensen & Ritz is the focus to build on for the work in Chapter 3, bearing in mind the improvement it offers over the other methods previously discussed.

1.5 Optimal Design Theory

When designing any trial in a clinical or pre-clinical context, the choice of sampling schedule is a very important one. The number of samples taken and the time that they are taken will have a great impact on the accuracy of the estimates of the PK parameters. Clearly it is desirable to have the most accurate estimates, and thus lower variability. How to measure this and how to optimize it under constraints of PK trials is one of much discussion in the literature. This section discusses the general types of optimality and their application to PK/PD studies.

1.5.1 Model Based Optimality

When considering optimal designs from a PK modelling approach, there are different types of optimality criteria. These are all based on the assumed PK model. This requires some initial notation to be defined. Observations $\{y_{ij}\}$ are assumed to satisfy:

$$y_{ij} = \eta(\mathbf{x}, \boldsymbol{\theta}) + \epsilon_{ij}, \quad i = 1, 2, \dots, n; \quad j = 1, \dots, r_i; \quad \sum_{i=1}^n r_i = N, \quad (1.27)$$

where $\eta(\mathbf{x}, \boldsymbol{\theta})$ is the response function of m unknown parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$ and design variables \mathbf{x} . Consider a linear model, $\eta(\mathbf{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{f}(\mathbf{x}_i)$ where $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x})]^T$ is a vector of known ‘‘basis’’ functions. The ϵ_{ij} are uncorrelated random variables with zero mean and constant variance.

The design of the experiment is denoted as the collection:

$$\xi_N = \left\{ \begin{array}{c} \mathbf{x}_1, \dots, \mathbf{x}_n \\ p_1, \dots, p_n \end{array} \right\} = \{\mathbf{x}_i, p_i\}_1^n, \quad \text{where } p_i = r_i/N. \quad (1.28)$$

The \mathbf{x}_i are the design points, the r_i are the number of samples taken at point \mathbf{x}_i for $i = 1, \dots, n$ and N is the total number of observations.

The information matrix $\mathbf{M}(\xi_N)$ of the design is given by the sum of information matrices corresponding to individual observations::

$$\mathbf{M}(\xi_N) = \sum_{i=1}^n r_i \boldsymbol{\mu}(\mathbf{x}_i), \quad \text{where } \boldsymbol{\mu}(\mathbf{x}) = \sigma^{-2} \mathbf{f}(\mathbf{x}) \mathbf{f}^T(\mathbf{x}), \quad (1.29)$$

such that

$$\mathbf{M}(\xi_N) \boldsymbol{\theta} = \mathcal{Y}, \quad (1.30)$$

where

$$\mathcal{Y} = \sigma^{-2} \sum_{i=1}^n r_i \bar{y}_i \mathbf{f}(\mathbf{x}_i). \quad (1.31)$$

It can be proved for large samples that under certain regulatory assumptions, the variance-covariance matrix of the MLE $\hat{\boldsymbol{\theta}}_N$ is approximated by the inverse of

the information matrix:

$$\text{Var}[\boldsymbol{\theta}] = \mathbf{D}(\xi_N, \boldsymbol{\theta}) \approx \mathbf{M}^{-1}(\xi_N, \boldsymbol{\theta}). \quad (1.32)$$

The equation:

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_N)^T \mathbf{M}(\xi_N) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_N) = R^2, \quad (1.33)$$

defines an ellipsoid of concentration, which generates confidence regions for normal linear models.⁵⁹ Therefore the “larger” the matrix $\mathbf{M}(\xi_N)$ (equivalent to the “smaller” the matrix $\mathbf{D}(\xi_N)$) is, the “smaller” the ellipsoid of concentration will be. Hence in order to optimize the precision of the estimator $\hat{\boldsymbol{\theta}}_N$, one wishes to “maximize” the matrix $\mathbf{M}(\xi_N)$ or equivalently “minimize” the matrix $\mathbf{D}(\xi_N)$.

This relationship can be shown by considering the following argument. In the case of a single parameter θ , one can construct an estimator $\hat{\theta}_N$ which is approximately normal with mean θ_t and variance V_N . The ratio $\frac{\hat{\theta}_N - \theta_t}{\sqrt{V_N}}$ is approximately standard normal and hence the approximate confidence interval with coverage probability $1 - \alpha$ is:

$$CI_{1-\alpha} = \left\{ \theta : \frac{|\theta - \theta_N|}{\sqrt{V_N}} \leq z_{\alpha/2} \right\}, \quad (1.34)$$

with $z_{\alpha/2}$ representing the $100(1 - \alpha/2)$ percentile of the standard normal distribution. So in the case of:

$$y_i = \theta_t + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad (1.35)$$

with σ^2 known, the sample mean $\hat{\theta}_N = \sum_{i=1}^N \frac{y_i}{N} \sim \mathcal{N}(\theta_t, \frac{\sigma^2}{N})$ has that:

$$CI_{1-\alpha, norm} = \left\{ \theta : \frac{|\theta - \theta_N|}{\sigma/\sqrt{N}} \leq z_{\alpha/2} \right\}. \quad (1.36)$$

Since the square of the standard normal random variable is equivalent to χ_1^2 , $z_{\alpha/2}^2 = \chi_{1, \alpha}^2$ where $\chi_{1, \alpha}^2$ is the $100(1 - \alpha)$ percentile of the χ_1^2 distribution. Hence:

$$\frac{(\theta - \hat{\theta}_N)^2}{\sigma^2/N} = N M (\theta - \hat{\theta}_N)^2 \sim \chi_1^2, \quad (1.37)$$

where $M = \sigma^{-2}$ is the Fisher information of the random variable y_i . Therefore an equivalent confidence interval to that of (1.36) is:

$$CI_{1-\alpha, \chi^2} = \left\{ \theta : N M (\theta - \hat{\theta}_N)^2 \leq \chi_{1, \alpha}^2 \right\}. \quad (1.38)$$

Extending to the case of an m -dimensional parameter θ , the direct analog of (1.38) is:

$$CI_{1-\alpha, \chi^2} = \left\{ \theta : (\theta - \hat{\theta}_N)^T \mathbf{M}(\xi_N) (\theta - \hat{\theta}_N) \leq \chi_{m, \alpha}^2 \right\}, \quad (1.39)$$

with $\chi_{m, \alpha}^2$ representing the $100(1 - \alpha)$ percentile of the χ^2 distribution with m degrees of freedom. Thus the ellipsoid of concentration is defined by the boundaries of this confidence region.

The general optimization problem is defined as finding the solution:

$$\xi_N^* = \{x_i^*, p_i^*\}_1^{n^*} = \arg \min_{x_i, p_i, n} \Psi [\mathbf{M}(\{x_i, p_i\}_1^n)], \quad (1.40)$$

where Ψ is a scalar known as the criterion of optimality . The possible solutions are any combinations of n support points out of the available choices ($x_i \in \mathfrak{X}$) and the number of replications r_i at x_i such that $\sum_{i=1}^n r_i = N$ and $n \leq N$. Thus this is a discrete optimization problem with respect to the frequencies r_i or equivalently the weights $p_i = r_i/N$

The most popular types of optimality criteria are described by Fedorov & Leonov²⁸ as the following:

- D -optimality:

$$\Psi = |\mathbf{D}| = |\mathbf{M}|^{-1}. \quad (1.41)$$

Often called the generalized variance criterion, this D -criterion seems a reasonable measure of the “size” of the ellipsoid of concentration defined in (1.33) because $|\mathbf{D}|^{\frac{1}{2}}$ is proportional to the volume of the ellipsoid:

$$\text{Volume} = V(m) R^m |\mathbf{D}|^{\frac{1}{2}}, \quad \text{where } V(m) = \frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2} + 1)}, \quad (1.42)$$

with R defined in (1.33) and $\Gamma(u)$ defined as the Gamma function.

- *E*-optimality:

$$\Psi = \lambda_{\min}^{-1} [\mathbf{M}] = \lambda_{\max} [\mathbf{D}], \quad (1.43)$$

where $\lambda_{\min} [\mathbf{B}]$ and $\lambda_{\max} [\mathbf{B}]$ are minimal and maximal eigenvalues of the matrix \mathbf{B} respectively. The length of the principal axis of the ellipsoid of concentration is $2\lambda_{\max}^{\frac{1}{2}} [\mathbf{D}]$, hence minimization of this *E*-criterion also leads to the reduction of the linear “size” of the ellipsoid.

- *A*- or linear optimality:

$$\Psi = \text{tr}[\mathbf{A}\mathbf{D}], \quad (1.44)$$

where \mathbf{A} is an $m \times n$ non-negative definite matrix known as a utility matrix and $\text{tr}[\mathbf{B}]$ is the trace of the matrix \mathbf{B} . For example if $\mathbf{A} = m^{-1}\mathbf{I}_m$ where \mathbf{I}_m is the $m \times m$ identity matrix, then the *A*-criterion is based on the average variance of the parameter estimates:

$$\Psi = m^{-1}\text{tr}[\mathbf{D}] = m^{-1}\text{tr}[\mathbf{M}^{-1}] = m^{-1} \sum_{i=1}^m \text{Var}(\hat{\theta}_i). \quad (1.45)$$

D-optimality tends to be the most popular criterion used by theoretical and applied researchers since *D*-optimal designs are invariant with respect to nondegenerate transformations of parameters, e.g. changes in the parameter scale. Also, they perform well according to other optimality criteria.²⁷

In order to compare two designs, one can use their relative *D*-efficiency:⁵

$$\text{Eff}_D(\xi_{N,1}, \xi_{N,2}) = \left[\frac{|\mathbf{M}(\xi_{N,1})|}{|\mathbf{M}(\xi_{N,2})|} \right]^{1/m}. \quad (1.46)$$

If \mathbf{M} is a diagonal matrix and $\mathbf{D} = \mathbf{M}^{-1}$ then

$$|\mathbf{M}|^{-1/m} = \left(\prod_{i=1}^m m_{ii} \right)^{-1/m} = \left(\prod_{i=1}^m d_{ii} \right)^{1/m}, \quad (1.47)$$

where m_{ii} and d_{ii} are the *i*th diagonal elements of the matrices \mathbf{M} and \mathbf{D} respectively. Then $|\mathbf{M}|^{-1/m}$ is the geometric mean of parameter variances. The further the relative *D* efficiency is from one, the more the loss of precision of parameters from the better design to the other.

1.5.2 Cost-based Designs

It is often the case that there is an associated cost with a given design. When the costs are different for the set of designs considered, this can be included in the following way.

The general optimization problem denoted in (1.40) may be viewed as an approximation to the following optimization problem:

$$\xi_N^* = \arg \min_{\xi_N} \Psi [NM(\xi_N)], \quad (1.48)$$

subject to

$$\sum_{i=1}^n r_i = N \leq N^*, \quad (1.49)$$

It is described by Fedorov & Leonov²⁸ that when measurements at a point \mathbf{x} are associated with some penalty or cost denoted as $\phi(\mathbf{x})$, the constraint of (1.49) may be replaced by:

$$\Phi(\xi_N) = \sum_{i=1}^n r_i \phi(\mathbf{x}_i) \leq \Phi^*, \quad (1.50)$$

where Φ^* is the constraint on the total cost. Equivalently in the continuous setting of the p_i 's:

$$N\Phi(\xi_N) \leq \Phi^*, \text{ where } \Phi(\xi_N) = \sum_{i=1}^n p_i \phi(\mathbf{x}_i). \quad (1.51)$$

This optimization problem, due to the monotonicity and homogeneity of criteria Φ , can therefore be written as:

$$\xi^* = \arg \min_{\xi} \Phi \left[\frac{\Phi^*}{\Phi(\xi)} M(\xi) \right] = \arg \min_{\xi} \Phi \left[\frac{M(\xi)}{\Phi(\xi)} \right]. \quad (1.52)$$

1.5.3 Application to PK/PD Studies

The standard model, slightly adapting the previous notation introduced in (1.27), used for observations in PK sampling is

$$y_{ij} = \eta(x_{ij}, \gamma_i) + \epsilon_{ij}, \quad i = 1, \dots, N; \quad j = 1, \dots, k_i, \quad (1.53)$$

where x_{ij} is the j th sampling time for subject i , k_i is the total number of measurements for subject i . γ_i is the vector of individual parameters for subject i . N is the total number of subjects in the study. The ϵ_{ij} are the measurement errors with zero mean.

In the model based approach to finding sampling sequences such as that described by Gagnon & Leonov,³⁰ the key is to find a closed form expression or approximation to the information matrix for a single multidimensional point \mathbf{x}_i , that is the $(k_i \times 1)$ vector of responses for subject i . A design region \mathfrak{X} must be defined, and then the construction of optimal designs is relatively straightforward using for example D -optimality.

As an illustration, Gagnon & Leonov³⁰ use the design region \mathfrak{X} formed by the combinations of r sampling times from a sampling sequence of 16 choices, and the cost function given by:

$$\phi(\mathbf{x}_k) = C_v + C_s k, \quad (1.54)$$

where k is the number of samples taken (the length of the sampling sequence), C_s is the cost of collecting/analysing a single sample and C_v is the cost of enrolling a single subject. Gagnon & Leonov³⁰ show that when $C_s > 0$ then it is possible for the sequences with smaller samples to become optimal. It is also noted that optimal designs may comprise of a mixture of sequences with different numbers of samples. It is important to realise that although this cost function refers to monetary cost, that is not necessarily the definition of the cost function. The 'cost' may be in terms of time, or stress to the animals, or even a statistical concept such as bias that is wished to be penalized.

It is often the case in early stages of drug development that non-compartmental analysis is preferred to the model based compartmental approach. This means that PK parameters such as AUC , C_{max} and t_{max} are estimated using an empirical approach. Fedorov & Leonov²⁶ use the model based approach as a benchmark to compare proposed empirical methods to in terms of choosing a sampling se-

quence.

The simplest case is considered first, in which all subjects have the same sampling schedule. That is, $x_{ij} \equiv x_j$ and $k_i \equiv k$ for all $i = 1, \dots, N$. Two types of empirical approaches are discussed by Fedorov & Leonov:²⁶

- Type I: Method E1 For each subject, find the individual t_{max} (\hat{T}_i) and C_{max} (\hat{C}_i):

$$\hat{T}_i = x_{j^*(i)}, \text{ where } j^*(i) = \arg \max_j y_{ij}, \hat{C}_i = y_{i,j^*(i)}. \quad (1.55)$$

To find \widehat{AUC}_i , numerical methods are used, then averaged over all subjects in the study, either using arithmetic or geometric means to obtain populations estimators \widehat{AUC}_{E1} , \hat{T}_{E1} and \hat{C}_{E1} . (The subscript E stands for empirical). It is noted that for large N this method will produce reasonable estimators, however in the case of sparse sampling, the next method may be more appropriate.

- Type II: Method E2 At each time point, average the response over all patients:

$$\hat{\eta}_j = \hat{\eta}_{jN} = \frac{1}{N} \sum_{i=1}^N y_{ij}, \quad j = 0, 1, \dots, k, \quad (1.56)$$

and build estimators for the population curve $\{\hat{\eta}_j\}$:

$$\hat{T}_{E2} = x_{j^*}, \quad \hat{C}_{E2} = \hat{\eta}_{j^*}, \quad \text{where } j^* = \arg \max_j \hat{\eta}_j. \quad (1.57)$$

Note that geometric means may also be used in place of these arithmetic means. Then numerical integration algorithms must be used to estimate AUC :

$$\widehat{AUC}_{E2} = \sum_{j=1}^k \int_{x_{j-1}}^{x_j} g(x, \mathbf{a}_j) dx, \quad (1.58)$$

where g is an interpolating function with parameters \mathbf{a}_j chosen such that g passes exactly through $\hat{\eta}_{j-1}$ and $\hat{\eta}_j$. These are expected to provide good estimators for the parameters for large N , but also applicable in the case of sparse sampling.

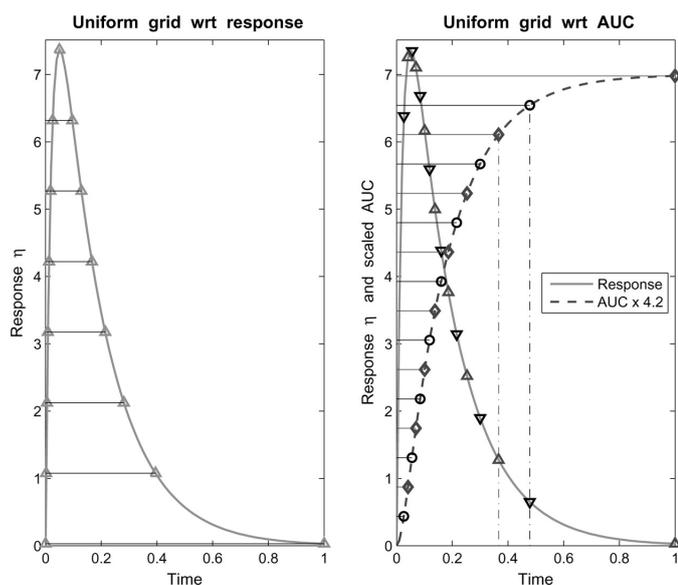


Figure 1.12: Two different sampling grids: Uniform with respect to response (left) and uniform with respect to AUC (right).²⁶

The choice of sampling grid, that is the possible time points that samples can be taken, is often influenced by prior estimates of the plasma concentration curve. It is often the case that more samples are taken towards the start of the sampling interval and then the frequency of sampling decreases after the anticipated t_{max} . Two types of sampling grid are proposed by Fedorov & Leonov,²⁶ as shown in Figure 1.12. The left hand panel shows a uniform grid on the vertical axis with respect to values of the response function, and then points are projected onto the horizontal axis for sampling times. The right hand panel shows a uniform grid on the vertical axis with respect to the accumulated AUC value, and then points are projected onto the horizontal axis for sampling times.

Of course this requires preliminary knowledge of the plasma concentration curve, but it is noted that so do traditional sampling schemes. It is especially common, as previously mentioned, to have a higher frequency of samples around the anticipated t_{max} , which may be different for different trials.

Fedorov & Leonov²⁶ conduct simulation studies to compare the model based approach, the empirical approach and an approach using splines to ap-

proximate the plasma concentration curve. This comparison is in terms of bias and variability of the three PK parameters previously discussed. The model based approach is found to perform best although the empirical approach also performs well.

A splitting of the sampling grid is also suggested, for example:

- Denote the single grid with $2k$ sampling points as $\{x_j : j = 1, 2, \dots, 2k\}$.
- Collect samples for the first $N/2$ subjects at times $\{x_{2j-1} : j = 1, \dots, k\}$.
- Collect samples from the second half of the study cohort at $\{x_{2j} : j = 1, \dots, k\}$.
- Using the method E2 for the estimation of AUC_2 , average the responses from each half of the cohort separately and then combine the series to obtain a population curve $\{\hat{\eta}_j\}$ and estimate the AUC using (1.58).

Again, simulations are conducted and it is found that the split-grid approach results in a rather small loss of precision compared to the single grid approach.

The mean squared error (MSE) of the estimate of AUC for the single and split grids are compared using the empirical approach with the trapezium rule. In this case the response is approximated by the 2nd order polynomial with random intercept:

$$\eta(x_j, \gamma_i) = \gamma_{0i} + \gamma_1 x_j + \gamma_2 x_j^2, \quad (1.59)$$

where $\gamma_i = (\gamma_{0i}, \gamma_1, \gamma_2)$, $\mathbb{E}(\gamma_{0i}) = \gamma_0$ and $\text{Var}(\gamma_{0i}) = u^2$. Let the MSE of the \hat{AUC}_{E2} for the single grid be MSE_1 and for the split grid be MSE_2 . Let $Bias_r$ and Var_r be the corresponding bias and variance terms in the following expression for MSE_r :

$$MSE_r = Bias_r^2 + Var_r, \quad r = 1, 2. \quad (1.60)$$

Assume a uniform grid $\{x_j = jT/(2k), j = 0, 1, \dots, 2k\}$ and without loss of generality assume $T = 1$. It is shown by Fedorov & Leonov²⁶ that for the single grid,

if the ϵ_{ij} are assumed to have variance σ^2 :

$$Bias_1 = \frac{\gamma_2}{6} \frac{1}{4k^2}, \quad Var_1 = \frac{1}{N} \left[\frac{\sigma^2(2k - 0.5)}{4k^2} + u^2 \right] \sim \frac{\sigma^2}{2Nk} + \frac{u^2}{N}, \quad (1.61)$$

and for the split grid:

$$Bias_2 \equiv Bias_1, \quad Var_2 = \frac{(2k - 1.5)2\sigma^2}{4k^2N} + \frac{u^2}{N} \left(1 - \frac{1}{2k^2} \right) \sim \frac{\sigma^2}{Nk} + \frac{u^2}{N}. \quad (1.62)$$

It follows from this that the measurement variability for the split grid is double that of the single grid, unsurprising since the number of samples taken is halved. However, the population component in the variance and the bias are the same for both grids. Hence when the population variance u^2 dominates the measurement variance σ^2 , then $MSE_1 \approx MSE_2$, and this is in spite the number of samples being halved. However the single grid will always be better, and this difference depends on the values of η'' , σ^2 and u^2 . If costs are introduced however, this may not be the case.

Continue the notation introduced in (1.54) for costs, and let C_{total} be the upper bound for the study budget. Then the overall costs are expressed as:

$$2kNC_s + NC_v \leq C_{total} \quad \text{for the single grid,} \quad (1.63)$$

$$kNC_s + NC_v \leq C_{total} \quad \text{for the split grid.} \quad (1.64)$$

This may be approached by either selecting N and finding the maximal admissible $k = k(N, C_{total})$, or to select k and find the maximal $N = N(k, C_{total})$. Note that the values of k and N are not independent. A simulation study is conducted by Fedorov & Leonov,²⁶ with the MSE of the estimate of AUC from this study plotted in Figure 1.13, showing that the split grid may outperform the single grid in terms of this MSE.

These results highlight the importance of considering split sampling grids. Although this work primarily focused on batch designs, (see Section 1.3.3), the

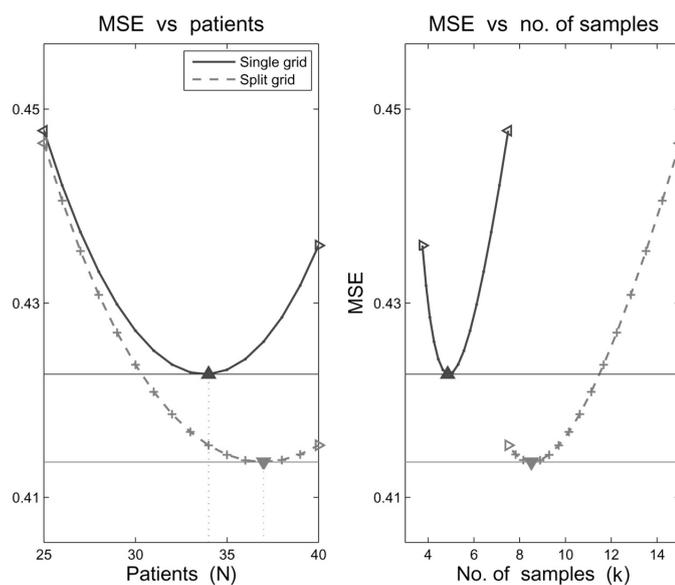


Figure 1.13: MSE as a function of N and k for $u = 2.4$, $\sigma = 9$ and $25 \leq N \leq 40^{26}$

principles discussed can be extended to both flexible and alternative sparse sampling schemes. Using the optimality criteria of the MSE is one way of including both the bias and the variance of the estimator of the AUC in the choice of optimal design. It does however raise questions concerning the optimal way to split the sampling grid. In fact, since in many PK studies, the total number of subjects N is fixed, the choice of sampling times and the flexibility for different subjects to be on different schedules gives rise to more complexity than can be incorporated in the analytical form of the MSE previously given. An advantage of this approach however is that it considers individual subjects having individual parameters. The following section also builds upon this assumption, although allowing for more flexible designs. This is however at the cost of using the PK modelling framework for estimation of PK parameters.

1.5.4 D-Optimality for multiple response non-linear mixed effect models

Since D-optimality is one of the most popular choices for optimality, much has been done in the area of development of methods for its application. One of the most notable perhaps, is the approximation of the Fisher Information Matrix for non-linear mixed effect models for use in PK/PD optimal trial designs by Retout et al.^{62,63} This can be applied in the software package PFIM,⁹ the theory behind which is explained in this section.

As this is within the PK modelling framework, a model must be assumed:

$$y_i = f(\theta_i, \xi_i) + \epsilon_i,$$

where:

- \mathbf{y}_i is the n_i -vector of observations of the i th individual amongst N .
- f is the known function describing the non-linear structural model.
- $\boldsymbol{\xi}_i = (t_{i1}, t_{i2}, \dots, t_{in_i})^T$ is the n_i vector of sampling times for individual i .
- $\boldsymbol{\theta}_i$ is the p -vector of individual parameters.
- $\boldsymbol{\epsilon}_i$ is the n_i -vector of random errors with $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$
- $\boldsymbol{\Sigma}_i$ are assumed to be $n_i \times n_i$ diagonal matrices.
- The model for the variance of the j th observation of individual i is either a constant variance model $\text{var}(\epsilon_{ij}) = \sigma^2$ or a constant coefficient of variation model $\text{var}(\epsilon_{ij}) = \sigma^2 f^2(\boldsymbol{\theta}_i, t_{ij})$.

For the mixed effects, let \mathbf{b}_i be the p -vector of random effects for individual i , and $\boldsymbol{\beta}$ be the p -vector of fixed effects. Then the expression for the inter-individual variability is given by $\boldsymbol{\theta}_i = \boldsymbol{\beta} + \mathbf{b}_i$ if the modelization of random effects is additive, or $\boldsymbol{\theta}_i = \boldsymbol{\beta} e^{\mathbf{b}_i}$ if the modelization is exponential.

Then we can rewrite $f(\boldsymbol{\theta}_i, \boldsymbol{\xi})$ as $f(\boldsymbol{\beta}, \mathbf{b}_i, \boldsymbol{\xi})$. The random effects \mathbf{b}_i are assumed to be distributed $\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\Omega})$, with $\boldsymbol{\Omega}$ a $p \times p$ diagonal matrix, each diagonal element ω_k the variance of the k th component of the random effects vector. The $\epsilon_i | \mathbf{b}_i$ for $i = 1, \dots, N$ are assumed to be independent from one subject to another, and for each subject, ϵ_i and \mathbf{b}_i are independent. The vector of population parameter is then given as $\boldsymbol{\Psi}^T = (\boldsymbol{\beta}^T, \omega_1, \dots, \omega_k, \sigma^2)$.

1.5.4.1 Individual Designs

Firstly, the expression for the elementary Fisher Information Matrix is developed, which corresponds to one individual with a vector of sampling times $\boldsymbol{\xi}$. The index i is therefore dropped in the following. The Fisher Information Matrix is defined as:

$$M_F(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \mathbb{E} \left(-\frac{\partial^2 \ell(\boldsymbol{\Psi}; \mathbf{y})}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T} \right)$$

where $\ell(\boldsymbol{\Psi}; \mathbf{y})$ is the log-likelihood of the vector of observation \mathbf{y} for the population parameters $\boldsymbol{\Psi}$.

The first case considered is the constant residual error variance model. Because of the non-linearity of the structural model f with respect to $\boldsymbol{\theta}$ in PK settings, there is no analytical expression for $\ell(\boldsymbol{\Psi}; \mathbf{y})$. Therefore a first order Taylor approximation of the structural model $f(\boldsymbol{\beta}, \mathbf{b}_i, \boldsymbol{\xi})$ around $\mathbf{0}$ (the expectation of \mathbf{b}) may be used:

$$y = f(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\xi}) + \epsilon \cong f(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi}) + \frac{\partial f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial \mathbf{b}} \mathbf{b} + \epsilon.$$

So that

$$\mathbb{E}(\mathbf{y}) \cong f(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})$$

$$\text{var}(\mathbf{y}) = \mathbf{V} \cong \frac{\partial f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial \mathbf{b}} \boldsymbol{\omega} \frac{\partial f(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial \mathbf{b}} + \sigma^2 \mathbf{I}_n,$$

where \mathbf{I}_n is the $n \times n$ identity matrix.

The log-likelihood is then approximated by:

$$\ell(\Psi; \mathbf{y}) \cong -\frac{1}{2} \left(n \log(2\pi) + \log |\mathbf{V}| + (\mathbf{y} - f(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi}))^T \mathbf{V}^{-1} (\mathbf{y} - f(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})) \right),$$

since \mathbf{b} and $\boldsymbol{\epsilon}$ are assumed to be normally distributed.

For the further notation, the population parameters are partitioned into mean and variance parameters, denoting the variance parameters $\boldsymbol{\lambda} = (\omega_1, \omega_2, \dots, \omega_p, \sigma^2)$, hence $\Psi^T = (\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)$. It is assumed that \mathbf{V} is independent of $\boldsymbol{\beta}$, and using the following:

$$\frac{\partial \log |\mathbf{V}|}{\partial \lambda_k} = \text{tr} \left(\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \lambda_k} \right) \quad \text{and} \quad \frac{\partial \mathbf{V}^{-1}}{\partial \lambda_k} = -\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \lambda_k} \mathbf{V}^{-1},$$

Second derivatives of minus twice the log-likelihoods can then be approximated.

$$\mathbb{E} \left(\frac{\partial^2 (-2\ell(\Psi; \mathbf{y}))}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right) \cong 2 \frac{\partial f^T(\boldsymbol{\beta}, 0, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} \mathbf{V}^{-1} \frac{\partial f(\boldsymbol{\beta}, 0, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}}.$$

$$\mathbb{E} \left(\frac{\partial^2 (-2\ell(\Psi; \mathbf{y}))}{\partial \boldsymbol{\beta} \partial \lambda_k} \right) \cong 0 \quad \text{for } k \text{ in } \{1, \dots, p+1\}.$$

$$\mathbb{E} \left(\frac{\partial^2 (-2\ell(\Psi; \mathbf{y}))}{\partial \lambda_k \partial \lambda_j} \right) \cong \text{tr} \left(\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \lambda_j} \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \lambda_k} \right) \quad \text{for } k \text{ and } j \text{ in } \{1, \dots, p+1\}.$$

The elementary Fisher Information Matrix is thus approximated by a block diagonal matrix as follows:

$$M_F(\Psi, \boldsymbol{\xi}) \cong \begin{bmatrix} \frac{\partial f^T(\boldsymbol{\beta}, 0, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} \mathbf{V}^{-1} \frac{\partial f(\boldsymbol{\beta}, 0, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} & 0 \\ 0 & \frac{1}{2} \mathbf{F} \end{bmatrix} \quad (1.65)$$

$$= \begin{bmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{bmatrix}, \quad (1.66)$$

where the elements of \mathbf{F} are defined as:

$$F_{jk} = \left(\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \lambda_j} \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \lambda_k} \right).$$

The block \mathbf{A} is the $p \times p$ symmetric matrix for fixed effects, and the block $\mathbf{B} = \frac{1}{2}\mathbf{F}$ is a $(p+1) \times (p+1)$ symmetric matrix for the variances. The elements of \mathbf{F} may be simplified as: (i) for j and k in $\{1, \dots, p\}$:

$$F_{jk} = \left(\frac{\partial f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial b_j} \mathbf{V}^{-1} \frac{\partial f(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial b_k} \right)^2$$

(ii) for k in $\{1, \dots, p\}$:

$$F_{(p+1)k} = \left(\frac{\partial f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial b_k} \mathbf{V}^{-2} \frac{\partial f(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial b_k} \right)^2$$

(iii)

$$F_{(p+1)(p+1)} = \text{tr}(\mathbf{V}^{-2})$$

When considering the case where the variance of the ϵ_i has a constant coefficient of variation, it can be approximated to a constant variance error using the log of the observations and of the model. In calculations this leads to the replacement of

$$\frac{\partial f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} \quad \text{by} \quad \frac{\partial f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} \times \frac{1}{f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})},$$

and

$$\frac{\partial f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial b_j} \quad \text{by} \quad \frac{\partial f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial b_j} \times \frac{1}{f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}.$$

These are equivalent to:

$$\frac{\partial \log f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} \quad \text{and} \quad \frac{\partial \log f^T(\boldsymbol{\beta}, \mathbf{0}, \boldsymbol{\xi})}{\partial b_j} \quad \text{for} \quad j = \{1, \dots, p+1\}.$$

1.5.4.2 Population Designs

So far, these approximations have been for an individual design. For a population design $\Xi = \{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N\}$, the Fisher Information Matrix $M_F(\boldsymbol{\Psi}, \Xi)$ is the sum of the N elementary Fisher Information Matrices $M_F(\boldsymbol{\Psi}, \boldsymbol{\xi}_i)$ for each subject i with design $\boldsymbol{\xi}_i$:

$$M_F(\boldsymbol{\Psi}, \Xi) = \sum_{i=1}^N M_F(\boldsymbol{\Psi}, \boldsymbol{\xi}_i). \quad (1.67)$$

The population design is usually composed of a limited number of Q groups of elementary designs. Each of these designs q is composed of a set ξ_q of n_q sampling times and is performed in N_q subjects. The elementary Fisher Information Matrices are identical for all subjects in a given design group q . Then the Fisher Information Matrix for this population design Ξ is:

$$M_F(\Psi, \Xi) = \sum_{q=1}^Q N_q M_F(\Psi, \xi_q). \quad (1.68)$$

Then the expected values of the standard errors for each population parameter are computed as the square root of the diagonal elements of the inverse of $M_F(\Psi, \Xi)$. These values are the lower bound of the standard errors of parameter estimation from the Cramer-Rao inequality.

A population design Ξ is D -optimal for a given population parameter value Ψ^0 if it minimizes the inverse of the determinant of the Fisher Information Matrix:

$$\Xi_D = \arg \min_{\Xi} \frac{1}{|M_F(\Psi^0, \Xi)|}.$$

1.5.4.3 Comparing Efficiency Between Designs

In order to compare efficiency between designs, a criterion, ϕ , the determinant standardized by the dimension of the vector Ψ^0 :

$$\phi(\Xi) = |M_F(\Psi^0, \Xi)|^{1/\dim(\Psi^0)}.$$

Then the efficiency of a population design Ξ_1 with respect to another population design Ξ_2 is:

$$\frac{\phi(\Xi_1)}{\phi(\Xi_2)}. \quad (1.69)$$

If this ratio is greater than 1, it indicates Ξ_1 is more efficient than Ξ_2 . This ratio is the factor of mean estimation of variance decrease of using Ξ_1 instead of Ξ_2 .

It is also the case that if Ξ_1 and Ξ_2 are identical apart from the number of subjects in each group of Ξ_2 is multiplied by a constant p , then the ratio in 1.69 is equal to $\frac{1}{p}$, indicating design Ξ_2 is p times more efficient than Ξ_1 .

It is this approach by Retout et al^{62,63} to which we compare the method in Chapter 4, since this is a well established procedure for finding optimal designs for PK studies.

1.6 Measurements that Cannot be Reliably Detected

Any analytic procedure for measuring the concentration of an analyte in a sample of bodily tissue has limitations. The main limitation is that samples of low concentration cannot be reliably detected. The following section details definitions associated with these limitations and methods for dealing with them.

1.6.1 Definitions

There are three defined limits of at what level a low concentration is too low to be reliably detected:

- Limit of Blank (LOB),
- Limit of Detection (LOD),
- Limit of Quantification (LOQ).

These limits are defined based on calibration of the measurement process using blank samples and low concentration samples.^{4,19} Figure 1.14 shows the relationship between these limits and how they are calculated. When replicate blank samples (samples containing no analyte) are tested, they can produce signals that indicate there is a low concentration of analyte in the sample. Since equipment will not output a negative value, there is an artificially increased frequency of samples at 0 concentration. The mean and standard deviation of these results is calculated and the LOB is then defined as:

$$\text{LOB} = \text{mean}_{\text{blank}} + 1.645 * (\text{SD}_{\text{blank}}).$$

This is set to control the type *I* error (α) at 5%. Therefore under the assumption of a Gaussian distribution of raw analytical signals from blank samples, this limit represents 95% of the observed values. The other 5% produce a signal that could be produced by a sample of low concentration analyte. This in turn affects the type *II* error (β), the proportion of low concentration samples that will falsely be reported as blank.

Replicate samples that truly contain a known low concentration of analyte are used to calibrate for the LOD. The mean and standard deviation of the results from these replicates are used in the following way to provide a provisional LOD:

$$\text{LOD} = \text{LOB} + 1.645 * (\text{SD}_{\text{lowconcentration sample}}).$$

With this definition, 95% of the low concentration replicates exceed the previously defined LOB, and 5% below the LOB. Once this provisional LOD is established, this value may be confirmed using samples with a known concentration equal to the provisional LOD. No more than 5% of these samples should give measurements below the LOB. If more than 5% are below this value then the LOD is too low and must be re-estimated by testing samples that contain a higher concentration of the analyte.

The LOQ is set to be the lowest concentration that the true concentration of the analyte can be reliably detected and also at a level that the predefined goals of bias and precision are achieved. The value of the LOQ may be more than or equal to the LOD. If the observed levels of bias and imprecision meet the predefined goals at the LOD then the value of LOQ equals that of the LOD. If these measurements do not meet the specific requirements for the total error, then samples with higher concentration of analyte must be used to determine the LOQ.

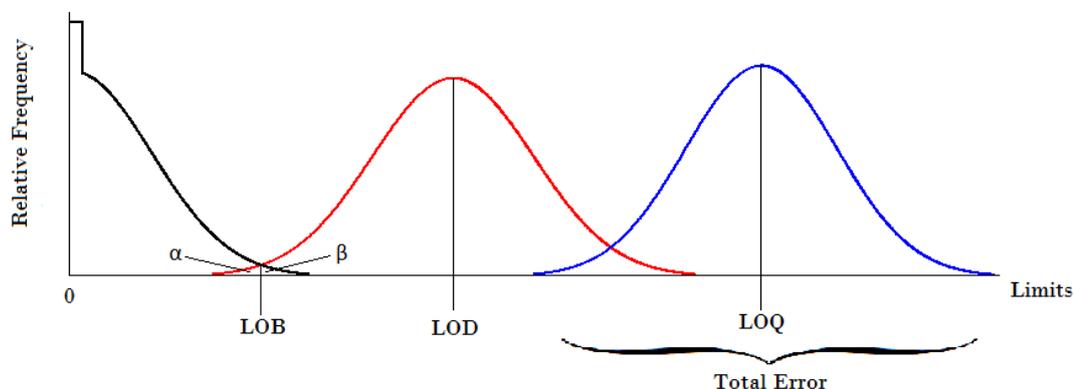


Figure 1.14: An illustration of the relationship between LOB, LOD, and LOQ, indicating relative frequencies of raw analytical signals at varying concentrations. The black line represents results from a blank sample. The red line represents the imprecision of results from a sample with a prespecified low concentration. The blue line represents the distribution of results for a specimen of low concentration meeting the target for total error (bias and imprecision).

1.6.2 Methods

Since any sample with a resulting measure of concentration below the limit of quantification will be reported as "BLOQ", and no numerical result given, methods must be established for the case when these BLOQ values are observed. Beal¹⁰ compares seven methods for dealing with BLOQ measurements in PK studies that use PK modelling as the method of analysis. Hing et al³⁵ also tackle the subject of dealing with BLOQ values from a PK modelling perspective, with applications in NONMEM. Further to these, Byon et al.¹⁴ and Ahn et al.¹ look at the impact on the PK model that using likelihood based methods has for scenarios with BLOQ responses. For each of these methods to be used, a PK model must be specified beforehand in order for it to be fitted. The seven methods proposed by Beal are as follows:

- **Method 1:** Discard BLOQ observations and apply extended least squares to the remaining observations.

- **Method 2:** Discard BLOQ observations and apply method of maximum conditional likelihood to the remaining observations.
- **Method 3:** Maximise the likelihood for all the data treating BLOQ data as censored.
- **Method 4:** As Method 4, but the likelihoods for data above and below the LOQ are conditioned on the observations being greater than 0.
- **Method 5:** Replace BLOQ observations with $LOQ/2$ and apply extended least squares estimation.
- **Method 6:** Replace the first BLOQ observation with $LOQ/2$ and discard the rest of them as in Method 1.
- **Method 7:** Replace the first BLOQ observation with 0 and discard the rest of them as in Method 1.

When applied to simulated data from a one-compartmental oral dose model in both single subject and population studies, it was found that Method 7 performed poorly and hence is not recommended for practical use¹⁰. Method 1 is suggested as a simple method for situations with low frequencies of BLOQ responses, however for larger frequencies in the single subject case, Method 5 is preferable. Methods 2, 3 and 4, whilst more sophisticated than Methods 5 and 6, do not show improvement on the simpler alternatives that Methods 5 and 6 offer. In fact, Beal suggests that there is little difference between the first four methods and hence propose simply using Method 1.

Whilst the procedure of discarding BLOQ observations may be the most favourable choice in this case, PK studies generally consider groups of subjects with different levels of dosing. Subjects with a lower dose are more likely to have higher frequencies of BLOQ observations and hence more values being discarded. Discarding a larger number of values will not only lead to biased

estimates of parameter, but also a vast overestimation of any variances. This is even more accentuated in the case of studies with smaller sample sizes, those tending to use NCA as opposed to PK modelling. Therefore the focus of Chapter 5 is on methods of dealing with BLOQ values in PK studies that use non-compartmental methods.

CHAPTER 2

Thesis Summary

This thesis concerns three separate but related research topics. These are developed in three academic papers labelled Paper A, B and C that comprise Chapters 3, 4 and 5 respectively. These three topics are all in the setting of PK studies, with the overall incentive of improving the accuracy of the PK outcome measures. Motivated by the novel blood sampling method of microsampling discussed in Section 1.2, these research topics also address the issues of the three Rs, focussing on refinement and reduction.

The first two papers are directly motivated by microsampling: the first paper developing a statistical method to detect equivalence between microsampling and traditional sampling in PK studies, and the second paper establishing a method for designing sparse sampling schedules in studies that use microsampling. The third paper introduces a method for dealing with BLOQ responses using kernel density estimation. The contents of each paper are summarized below.

Paper A: Comparing sampling methods for pharmacokinetic studies using model averaged derived parameters. Pharmacokinetic (PK) studies aim to study how a compound is absorbed, distributed, metabolised and excreted (ADME). The concentration of the compound in the blood or plasma is measured at different time points after administration and pharmacokinetic parameters such as the area under the curve (AUC) or maximum concentration (C_{max}) are derived from the resulting concentration time profile. In this paper we want to compare different methods for collecting concentration measurements (traditional sampling versus microsampling) on the basis of these derived parameters. We adjust and evaluate an existing method for testing superiority of multiple derived parameters that accounts for model uncertainty. We subsequently extend the approach to allow testing for equivalence. We motivate the methods through an illustrative example and evaluate the performance using simulations. The extensions show promising results for application to the desired setting. (*Ap-*

pendices A - E supplement Paper A.)

Paper B: Optimal Designs for Non-Compartmental Analysis of Pharmacokinetic Studies. In traditional toxicology trials, pharmacokinetics (PK) is investigated in the satellite group of animals, and the pharmacodynamics (PD) is investigated in the study group of animals. The new blood sampling method of microsampling opens up the opportunity to investigate both PK and PD in the same animals. To avoid excessive burden on the animals from the required blood sampling, sparse sampling schemes are typically utilized. Motivated by this application, this paper introduces a procedure to choose an optimal sparse sampling scheme and sampling time points using non-compartmental methods but which can be applied to further settings beyond this. We discuss how robust designs can be obtained and we apply and evaluate the approach to a range of scenarios to give an example of how it may be implemented. The results are compared to optimal designs for model based PK. (*Appendices F - G supplement Paper B.*)

Paper C: Methods for Non-Compartmental Pharmacokinetic Analysis with Observations below the Limit of Quantification. Pharmacokinetic (PK) studies are conducted to learn about the absorption, distribution, metabolism and excretion (ADME) processes of an externally administered compound by measuring its concentration in bodily tissue at a number of time points after administration. Two methods are available for this analysis, modelling and non-compartmental (NCA). When concentrations of the compound are low, they may be reported as below the limit of quantification (BLOQ). This paper compares seven methods for dealing with BLOQ responses in the NCA framework for estimating the area under the concentrations versus time curve (AUC). These include simple imputations that are currently used, maximum likelihood methods, and introducing an algorithm that uses kernel density estimation to impute values onto BLOQ responses. Performance is evaluated using simulations for a

range of scenarios. We find that the kernel based method performs well for most situations. (*Appendix H supplements Paper C.*)

CHAPTER 3

**Paper A: Comparing sampling methods for pharmacokinetic studies using
model averaged derived parameters**

3.1 Introduction

The purpose of this paper is to construct a method for comparing the traditional method of collecting concentration samples in pharmacokinetic (PK) studies and the recently developed method of microsampling.¹⁶ The two blood sampling methods differ in both the volume of blood collected and the method of analysis. The use of microsampling offers many economic and ethical advantages. The reduction in blood volume of samples not only allows for the possibility of elimination of satellite subjects in toxicokinetic studies, but also the opportunity to redesign the sampling scheme to further reduce the number of subjects needed in both these and pharmacokinetic/pharmacodynamic (PK/PD) studies. The nature of the comparison considered is both superiority and equivalence. While both approaches are possible, testing for equivalence is preferred for a study to provide evidence to support microsampling as a valid blood sampling method.

Since PK studies measure the absorption, distribution, metabolism and excretion (ADME) processes over time using multiple parameters derived from the estimated functional relationship, a simultaneous comparison of these multiple derived parameters between the two sampling methods will provide a comparison between the two methods themselves.

These parameters may be estimated using non-compartmental analysis^{42,43} or using compartmental methods such as fitting non-linear mixed effects models.²⁰ However, in the latter approach there may be uncertainty in the choice of model and hence the use of model averaging is an apt idea to incorporate this. Since there are multiple parameters to compare simultaneously, a multiplicity adjustment must be made. In addition, the variance of the derived parameter estimates in many cases cannot be directly calculated so we must rely on approximations. In order to incorporate these properties, we use the method described

by Jensen & Ritz⁴⁷ as a starting point. We then investigate and improve the performance of the method when testing for superiority before we adapt it for the more suitable case for our question of testing for equivalence. In this case, testing for equivalence is more relevant in order to give opportunity to find evidence that the two sampling methods give equivalent results.

The motivation behind such a procedure is to conduct a study comparing the two sampling methods, such as the following example. This study conducted by Janssen Pharmaceutica examines the PK profile a novel compound that is administered intravenously. Plasma concentrations are taken from the same 5 rats at 3 time points using both microsampling and traditional sampling, a total of 30 observations (Figure 3.1). Blood sampling and analysis was undertaken by the same analyst and all animals entered the study on the same day.

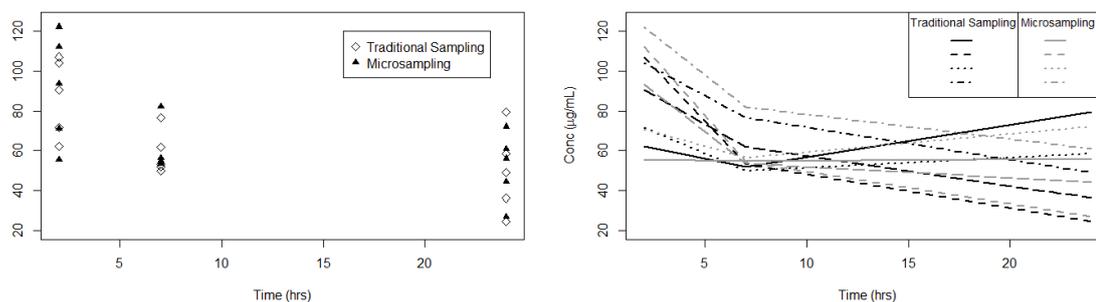


Figure 3.1: Example dataset with individual concentrations (left) and spaghetti plot (right).

3.2 Superiority Testing

3.2.1 Baseline Method

Jensen & Ritz⁴⁷ discuss simultaneous inference after model averaging parameters derived from a parametric function, specifically for the use of finding the derived parameters Bench Mark Dose (BMD) and the lower limit of the confidence interval for this (BMDL) in non-linear dose response modelling. A brief

outline of the method is as follows: A set of candidate models is fitted to the data and (approximate) estimates and variances of the derived parameters of interest are calculated. These are combined using model averaging¹² in order to account for uncertainty in the choice of model and simultaneous confidence intervals are calculated.⁵⁷ The purpose of using model averaging is to include multiple candidate models, thus taking into account model uncertainty and reducing the possibility of model mis-specification. Of course this introduces the implication that some of the set of candidate models must be incorrect, but we vastly improve our chances of including a good model by considering many candidates.

More specifically, let the K candidate models be parametrized by $\tau = (\tau_1, \dots, \tau_K)$ of dimensions (p_1, \dots, p_K) with $p = \sum_{k=1}^K p_k$. Consider the case where for model $k = 1, \dots, K$ L derived parameters are of interest: $\theta_{1k}, \dots, \theta_{Lk}$. It is assumed these are differentiable functions of the model parameters, so that $\theta_{1k} = f_{1k}(\tau_k), \dots, \theta_{Lk} = f_{Lk}(\tau_k)$. To obtain an overall estimator across the K models, the following weighted average of parameter estimates from each of the K candidate models is used⁴⁷ :

$$\hat{\theta}_{l,MA} = \sum_{k=1}^K w_k \hat{\theta}_{lk} = \sum_{k=1}^K w_k f_{lk}(\hat{\tau}_k),$$

where the w_k 's are the model specific weights such that $\sum_{k=1}^K w_k = 1$. Buckland et al.¹² suggest to use:

$$w_k = \frac{\exp(-I_k/2)}{\sum_{k=1}^K \exp(-I_k/2)},$$

where I_k is some information criterion based on model fit for candidate model k . For the following applications, we use these weights with information criterion AICc,¹³ although the method is valid for any reasonable specification of weights.

To conduct a test of superiority the multiple null hypotheses of interest are:

$$H_{0_1} : \theta_{1,MA} = \kappa_1, \dots, H_{0_l} : \theta_{l,MA} = \kappa_l, \dots, H_{0_L} : \theta_{L,MA} = \kappa_L,$$

where $\kappa_1, \dots, \kappa_L$ are the values the parameters are tested against. If κ_l is in

each corresponding confidence interval with simultaneous error rate α , we fail to reject the null hypothesis at simultaneous level α . If at least one κ_l is not in its corresponding interval, then the null hypothesis can be rejected. Hence the alternative hypothesis of this two-sided superiority test is that at least one of the $\theta_{l,MA}$ is not equal to the corresponding κ_l .

In order to calculate simultaneous confidence intervals for the multiple derived parameters with simultaneous coverage $1 - \alpha$, the methods of Pippert et al.,⁵⁷ described below, which depends on the p_k -dimensional asymptotic expansion of the maximum likelihood estimator $\hat{\tau}_k$ for each of the K models is used.

An asymptotic approximation for a single model-averaged estimate may be obtained by use of the delta method which uses a first order Taylor expansion to approximate the estimate for the model averaged parameter and its variance. The variance of the model-averaged estimate, $\hat{\theta}_{l,MA}$, is approximated as⁴⁷ :

$$\text{var}(\hat{\theta}_{l,MA}) \approx n^{-1} \left\{ \left(\frac{df_{lk}}{d\tau} \right) \mathbf{w} \right\}^T \hat{\Sigma} \left(\frac{df_{lk}}{d\tau} \right) \mathbf{w},$$

where $\mathbf{w} = (w_1, \dots, w_k)$ and $\frac{df_{lk}}{d\tau}$ is the $K \times p$ matrix $(\frac{d}{d\tau} f_{l1}(\hat{\tau}_1)^T, \dots, \frac{d}{d\tau} f_{lK}(\hat{\tau}_K)^T)$ for $l = 1, \dots, L$ and n is the number of observations. The covariance matrix $\hat{\Sigma} = n^{-1} \sum_{i=1}^n \hat{\Psi}_i^T \hat{\Psi}_i$ where the $\hat{\Psi}_i$ are estimates of $\Psi_i = \Psi_{1i}, \dots, \Psi_{Ki}$. $\Psi_{ki} = (I_k^{-1}) \tilde{\Psi}_{ki}$ with I_k^{-1} being the inverse Fisher information matrix for model k and $\tilde{\Psi}_{ki}$ the score function for observation i of model k , both evaluated at the parameter value τ_k .

To obtain $1 - \alpha$ simultaneous confidence intervals for all of the derived parameters, intervals of the form $\hat{\theta}_{l,MA} \pm Q \sqrt{\text{var}(\hat{\theta}_{l,MA})}$ are used. The quantile Q is calculated so that an L -dimensional random vector \mathbf{X} with standard multivariate z distribution with correlation structure the same as the parameter estimates' satisfies:

$$\mathbb{P}\left(\bigcap_{l=1}^L (-Q \leq X_l \leq Q)\right) = 1 - \alpha,$$

for a $100(1 - \alpha)\%$ confidence interval.

Jensen & Ritz⁴⁷ use a Normal distribution is used to calculate the quantiles for the simultaneous intervals, although the t -distribution may be more appropriate for the smaller sample sizes and estimations of variance using these small samples. This relies on an estimate of the degrees of freedom associated with the parameter estimators, which can be obtained from the residual degrees of freedom when fitting the model to the data; an estimate that depends on the number of data points and parameters in the fitted model.

3.2.1.1 Example

To illustrate the performance of the approach described above we begin by estimating PK parameters of interest, $t_{\frac{1}{2}}$, the time taken for the concentration to reach half its initial value and C_{max} , the maximum concentration, in this example, for both sampling methods in the example study data shown in Figure 3.1. For this dataset we will model the concentrations in a single model to ensure that the residual error variance is the same for both methods:

$$\mathbb{E}(Y_t) = (1 - \mathbb{I}_M)\mathbb{E}(g(\tau_S, t)) + \mathbb{I}_M\mathbb{E}(g(\tau_M, t)),$$

where \mathbb{I}_M is the indicator variable for the use of microsampling, τ_S and τ_M are the model parameters for standard sampling and microsampling respectively, t is time and g is the PK model, the same for standard and microsampling. Note, however, that such an assumption is not necessary in order to apply the method and one could equally use two separate models instead. An example using two separate models is discussed later.

To account for the uncertainty about the models form we consider the following two candidate models:

Candidate Model 3.1: Log Linear

$$g(\tau, t) = \beta_0^t \exp(\beta_1 t) + \epsilon. \quad (3.1)$$

Candidate Model 3.2: Log Log Linear

$$g(\tau, t) = \exp(\beta'_0 \exp(\beta_1 t)) + \epsilon. \quad (3.2)$$

We assume $\epsilon \sim N(0, \sigma^2)$ for model fitting. Based on these models we can then derive the analytical form of the PK parameters of interest as functions of the model parameters (See Appendix A for the full details). In order to compare the sampling methods, we are then interested in the differences between each parameter when using microsampling and standard sampling. These difference may be expressed as $t_{\frac{1}{2}}^{(S)} - t_{\frac{1}{2}}^{(M)}$ and $C_{max}^{(S)} - C_{max}^{(M)}$ where the superscripts S and M indicate standard sampling and microsampling respectively. For both candidate models, these derived parameters can be expressed as differentiable functions of the respective model parameters, and importantly have the same interpretation for each model. In this case, candidate model 1 has weight 0.461 and candidate model 2 has weight 0.539.

For this particular example, applying the method described by Jensen & Ritz⁴⁷ gives the 95% model averaged confidence intervals as (-38.2 , 37.1) and (-33.2 , 27.3) for $t_{\frac{1}{2}}^{(S)} - t_{\frac{1}{2}}^{(M)}$ and $C_{max}^{(S)} - C_{max}^{(M)}$ respectively. Since 0 is in both intervals, both fail to reject the null hypothesis suggesting that the two methods are similar. It is clear however that this does not show equivalence of the methods and it is perhaps worth observing that these intervals are rather wide, likely due to the small sample size of the study.

3.2.1.2 Simulation Studies

For a first evaluation of the performance of the method, we find the empirical overall type I error rate by simulating 10,000 datasets under the null hypothesis. The same derived parameters of interest and candidate models as used in the previous example are considered in these simulation studies. Data are generated from the same data generating model for both sampling methods, standard and microsampling. The data generating model used is the best model fit to the

example data as judged by the AICc¹³ which corresponds to model 3.2 and parameter values $\beta'_0 = 4.436733$, $\beta_1 = -0.006318$. An additive normally distributed error with $\sigma = 5$ is used in order to replicate the variation seen in the example dataset. For simplicity, we assume independent subjects at each time point for the time being, and return to longitudinal data at a later point.

Different numbers of subjects, $n = 5, 10, 100, 1000$, are simulated independently at each time point. The smaller sample sizes are more realistic in terms of conducting these studies but the larger sample sizes show the asymptotic behaviour of the method. Between 3 and 10 timepoints (see Table B.1 in Appendix B for details) are considered. For a true coverage of 95%, the estimate of familywise type I error rate is expected to be between 0.0457 and 0.0543 due to simulation error. Figure 3.2 shows how the number of time points and number of subjects affects the coverage of the simultaneous confidence intervals based on normal and t-quantiles. It is apparent that the method is conservative when small sample sizes are used while inflated type I error rates are observed for large sample sizes (and a large number of time points). The use of t quantile results in type I errors closer to target, but even there conservatism for small sample sizes and anticonservatism is seen for a large number of data points, as seen in Figure 3.2. In terms of bias of the model averaged derived parameters (shown in Table D.1 in Appendix D), we see as expected, larger bias for the smaller sample sizes and smaller number of time points. For example, we see an average bias of -0.06 for C_{max} for 1000 subjects at 10 time points, but an average bias of -0.32 for 5 subjects at 3 time points. Please refer to Table D.1 in Appendix D for the full range of values. All things considered, this amount of bias is not too concerning for the procedure.

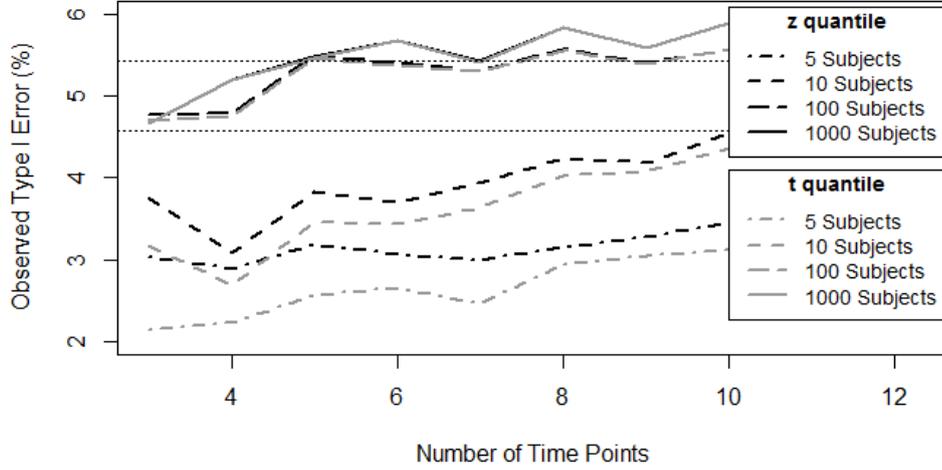


Figure 3.2: Comparison of observed type I error rate for varying number of time points and subjects for use of z and t quantile. Horizontal dotted lines show error bounds for 10,000 simulations

3.2.2 Extension

Based on the brief simulations above we are now interested in evaluating if a better approximation of the variance, the current method uses a first order delta method, can be used to improve the results of the method. To do so we consider a second order Taylor approximation to estimate the variance of a function of parameter estimates which can be found as follows: τ is a vector of model parameters with $\hat{\tau}$ the estimates from the model; $\mu = E[\hat{\tau}]$, a vector of the expectations; $\hat{\theta} = f(\hat{\tau})$, the derived parameter estimates as a differentiable function of the model parameter estimates (including the weights from the model averaging step); $\hat{\Sigma}(\hat{\tau})$ is the covariance matrix of $\hat{\tau}$; $\mathbf{D} = \partial_{\tau} f(\hat{\tau})$ is the gradient of $f(\tau)$ evaluated at μ and $\mathbf{H} = \partial_{\tau}^2 f(\hat{\tau})$ is the Hessian of $f(\tau)$ evaluated at μ .

Then a second order Taylor approximation of f is given by:

$$f(\hat{\tau}) \approx f(\mu) + \mathbf{D}^T(\hat{\tau} - \mu) + \frac{1}{2}(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu)$$

This gives as approximations of the expectation and variance of f :

$$E[f(\hat{\tau})] \approx f(\mu) + \frac{1}{2}tr\{\mathbf{H}\hat{\Sigma}(\hat{\tau})\}. \quad (3.3)$$

and

$$\text{Var}[f(\hat{\tau})] \approx \mathbf{D}^T \hat{\Sigma}(\hat{\tau}) \mathbf{D} + \frac{1}{2}tr\{(\mathbf{H}\hat{\Sigma}(\hat{\tau}))^2\}.$$

See Appendix C for the full derivations.

3.2.2.1 Example revisited

When applied to the example dataset, the use of the second order approximation yields the new intervals $(-44.1, 43.2)$ for $t_{\frac{1}{2}}^{(S)} - t_{\frac{1}{2}}^{(M)}$ and $(-33.3, 27.4)$ for $C_{max}^{(S)} - C_{max}^{(M)}$ compared to the previous $(-38.2, 37.1)$ and $(-33.2, 27.3)$ respectively. The interval for the $t_{\frac{1}{2}}$ difference is noticeably wider for the use of the second order approximation, and the interval for the C_{max} difference is slightly wider. Since they both still contain 0, we remain with the same conclusion with this modification, there is no evidence of a difference between the two methods.

3.2.2.2 Simulations

To explore the difference in the confidence intervals based on first and second order approximations a small simulation trial is conducted. The confidence intervals for the half-life difference $t_{\frac{1}{2}}^{(S)} - t_{\frac{1}{2}}^{(M)}$, one of the derived parameters considered previously, are considered. The introduction of the second order term increases the estimate of the variance, and so the width of the intervals is increasing from the use of first to second order approximation for all simulation runs. Note that the introduction of the second order term also slightly changes the actual point estimate due to the second term in (3.3).

The implications of these changes can be seen in Figure 3.3 that shows the difference between type I error rate when using the first and second order approximation and a t quantile. We find that the type I error rate is closer to the

target level for small sample sizes when using the second order approximation (grey lines) compared to the first order approximation (black lines), while the difference is negligible for large sample sizes. Most notable, the trend of increasing observed type *I* error rate for increasing number of time points is still apparent. This suggests that, while the use of the second order approximation of the variance and the estimate is an improvement, it does not remove the type *I* error inflation for large sample sizes. For typical PK studies, it is the smaller sample sizes that are of primary interest. Note also, that the calculation of the second derivatives Hessian matrix for each simulation increased the computational cost of the method.

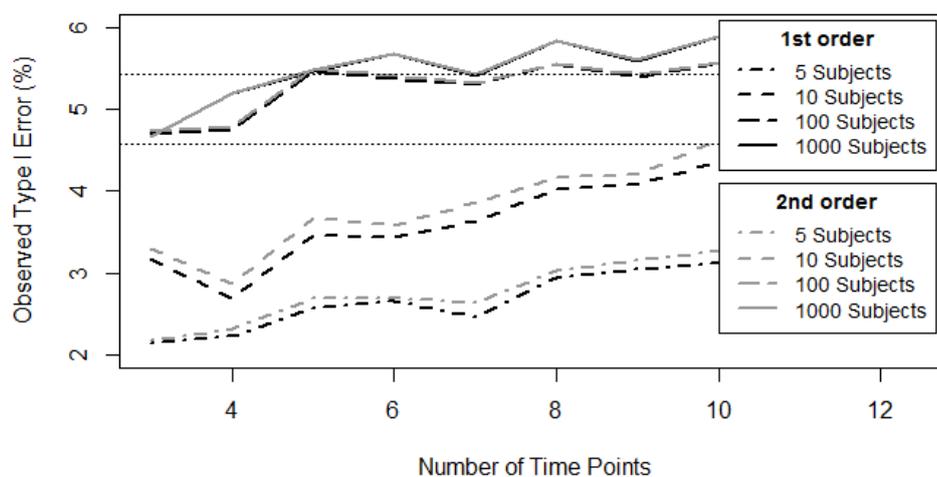


Figure 3.3: Comparison of observed type *I* error for varying number of time points and subjects for use of 1st and 2nd order approximation. Horizontal dotted lines show error bounds for 10,000 simulations.

3.3 Equivalence Testing

After this initial exploration of the properties of the procedure by Jensen & Ritz,⁴⁷ we now focus on the main setting of interest. In the application for the comparison of traditional and microsampling, we are in fact interested in

evidence of equivalence of the two sampling methods rather than superiority. This changes the nature of the tests we conduct, an overview of the differences is explored by Schuirmann.⁶⁶ The null hypotheses now state that the difference between parameters are at least as big as some given limit. Therefore the multiple test procedure breaks down into the following multiple two one sided null hypotheses:

$$\begin{aligned}
 &H_{01_1} : \theta_1 \leq \gamma_1 \text{ or } H_{02_1} : \theta_1 \geq \gamma_2, \\
 &\quad \vdots \\
 &H_{01_l} : \theta_l \leq \gamma_1 \text{ or } H_{02_l} : \theta_l \geq \gamma_2, \\
 &\quad \vdots \\
 &H_{01_L} : \theta_L \leq \gamma_1 \text{ or } H_{02_L} : \theta_L \geq \gamma_2,
 \end{aligned}$$

where $\gamma_2 > \gamma_1$ are the equivalence margins.

The main difference in the method proposed required for this setting is how confidence intervals are constructed and more specifically how the family-wise type I error rate is controlled. We will use the method proposed by Quan et al.⁶⁰ for assessment of equivalence of multiple correlated endpoints which uses the following adjustment. Assuming the endpoints, or in this case, derived parameter estimates $\hat{\theta}$ have a multivariate normal distribution with mean θ and variance Σ , with $\sigma_l = \sqrt{\Sigma_{ll}}$, where Σ_{ll} is the variance of the l th derived parameter estimate. Then for L derived parameter estimates, since we are considering two one-sided hypotheses, we require for $100(1 - 2\alpha)\%$ simultaneous intervals

$$\mathbb{P} \left(\bigcap_{l=1}^L (-\gamma \leq \hat{\theta}_l \pm Q\sigma_l \leq \gamma) \mid H_0 \right) = \alpha,$$

which is equivalent to

$$\mathbb{P} \left(\bigcap_{l=1}^L (Q\sigma_l - \gamma \leq \hat{\theta}_l \leq \gamma - Q\sigma_l) \mid H_0 \right) = \alpha.$$

Therefore replacing the unknown Σ with its estimate $\hat{\Sigma}$, then we find C such

that

$$\mathbb{P} \left(\bigcap_{l=1}^L (-C \leq \hat{\theta}_l \leq C) \mid H_0 \right) = \alpha.$$

In order to control the familywise type I error, find the maximum of the $Q_l = \frac{\gamma - C}{\hat{\sigma}_l}$ to give the adjusted quantile Q . With that adjustment in place we may now apply the standard procedure for equivalence testing: If all intervals fall within the previously stated bounds then we reject the null, there is evidence of equivalence. If at least one interval falls outside the specified range then we fail to reject.

3.3.1 Example revisited

Since the equivalence margins must be prespecified, a sensible approach is to follow a standard setup of a test for bio-equivalence such as that described by.¹⁸ We calculate the log of the ratio of the two parameters, in this case we consider $\log(C_{max}^{(S)}) - \log(C_{max}^{(M)})$ and $\log(t_{\frac{1}{2}}^{(S)}) - \log(t_{\frac{1}{2}}^{(M)})$ as the derived parameters of interest. The approximations of the estimates and their variances are calculated using the same methods previously discussed for superiority testing and simultaneous confidence intervals are calculated using this and the adjusted quantile Q . We then observe whether the entire confidence interval for this lies between $\gamma_1 = \log(0.8)$ and $\gamma_2 = \log(1.25)$, the standard bio-equivalence margins.

Although testing for equivalence between two sampling methods is of course different to testing for bio-equivalence, it suffices to use the same equivalence margins as a basis for analysis. Should one wish to be harsher or more lenient with one's definition of equivalence, then alternative equivalence margins may be used (with the possibility of taking $\gamma_1 \neq -\gamma_2$), however this would not affect any conclusions drawn from simulation results.

The equivalence test is applied to the example dataset using the same candidate models as used previously in the superiority test. The simultaneous confidence intervals of the log of the ratio of derived parameters are (-0.421

,0.387) and (-0.135,0.066) for $t_{\frac{1}{2}}$ and C_{max} respectively. The first is clearly outside $(\log(0.8), \log(1.25)) = \pm 0.223$. Hence there is insufficient evidence to reject the null hypothesis, there is no evidence of equivalence. Although this is possibly due to the small sample size being used in the trial.

3.3.2 Simulation Studies

To evaluate the method for testing equivalence more formally, we still wish to simulate data under the null hypothesis in order to evaluate the type I error rate. To do so we must consider a difference between the parameters and therefore generate data for each sampling method from the same models but differing parameter values such that the ratio of one pair of parameters is equal to 0.8, $(\log(t_{\frac{1}{2}}^{(S)}) - \log(t_{\frac{1}{2}}^{(M)}) = \log(0.8))$, and the other equal to 1, $(\log(C_{max}^{(S)}) - \log(C_{max}^{(M)}) = 1)$. In order to observe the maximal type I error, we use the situation under the null that is closest to the alternative - that is only one ratio is outside equivalence and that it is on the border.

Identical candidate and data generating models (Models 3.1 and 3.2) are used as previously when testing for superiority. In order to extend the application of the method, we have also evaluated the procedure under the assumption of a multiplicative error framework. Therefore in addition to the additive error framework we have previously assumed, we repeat the simulations under the assumption of a multiplicative error framework using the following two candidate models:

Candidate Model 3.4: Log Log Linear with Multiplicative Error

$$g(\tau, t) = \exp(\beta'_0 \exp(\beta_1 t)) \epsilon, \quad (3.4)$$

Candidate Model 3.5: Log Linear with Multiplicative Error

$$g(\tau, t) = \beta'_0 \exp(\beta_1 t) \epsilon, \quad (3.5)$$

where $\log(\epsilon) \sim N(0, \sigma^2)$. Model 3.4 is used in the data generation with $\sigma = 0.05$.

Using the new rejection criteria, we can observe the coverage. That is we observe the percentage of cases in our simulation that all intervals do not fall within the equivalence margins. Since this is a two one-sided test procedure, constructing a 95% confidence interval and assessing its position is equivalent to each one sided test having nominal level of 2.5%. Therefore since we are only simulating from one side of the null interval, we now expect a type I error of 2.5% instead of the previous 5%.

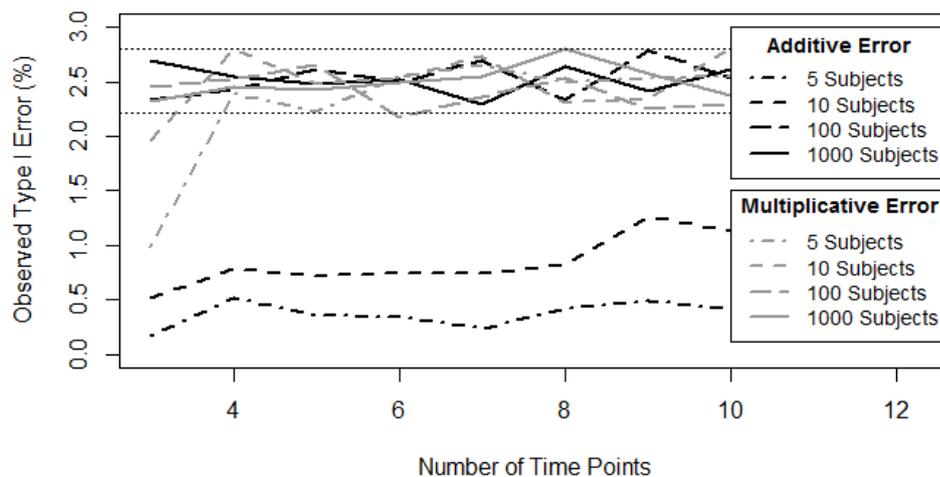


Figure 3.4: Comparison of observed type I error for varying number of time points and subjects for equivalence testing. Horizontal dotted lines show error bounds for 10,000 simulations.

Figure 3.4 shows the observed type I error rate using the 2nd order approximation and z quantile for both the additive and multiplicative error structure. A similar trend presents itself here to previously, with smaller type I error rate for the smaller numbers of subjects for the case of the additive error structure. For the multiplicative error structure, only for the small number of subjects and

time points does the type I error rate fall below the simulation error interval.

It is also important to consider the power of the procedure to detect equivalence when the underlying ratio of parameters is within the equivalence bounds. As an illustration of the difference in power curve of the procedure between the differing sample sizes when varying the underlying ratio to different values within the equivalence bounds, Figure 3.5 shows the curve for 5, 20 and 100 subjects at each time point. For the larger sample sizes the procedure shows high power in detecting equivalence, however for the smaller numbers of subjects, the power is noticeably lower.

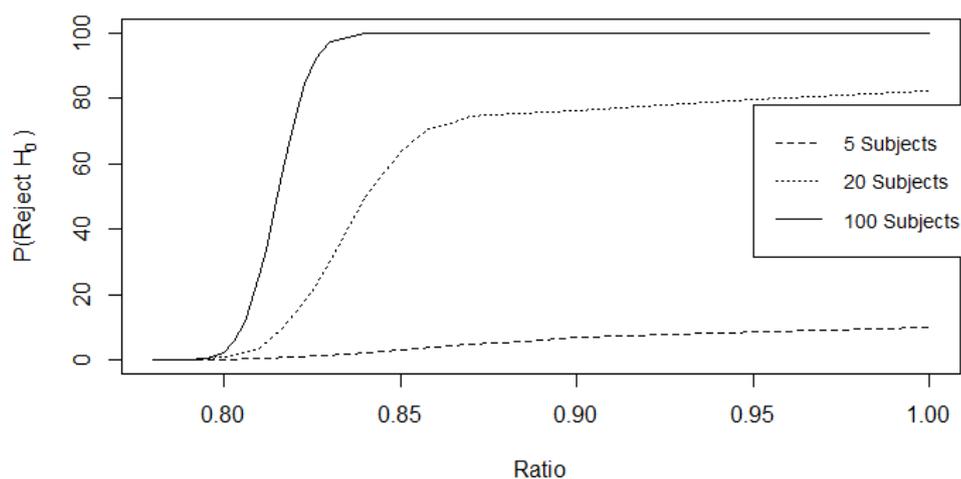


Figure 3.5: Power of procedure for equivalence testing with 5, 20 and 100 subjects at 3 time points.

In order to evaluate further this method for testing for equivalence between the two blood sampling methods, simulations are also conducted under the assumption of an oral administration of a compound.

For candidate models, the following standard one-compartmental oral dose model is used with two different error structures:

Candidate Model 3.6: Oral Dose with Multiplicative Error

$$g(\tau, t) = \left(\frac{k_a F D}{V(k_a - k_e)} (e^{-k_e t} - e^{-k_a t}) \right) \epsilon, \quad (3.6)$$

with $\epsilon \sim N(0, \sigma^2)$.

Candidate Model 3.7: Oral Dose with Additive Error

$$g(\tau, t) = \frac{k_a F D}{V(k_a - k_e)} (e^{-k_e t} - e^{-k_a t}) + \epsilon, \quad (3.7)$$

with $\log(\epsilon) \sim N(0, \sigma^2)$.

Candidate model 3.6 is used for data generation where we take $k_e = 0.0693$, $k_a = 0.231$, $V = 10$, $D = 500$ and $F = 1$ as described by Gibaldi & Perrier,³² assuming the $\log(\epsilon) \sim N(0, \sigma^2)$ with $\sigma = 0.05$. σ is chosen larger than that considered by Lunn & Aarons⁵¹ and identical to one example evaluated by Tod et al.⁶⁹ We believe that this is a reasonable standard deviation to use in the simulations to ensure the models of the two sampling methods are well separated. A larger standard deviation would mean the models were less well separated and although the procedure still works with differing values, this level of variation balances between separating the models well, and being realistic in terms of the variation expected between subjects at these concentrations.

Possible PK parameters to consider are C_{max} , t_{max} and AUC_{24} and simultaneous inference is performed on all combinations of pairs of these parameters and all three. See Appendix A for the form of these parameters. In the same process as previously, simulations are conducted to investigate the type I error rate and power of the procedure. In order to simulate under the null hypothesis, as previously, we generate the data so that the true log ratio of one PK parameter for the two sampling methods is equal to 0.8 and all others are equal to 1. In these simulations, the scenario with three time points has been omitted due to insufficient data to fit the more complex model. For the simulations to investigate the power of the procedure, the true log ratio of one PK parameter for the

two sampling methods is varied between 0.8 and 1, and the scenario with seven time points is considered.

Results show that type *I* error rate is controlled well for all combinations of these PK parameters, see Tables E.1 and E.2 in Appendix E. The power is also sufficient for a reasonable sample size. Figure 3.6 shows the power of the procedure, indicating that even with only 5 subjects per time point, an adequate power is achieved. For this case, the multiplicative error candidate model has average weight 0.9 over the 10,000 simulations and the additive error candidate model had average weight of 0.1. Around 18% of the simulations resulted in weights between 0.1 and 0.9 for both models, so it is clear that we are not consistently in the situation where we have weights 1 and 0, supporting the need for the model averaging in the procedure.

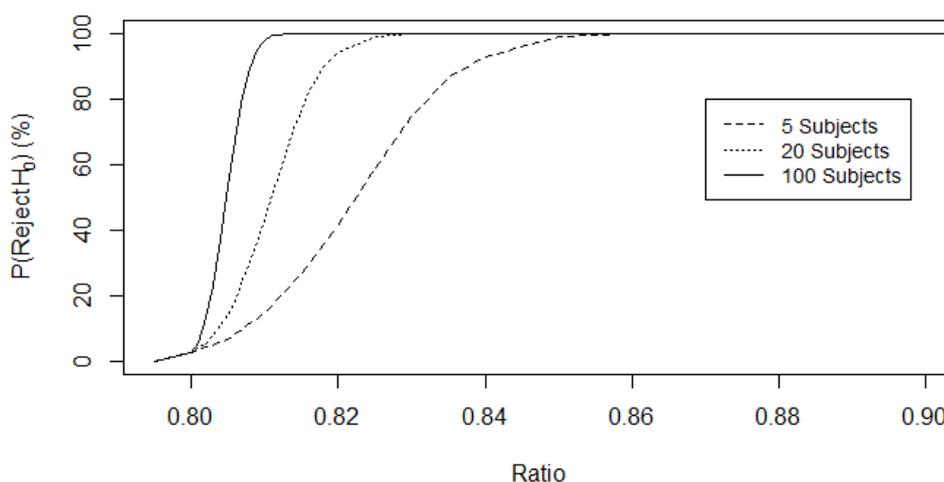


Figure 3.6: Power of procedure for equivalence testing with 5, 20 and 100 subjects at 7 time points for an oral administration of a compound.

Therefore for the purpose of our comparison between microsampling and traditional sampling, the proposed method for equivalence testing is applicable. Although in some cases conservative for the smaller sample sizes, the type

I error rate is controlled, with expected asymptotic behaviour for the larger sample sizes. An adequate power is achievable for a sample size that is manageable for a one-off study to confirm equivalence between the two sampling methods. Hence the suggested procedure provides an effective approach to the considered problem of providing a comparison between microsampling and traditional sampling.

3.3.3 Extension to Longitudinal Data

Although in these previous studies it is assumed that different subjects are sampled at each time point, extending to longitudinal data and fitting non linear mixed-effects models can be done without further complication. The derivations for this extension are not too dissimilar to those for the previous method and therefore are not included here. The ingredients are essentially the same; we use the estimates and variance matrix for the fixed effects from the fitted model. For the interested reader, we refer you to further details.^{46,64} We now consider the same subjects to be sampled at each time point, which may help somewhat to reduce the sample size needed for a realistic power to detect superiority or equivalence. In this section, sample size now refers to the total number of subjects in the study. This is an important extension in order to be able to conduct a study with fewer overall subjects to confirm equivalence between the sampling methods.

One may also wish to use separate models for the two sampling methods instead of combining them in a single model. This removes the restriction on the residual variances. This is implemented in the following simulations. However, fitting non linear mixed effects models greatly increases the computational intensity required to conduct simulations and therefore the number of simulations has been reduced to 1000. For this same reason, the scenario with 1000 subjects has been eliminated from these simulations. A reduced range of time points is

also considered.

The population model used is identical to that used in previous simulations for an oral administration of a compound for both the data generating model and the candidate models. Normally distributive additive random effects are assigned to V , k_e and k_a identically for the two sampling methods, with the ratio of the underlying population PK parameters of interest for each sampling method (AUC and C_{max} in this case) fixed as previously.

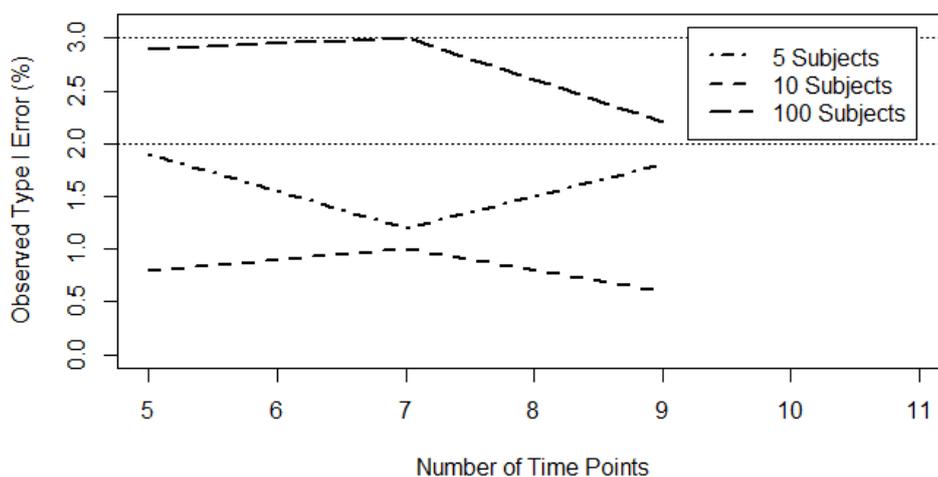


Figure 3.7: Type I error rate for varying numbers of time points for 5, 10 and 100 total subjects considering AUC and C_{max} as PK parameters for an oral administration of a compound. Horizontal dotted lines show error bounds for 1000 simulations. (Equivalence testing)

It is clear from Figure 3.7 that the Type I error rate is within simulation error bounds for 100 subjects, and is conservative when 5 and 10 total subjects are used. Figure 3.8 shows that even for 5 and 10 total subjects, we achieve a reasonable power for underlying ratios between 0.9 and 1. Therefore this extension, if physically feasible to collect such samples, provides a framework that could indeed be used to detect equivalence between the two blood sampling methods for a small sample size.

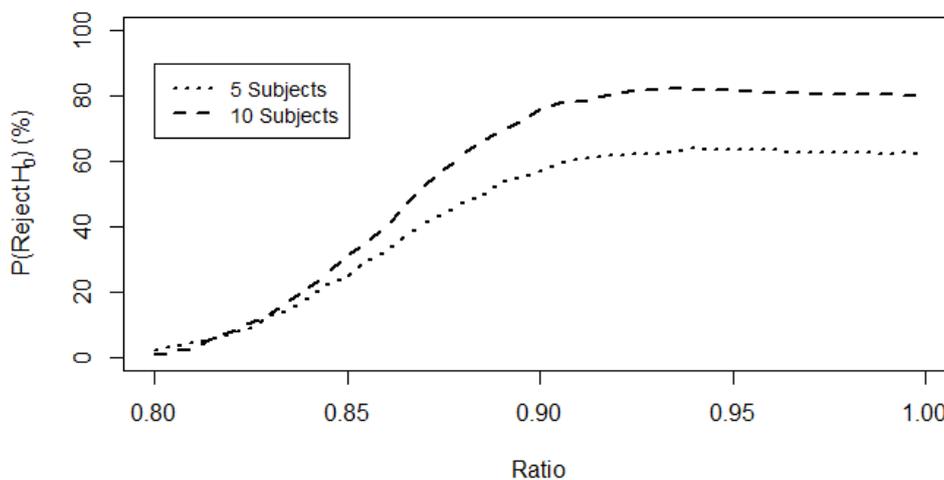


Figure 3.8: Power of procedure for equivalence testing for 5 time points for 5 and 10 total subjects considering AUC and C_{max} as PK parameters for an oral administration of a compound.

3.4 Discussion

The method by Jensen & Ritz⁴⁷ showed varying performance in terms of coverage in the simulation studies conducted in the application of comparison of PK parameters. It is noted that these methods⁴⁷ do need adjustment in order to be applied to our desired setting. However since our choice of designs are different to those discussed⁴⁷ then of course it is expected to be necessary to make such adjustments. The simulation studies are conducted in order to evaluate the power and type I error rate of varying designs so that we are able to make recommendations for such a study needed to detect equivalence of the sampling methods. Some noteworthy relationships between the coverage and the number of time points and the size of the sample emerged in this analysis.

The use of the second order delta approximation instead of the first order delta approximation showed coverage closer to 95% in the simulations, and in spite of the increase in computational cost, would be preferred over the first or-

der approximation. For the smaller sample sizes that are more representative of the type of studies this methodology is aimed at, the second order approximation does offer an improvement. For the larger sample sizes, it does not offer much improvement. However since the smaller sample sizes are typical of such studies, we have continued to recommend the use of second order approximation.

The extension to equivalence testing gives promising results that this could be applied in the desired setting. The method can be applied to different error structures and can even be used to account for uncertainty in the error structure. It can be applied to both simple and complex models, with both giving encouraging results. The simulations also give strong indications of the required sample size for such studies to have the power to detect equivalence between the two blood sampling methods.

When applied to the example dataset, neither the superiority or equivalence test found evidence to reject the null hypothesis. Hence with this particular example results were inconclusive, due to the small sample size. However, if a study were designed for this purpose with the power to detect equivalence, then it is hoped that results will be useful for the comparison of the two sampling methods. Thus the method developed to detect equivalence is an advantageous tool in this practical application of comparison between microsampling and standard sampling.

The extension to longitudinal data shows great promise to reduce the total sample size needed for such a study. However there may be physical restrictions on the collection of such samples using traditional methods, since the total volume of blood needed for the collection of multiple samples from the same subject may exceed protocols dictating the total volume of blood allowed to be taken within a certain time period. Therefore one may want to introduce a sparse sampling scheme, where not all subjects are sampled at each time point. This may

provide some middle ground between the two scenarios previously considered, balancing between a practical total sample size and a practical sampling schedule for each subject in terms of blood volume sampled.

The equivalence margins used are standard for testing for population bioequivalence of a test formulation of a drug against a reference formulation. However one may want to be more stringent when comparing sampling methods in order to reduce compounding of error. As a possible extension, one may also want to consider following the procedure for testing for individual bioequivalence as opposed to population bioequivalence. This may also enforce stricter conditions on claiming equivalence, which for practical reasons ought to be considered.

CHAPTER 4

Paper B: Optimal Designs for Non-Compartmental Analysis of Pharmacokinetic Studies

4.1 Introduction

Investigating pharmacokinetics (PK) is an indispensable task in drug development in order to understand how externally administered compounds are absorbed, distributed, metabolised and excreted by an organism. PK behaviour is typically assessed by measuring the substance's concentration in blood, plasma or other tissues at a number of time points after administration. Based on the resulting concentration-time profile, a variety of pharmacokinetic parameters such as the area under the concentration versus time curve (AUC) are considered. Two strategies, the non-compartmental approach^{15,32} or a modelling approach,^{11,20} are available for estimating PK parameters, describing the underlying exposure. Modelling allows for unstructured sampling schemes, which comes at the cost of uncertainty about the correct model to use, and technical difficulties in model fitting may arise. The non-compartmental approach on the other hand uses minimal assumptions about the data-generating process. The drawback is, however, that a more structured sampling design and some approximation (usually linear) of the curve between two observed time points is necessary.

The recent introduction of the new blood sampling method microsampling has led rise to the potential superfluity of satellite animals in such PK studies. The practice of microsampling involves taking a much reduced volume of blood per sample due to improvement in the sensitivity of bioanalytical techniques which mean that analysis can be conducted with smaller samples. Whereas previously volumes $\geq 200 \mu\text{L}$ were needed per sample, the practice of microsampling requires only approximately 25 - 30 μL per sample.¹⁶ As a consequence of the high volume drawn previously, the PK had to be studied in a group of animals called the satellite group and the pharmacodynamics (PD) was studied in a group called the study group due to the blood volume needed for both

measurements to ensure blood sampling did not adversely affect PD measurements. The extra subjects required for a satellite group are now superfluous as the smaller volume of blood taken in the sample means that we can sample both requirements from the same study animals. This is a big advantage in terms of the Reduction principle of the 3 R's of non-clinical research.⁶⁵ Since a rat can provide up to 250 μL in total over 24 hours,⁵³ the physical and time restrictions of the scientists performing the sampling are stricter than the ethical and physical restrictions on the amount of blood provided by a rat. One must also consider that the PD outcomes must not be influenced by this sampling. The number of samples taken per timepoint remain constant as previously, but there is now a larger pool of subjects to sample from, hence all animals cannot be sampled at all timepoints. Therefore what previously was a full sampling scheme in the satellite group results in a sparse sampling scheme in the study animals. There are of course many other instances, such as sampling from human babies and children, in which sparse sampling schemes are used. One proposed scheme (Table 4.1) that has arisen from discussion of the implementation of microsampling is the example introduced by Chapman et al.¹⁶

This example is just one of many possible sparse sampling schemes. The question is then, out of all possible schemes, how does one choose the best?

Much literature on optimal designs presents methods from a PK modelling perspective, with much focus on optimality criteria based on the Fisher Information Matrix.^{62,63,69} The most popular of which is *D*-optimality, minimizing the determinant of the inverse of the Fisher Information Matrix. The application of this⁹ is widely used for optimal sparse sampling schemes. For serial sampling, Gagnon & Leonov³⁰ discuss the process of finding such optimal designs. A large range of procedures for choosing optimal designs is explored by Fedorov & Leonov.²⁸

Since, due to regulatory requirements, such studies primarily make use of

Table 4.1: Sparse Sampling Scheme as suggested by Chapman et al.¹⁶ All main study animals are sampled, with 10 animals per sex per group. A total of 30 samples per sex per group are taken, with 3 per each of the 10 animals and 5 per each of the 6 timepoints.

Animal number	Sampling timepoint					
	#1	#2	#3	#4	#5	#6
1	X		X			X
2	X			X		X
3		X	X			X
4		X	X		X	
5		X			X	X
6		X		X	X	
7		X		X		X
8	X		X	X		
9	X			X	X	
10	X		X		X	
	n=5	n=5	n=5	n=5	n=5	n=5

non

-compartmental methods (NCA) in their analysis,^{29,61} we will focus on optimal designs for non-compartmental methods here. The traditional approach using the Fisher Information Matrix is not applicable here, since this requires the specification of a PK model. We also ensure robustness across possible scenarios by applying a minimax procedure.

Depending on the type of sampling, non-compartmental theory presents different methods for estimation of PK parameters and their variance. The area under the concentration versus time curve (AUC) is the most explored in literature, but others such as the maximum concentration (C_{max}), the time the maximum concentration occurs (t_{max}) and the time taken for the concentration to decrease to half its value ($t_{\frac{1}{2}}$) are also important in describing a population PK profile.

A standard method for estimating the AUC is the simple trapezoidal rule,

which uses linear interpolation between observed concentration measurements to approximate the concentration versus time curve. This is explored in depth in the literature.^{7,25,50} There have been many suggested extensions building upon this, such as the hyperbolic trapezoidal rule⁵⁴ and the log trapezoidal rule³¹ for destructively obtained measurements.

Whilst the estimation of point estimates of PK parameters do not depend on the sampling scheme, estimation of the variability of the estimator does. There are different types of sparse sampling schemes that can be used in such studies, with NCA methods available for calculating this variability. For serial sampling, where one blood sample is taken per subject, there are various methods for calculating the variance of the *AUC* estimator^{6,44,55,71} depending on the scenario. Alternative approaches exist for batch designs,^{37,41,75} where each animal is measured at multiple (but not all) time points. The animals are divided into batches with each batch of animals all sampled at the same time points. No other animals are sampled at any of these time points and hence the time points for each batch form a partition of the set of all time points. For more flexible designs, where the animals are not separated into disjoint batches by their time points, methods described by Jaki & Wolfsegger⁴³ can be applied. However for each individual set of sampling times, there must be at least two animals following this. For all such types of design, the R package PK⁴² can be used.

For other PK metrics, Fedorov & Leonov²⁶ compare the empirical estimators based on numerical integration methods to the model based estimators using a measure of MSE as the comparator. Wolfsegger & Jaki⁷² also investigate non-compartmental estimates for PK parameters, but focus on serial sampling designs.

Although reporting various advantages and disadvantages for their respective methods, it is made clear by all authors that the approach for estimating the PK parameters and their variances must be appropriate for the sampling

trial design. For each of the methods discussed, there are certain criteria that the sampling scheme must satisfy, for example the partitioning of schedules in batch designs or repeats on individual schedules in flexible designs. The example sparse sampling scheme¹⁶ however does not satisfy these criteria and hence these methods for estimating the variances are not appropriate. Therefore in the following, rather simple non-compartmental metrics are used since they are appropriate for such small sparse sampling designs, with variances estimated using simulations although exact variance estimators can be used if available for the class of designs of interest.

In this work we develop a general procedure for determining the optimal sparse sampling scheme for a given optimality criterion. We consider fixed time points as well as optimizing time points and discuss robust designs.

4.2 Method

The following procedure uses simulations (labelled $i = 1, \dots, M$ for very large M) to choose the optimal sparse design. We start by choosing designs assuming fixed time points, then will later move on to optimizing these time points. In order to define the set of all feasible designs S , we consider designs with J time points: t_j for $j = 1, \dots, J$ and N subjects $n = 1, \dots, N$. Each scheme is labelled S_k for $k = 1, \dots, K$ where K is the total number of schemes that satisfy our design constraints. We introduce an indicator variable, γ_{jn, S_k} , which takes the value 1 if subject n is sampled at time t_j in sparse scheme S_k , and 0 otherwise, with constraints on these γ_{jn, S_k} introduced. To choose the optimal design from the set of feasible designs, the optimality criterion used can be written as $\Psi = f(\hat{\theta})$ where $\hat{\theta}$ is an M -vector of the PK parameter estimates of interest across all simulations, that is the vector is made up of a parameter estimate from each of the M simulations.

The approach is outlined as follows:

Algorithm 1:

- (1) Define the set of all feasible sparse designs S .
- (2) Assume an underlying PK model, correlation and variance structure.
- (3) Using the model in step 2, generate M complete datasets and extract data according to each feasible sparse design.
- (4) Compute the optimality criterion for each feasible design S .
- (5) Select the best sparse design based on chosen optimality criterion.

By feasibility of a sparse design, we mean that it satisfies any constraints given by the user. We will later go on to specify particular constraints that we consider in our simulations, but at this point we keep the algorithm as general as possible in order to keep its range of applicability open.

In fact, we are not only able to select the best scheme, but can rank the sparse designs based on the chosen optimality criterion Ψ (examples below), with rank 1 given to the best design. Although this Ψ may relate to any PK parameter, we only consider using NCA as alternative approaches to find optimal designs for model based estimation exist (see²⁸).

In order to choose the optimal sparse sampling scheme out of the set of S_k , we use a measure of variability of the PK parameter of interest θ .

For example we may take

$$\Psi(S_k) = \text{var} \left(\widehat{\theta}_{S_k} \right),$$

where $\text{var} \left(\widehat{\theta}_{S_k} \right)$ is the variance of the estimator of the $\widehat{\theta}_{S_k}$ for scheme S_k . We then rank the designs from 1 to K by their value of $\Psi(S_k)$. The best design has the smallest value of $\Psi(S_k)$ and rank 1.

We use M simulated full datasets (observations from N subjects all at J time points). These datasets are indexed $i = 1, \dots, M$. We estimate this variance

of the PK parameter estimate for each scheme S_k by taking the variance of the estimates $\widehat{\theta}_{S_k}$ over $i = 1, \dots, M$ as follows. A non-compartmental estimate of θ is calculated for each scheme for each dataset, $\widehat{\theta}_{S_k,i}$. With $\bar{\theta}_{S_k} = \frac{1}{M} \sum_{i=1}^M \widehat{\theta}_{S_k,i}$, then:

$$\Psi(S_k) = \frac{1}{M-1} \sum_{i=1}^M \left(\widehat{\theta}_{S_k,i} - \bar{\theta}_{S_k} \right)^2 = \text{var} \left(\widehat{\theta}_{S_k} \right)$$

is our (unbiased) sample variance of the θ estimate for scheme S_k .

By standard results, this sample variance is an unbiased and consistent estimator of the true variance $\sigma_{S_k}^2$ (Lemma 4.2.1), a fact we can use to show that the above procedure will lead to the optimal ranking as the number of simulations increases.

Lemma 4.2.1. *For $\epsilon > 0$:*

$$\mathbb{P} \left(\left| \text{var} \left(\widehat{\theta}_{S_k} \right) - \sigma_{S_k}^2 \right| \geq \epsilon \right) \rightarrow 0 \text{ as } M \rightarrow \infty$$

for each $k = 1, \dots, K$.

Hence by using a sufficiently large M , our sample variance converges to the true variance.

Lemma 4.2.2. *The true ranking depends on the differences between variances, i.e. $\sigma_{S_k}^2 - \sigma_{S_{k'}}^2 > 0$ implies scheme S_k is ranked lower (worse) than $S_{k'}$, $\sigma_{S_k}^2 - \sigma_{S_{k'}}^2 < 0$ implies scheme S_k is ranked higher (better) than $S_{k'}$, $\sigma_{S_k}^2 - \sigma_{S_{k'}}^2 = 0$ implies scheme S_k is ranked equally to $S_{k'}$. Therefore as a consequence of Lemma 4.2.1:*

$$\text{var} \left(\widehat{\theta}_{S_k} \right) - \text{var} \left(\widehat{\theta}_{S_{k'}} \right) \rightarrow \sigma_{S_k}^2 - \sigma_{S_{k'}}^2 \text{ as } M \rightarrow \infty,$$

for all schemes $S_k, S_{k'} \in S$.

Hence the rankings based on our estimates will approach the true rankings as $M \rightarrow \infty$.

4.3 Results

4.3.1 Set Up

In order to apply the procedure, we consider schemes based on the example given by Chapman et al.¹⁶ We construct our sparse sampling designs by introducing constraints on the γ_{jn,S_k} . Following the example by Chapman et al,¹⁶ we use the following:

$$\sum_{j=1}^J \gamma_{jn,S_k} = T \leq J \quad \text{for all } n = 1, \dots, N,$$

that is, each subject is sampled at exactly T time points, and

$$\sum_{n=1}^N \gamma_{jn,S_k} = U \leq N \quad \text{for all } j = 1, \dots, J,$$

stipulating that U measurements need to be available at each time point. These constraints help keep the number of viable schemes small, but may be relaxed if one wishes to increase the amount of viable schemes K .

There are $H = \binom{J}{T}$ possible individual subject schemes that we label s_h for $h = 1, \dots, H$. Therefore each group sampling scheme S_k can be thought of as a unique collection of N out of the H individual subject schemes s_h . The constraints on the γ_{jn,S_k} obviously reduce this. We represent the group scheme S_k by a $N \times J$ matrix of 1's and 0's. Permutations of rows do not alter the structure of the scheme.

We use $N = 10$ subjects and $J = 6$ time points restricted to sampling exactly $U = 5$ subjects at each time point and each subject being sampled at exactly $T = 3$ time points. For this situation there are $K = 1044$ feasible sparse sampling schemes. We do not allow for repeated individual schemes s_h and place no other restrictions on sampling schemes, although it is possible to do so (for example one may want to disallow sampling the same subject for 3 consecutive time points).

Our assumed PK model is a one-compartmental first order kinetics with oral administration:

$$Y_{jn} = \frac{k_{an}FD}{V_n(k_{an} - k_{en})}(e^{-k_{en}t_j} - e^{-k_{an}t_j}) + \epsilon_{jn},$$

with mixed effects implemented as follows. The values of F and D are fixed at $F = 1$ and $D = 100$. The other three parameters, V , k_a and k_e have exponential random effects $(V_n, k_{an}, k_{en}) = \beta \exp(\mathbf{b}_n)$ where (V_n, k_{an}, k_{en}) is the individual parameter vector for subject n , β is the fixed effect and \mathbf{b}_n is the random effect for individual n , with $\mathbf{b}_n \sim N(\mathbf{0}, \Omega)$. Ω is defined as a diagonal matrix of the variances of the random effects for each of the three parameters, V_n, k_{an}, k_{en} . We assume values for the fixed effects, $\beta = (15, 2, 0.25)$ and the variances for the random effects $\text{var}(\mathbf{b}_n) = (0.1, 1, 0.25)$. The ϵ_{jn} are Normally distributed with constant coefficient of variation 0.15. The time points used are (0.5, 1, 2, 4, 9, 12). This follows an example described by Bazzoli et al.⁹ Figure 4.1 shows the population PK model with the time points considered. We generate $M=100,000$ full datasets of $N = 10$ subjects sampled at $J = 6$ timepoints in order to apply the method.

4.3.2 Initial Results

Using the $\theta = AUC$ calculated using the trapezoidal rule as the PK parameter of interest, we see the results for the top 10 ranking designs in Table 4.2. Schemes are referred to by their reference number (from 1 to 1044) throughout the results as it is not possible to provide the actual design for each choice. Although there may be little difference between the optimality criterion value for these top designs, the worst ranking designs have variances over 15% larger than the best design. This provides evidence that out of the set S , there is vast differences between the designs S_k . The example design by Chapman et al¹⁶ is ranked 84 out of 1044, with a variance of 6.856, not one of the worst designs but

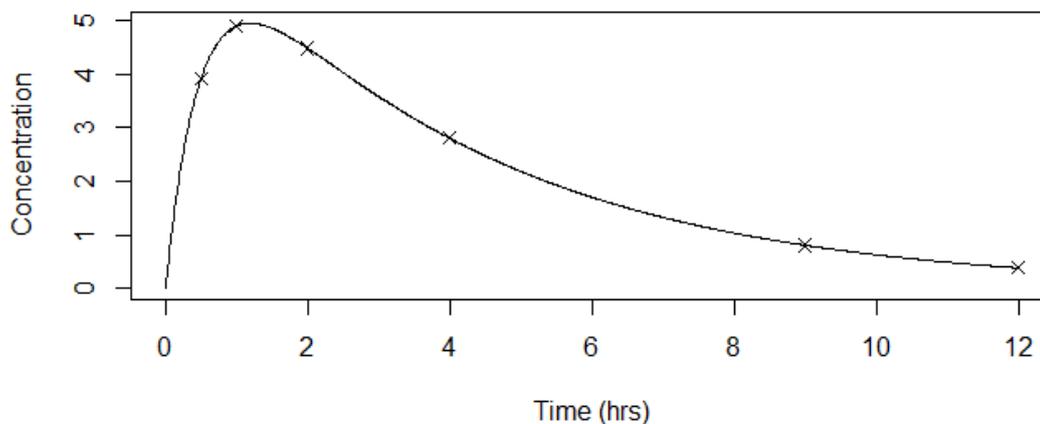


Figure 4.1: An illustration of the population PK model described and the sampling time points.

by no means the best, with over a 4% increase in variance over the best design.

Table 4.2: The top 10 and bottom 10 ranking designs according to the minimization of variance of the AUC estimate

Rank	Scheme #	AUC Var	Rank	Scheme #	AUC Var
1	1003	6.583	1035	73	7.527
2	1001	6.655	1036	20	7.532
3	775	6.661	1037	6	7.548
4	746	6.663	1038	236	7.553
5	809	6.680	1039	52	7.570
6	1026	6.696	1040	19	7.574
7	993	6.711	1041	44	7.580
8	764	6.714	1042	270	7.592
9	1018	6.730	1043	299	7.602
10	750	6.731	1044	42	7.656

It is natural to be curious about how this ranking differs to that which would be obtained using $\Psi(S_k) = \text{MSE}(\widehat{\theta}_{S_k}) = \text{bias}(\widehat{\theta}_{S_k})^2 + \text{var}(\widehat{\theta}_{S_k})$ as optimality criterion, as we have so far ignored the bias of the estimate of the PK parameter. It has already been established that sparse sampling schemes produce unbiased estimators for PK parameters⁴³ and hence it is unsurprising that

the rankings based on the MSE and the variance alone are very similar (see Figure 4.2). Consequently we will not consider the bias when choosing the sparse sampling scheme in our subsequent illustrations, but instead reserve this for our revisit when considering the choice of time points discussed later.

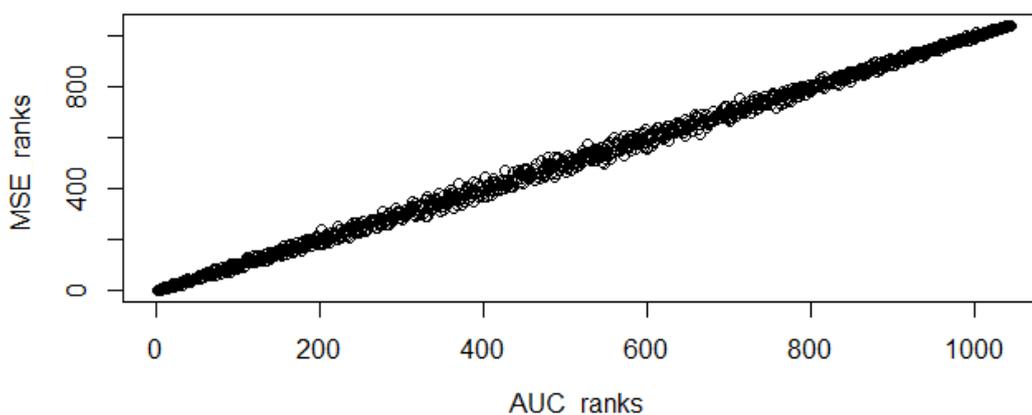


Figure 4.2: The relationship between ranks given to schemes using MSE vs variance of the AUC estimate as optimality criterion.

It is also worth noting that although we have considered only sampling schemes that do not include repeated schemes, it is possible to do so. For illustrative purposes we have included results for such a scenario, where individual schemes are allowed to be repeated once. That is, any one of the 20 individual schemes that we consider can appear up to a maximum of twice in the sampling scheme, which results in a total of 10374 possible schemes. This yields slight improvement in the value the optimality criterion for the top design, reducing the variance by nearly 2%. However for simplicity and computational time we illustrate further examples without considering replication of individual designs.

However in such trials, more than one PK parameter is typically of interest. Most commonly, AUC and C_{max} are used together to describe the PK profile. Therefore it seems appropriate to use an optimality criterion that incorporates

the variation of both parameters. We suggest the following:

$$\Psi = w_1 \text{var} \left(\widehat{AUC} \right)^* + w_2 \text{var} \left(\widehat{C}_{max} \right)^*$$

Where w_1 and w_2 are prespecified weights assigned based on the importance of the relevant PK parameter, and * indicates the variances are scaled from 0 to 1 by dividing the variance by the maximum variance across all schemes. The scaled variances are used as opposed to the actual variances as the PK parameters themselves are on very different scales, and so for the weights to have a meaningful interpretation they must be on the same scale. The choice of these weights is left to the user so that the level of interest in that particular PK parameter can be taken into account. In all further examples we use the weights $w_1 = w_2 = 0.5$.

Further evidence to support this criterion is the relationship between rankings given to the set S when using the variances of the individual PK parameter estimates shown in Figure 4.3. We see a slight positive relationship, in that the best designs tend to be shared by both criteria and likewise for the worst designs. However in the middle rankings there is a lot of discrepancy between the two criteria. The zoomed in figure makes it clear that from the top 50 schemes according the AUC ranking are mainly ranked in the top half by C_{max} , and in fact many are ranked in the top 50.

Table 4.3 shows the results using this criterion, including the ranks and efficiency measures given to each of the top designs when using the variance of the individual parameter estimates as the optimality criterion. This suggests that designs that perform well according to both measures, such as 1018, are more desirable than those which perform better for one and not for the other. The efficiency measure also highlights how even though the top design, 1003, has rank 81 in the C_{max} measure, it is still close to the best value, with an efficiency measure of 1.012. This is in fact the same as the efficiency of the 4th best scheme

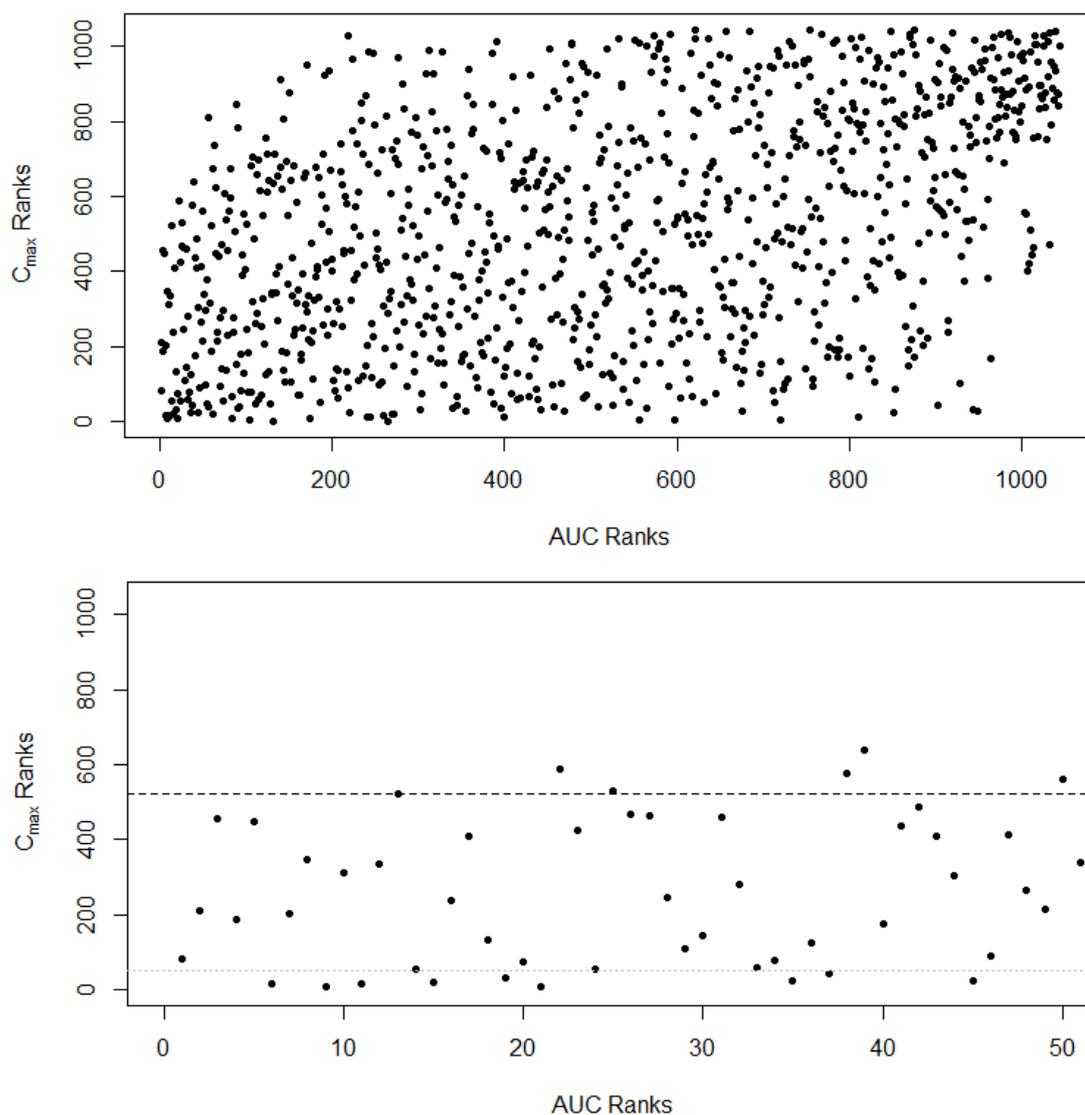


Figure 4.3: The relationship between ranks given to all schemes using the variances of \widehat{AUC} and \widehat{C}_{max} (top). The relationship between ranks given to schemes using the variances of \widehat{AUC} and \widehat{C}_{max} for the top 50 schemes according to the AUC ranks. Horizontal dashed line represents the top half of the ranks, dotted line represents rank 50 (bottom).

according to the AUC criterion alone.

4.3.3 Comparison to Model Based Optimal Designs

The proposed methods described for finding an optimal sparse sampling scheme made as few assumptions on the underlying model as possible, making

Table 4.3: The top 10 ranking designs according to the minimization of variance of $\Psi = w_1 \text{var} \left(\widehat{AUC} \right)^* + w_2 \text{var} \left(\widehat{C}_{max} \right)^*$ with $w_1 = w_2 = 0.5$. Efficiency Measure for a given scheme is the ratio between the variance for that scheme and the best scheme, the larger the value, the less efficient the scheme.

AC Rank	Scheme #	<i>AUC</i> Rank	<i>AUC</i> Eff. Measure	C_{max} Rank	C_{max} Eff. Measure	Ψ
1	1003	1	1.000	81	1.012	0.914
2	1026	6	1.017	17	1.005	0.918
3	1018	9	1.022	8	1.004	0.919
4	1039	11	1.023	15	1.005	0.920
5	1001	2	1.011	212	1.017	0.920
6	746	4	1.012	186	1.016	0.921
7	1006	15	1.024	19	1.006	0.921
8	1035	21	1.029	7	1.004	0.922
9	1025	14	1.024	54	1.010	0.923
10	1027	19	1.027	30	1.008	0.923

use of non-compartmental methods as optimality criterion. One of the fundamental assumptions used, however, is the underlying data generating model and hence it is natural to compare this to the model based criterion. This method is proposed as an alternative to model based methods, which use some measure of the Fisher Information Matrix for a particular model to judge the suitability of a sampling scheme. One of the most popular model based optimality criteria is D -optimality, which seeks to minimize the inverse of the determinant of the Fisher Information Matrix.

In order to compare the proposed non-compartmental method to the model based method, we use the approximation to the Fisher Information Matrix for non linear mixed effects models described by Retout et al.⁶³ The model used to generate data in the previous simulations is assumed as the model, with the same parameter values. The Fisher Information Matrix under this model is calculated for each of the 1044 sparse sampling schemes.

Figure 4.4 shows the relationship between the ranks given using our proposed non-compartmental methods and those given using D -optimality. We see that there is no real relationship between the two sets of ranks, unsurprising since they optimize two very different criteria. However this does highlight the

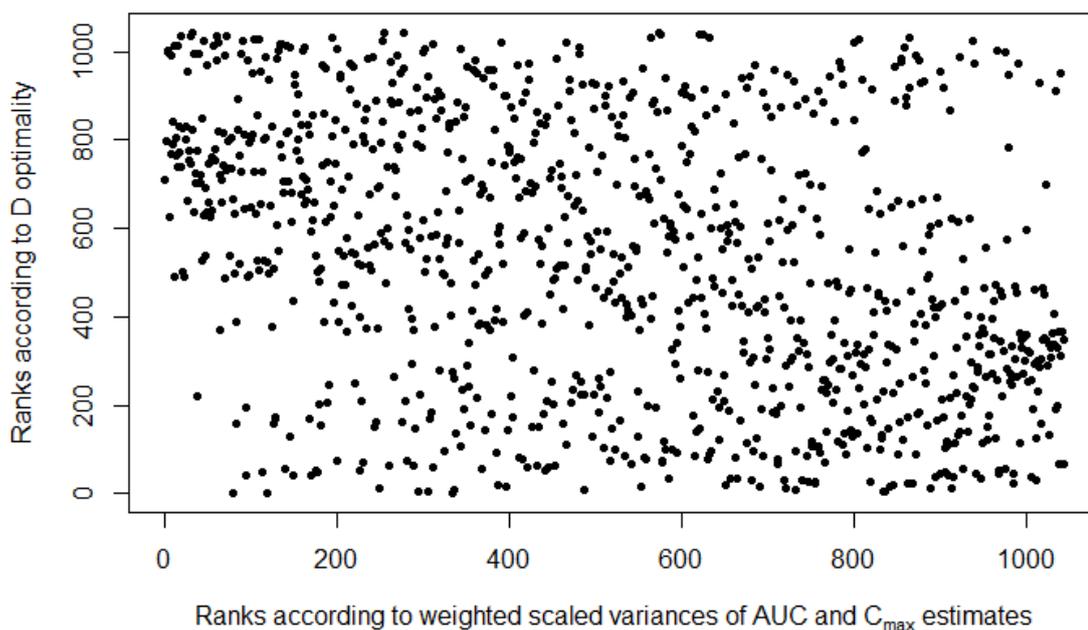


Figure 4.4: The relationship between ranks given to schemes using the weighted sum of scaled variances of \widehat{AUC} and \widehat{C}_{max} , and D -optimality.

importance of using optimality criteria that are relevant to the measured outcome of the trial the scheme is applied to.

4.3.4 Application of Minimax Criterion

So far we have considered just one scenario, i.e. one underlying PK model with one set of parameters. However it is most likely that there is substantial uncertainty over the model. It is desired that the chosen scheme will be robust over this uncertainty, hence we consider a worst case scenario - our chosen scheme should perform well over all scenarios considered. We prefer a scheme that has reasonable results for all scenarios than one that has excellent results for some and poor results for others. Hence we use the minimax criterion.

Given $p = 1, \dots, P$ different scenarios, each scenario could be different correlation structure or models etc. Each of these scenarios will give a ranking of the K sparse schemes. Let R_s be defined as $R_s(S_k, p)$, the rank of S_k for the

p th scenario. Then our robust choice of optimal design is the solution to:

$$\arg \min_{S_k} \max_p R_s(S_k, p).$$

In particular, there is generally uncertainty over the variance associated with the random effects of the parameters. Hence to apply the minimax procedure, for each of the three model parameters that we have assigned random effects to, we consider a high (H) and low (L) variance. All combinations of these give the 8 scenarios with details given in Appendix F. Of course one may choose for example to consider different values for the fixed effects across scenarios if it is believed that there is a potential for this to be apparent in the study of interest.

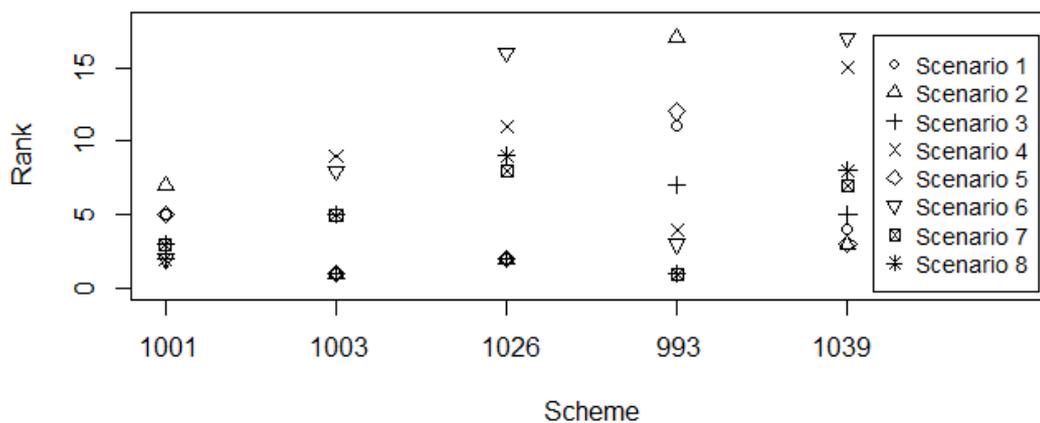


Figure 4.5: Top 5 overall schemes according to minimax criterion applied to equally weighted scaled sum. Ranks for each of the eight scenarios are plotted.

Table G.1 in Appendix G and Figure 4.5 show the results of the application of the minimax procedure. We see that the overall best scheme is chosen to be scheme 1001, although interestingly it is never chosen to be the best scheme in any scenario. Scheme 1003 is a close second, ranking 1st in half of the scenarios, but overall performs worse across the board. The variation across scenarios increases for overall lower ranking designs, thus the higher ranking designs are

also more consistent in their rankings than lower ranking designs.

It is also worth considering how these choices would change were we to consider the efficiencies of the schemes, rather than the ranks, in the minimax procedure. Table G.2 in Appendix G shows these results. The top choices of scheme are indeed very similar, although with some permutation in the overall position. This is due to the small differences in the efficiency measure between scenarios. For example, the 7th ranking design in scenario 2 is less efficient than the 9th ranking design in scenario 2. Hence by using the efficiency measure we obtain slightly different results to using the ranks. However, for this particular example, the difference is not so great and a good overall design would be picked regardless of the use of the efficiency measure or rank. The ranks are used for the remainder of the examples, however the extension to using efficiencies is simple if it is required by the user.

4.4 Choice of Time Points

So far we have assumed a fixed set of time points available for blood sampling: $(0.5, 1, 2, 4, 9, 12)$. We now extend to the case where the choice of these time points is possible. To include the choice of time points, we suggest two approaches. Ideally, we extend the set of feasible schemes to include the opportunity for different sampling time points, but the rest of the approach is identical to Algorithm 1. The challenge with this approach (formalized as Algorithm 2) is that the number of feasible schemes becomes so large that it is computationally intractable to find the optimal design. Therefore the second approach (Algorithm 3) breaks this down into a two stage procedure, which is less computationally intensive and will approximate the results of Algorithm 2. We provide a toy example to illustrate the application of Algorithm 2, then further evaluate Algorithm 3 in a manner it could be used in practice.

Algorithm 2:

- (1) Define the set of all feasible sparse designs applied to feasible time choices $S_{(\tau)}$.
- (2) Assume an underlying PK model, correlation and variance structure.
- (3) Using the model in step 2, generate M complete datasets and extract data according to each feasible sparse design evaluated at feasible time choices.
- (4) Compute the optimality criterion for each $S_{(\tau)}$.
- (5) Select the best sparse design and time choices based on chosen optimality criterion.

In order to apply this, we consider the same set of sparse schemes as previously, applied to the following four possible time point choices:

- (1) (0.5, 1, 2, 4, 9, 12)
- (2) (0.5, 1.5, 2, 4, 9, 12)
- (3) (0.5, 1, 2.5, 4, 9, 12)
- (4) (0.5, 1.5, 2.5, 4, 9, 12)

The results of this application to the toy example are shown in Table 4.4, using the variance of the AUC as optimality criterion. The underlying PK model used is identical to that used in scenario 1. It can be seen that we can indeed improve upon the previous best design by 2.5% by altering the time points. Interestingly, for this particular example, the top ranking scheme is the same, but the choice of time points is changed in order to reduce the variance. It is also worth noting the bottom ranking designs are using time point choice 1 (that is the original time points: 0.5,1,2,4,9,12h); a further indication of how important

it is to include the choice of time points in our design instead of choosing them arbitrarily.

Table 4.4: Top 10 rankings of sparse schemes with choice of time points

Rank	Time Points #	Scheme #	AUC Var
1	3	1003	25.67
2	3	809	25.79
3	2	1003	25.86
4	3	775	25.91
5	3	1039	25.94
6	3	746	25.98
7	3	1001	26.04
8	1	1003	26.05
9	3	750	26.05
10	3	1026	26.10

Table 4.5: Top 10 rankings of sparse schemes with choice of time points by $\Psi = w_1 \text{var}(\widehat{AUC})^* + w_2 \text{var}(\widehat{C}_{max})^*$ with $w_1 = w_2 = 0.5$.

Rank	Time Points #	Scheme #	Ψ
1	3	1003	1.744
2	3	809	1.746
3	3	775	1.751
4	3	1039	1.754
5	3	750	1.754
6	3	796	1.757
7	3	746	1.758
8	3	805	1.759
9	3	1001	1.760
10	3	1026	1.760

We see that when using the same criteria $\Psi = w_1 \text{var}(\widehat{AUC})^* + w_2 \text{var}(\widehat{C}_{max})^*$ as previously discussed with $w_1 = w_2 = 0.5$, the results in Table 4.5 are obtained. From this we see that the choice 3 out of the 4 time points dominates the top 10 schemes chosen.

In order to compute this within any reasonable time, only a small set of possible time points can be used. Ideally, we would want to consider a larger range of time point choices and so in order to approximate this, we reduce the

approach to a two stage procedure. First, choose the optimal set of time points at which to sample. Then using these time points as fixed, we apply Algorithm 1 to find the optimal sparse sampling scheme. It is crucial to emphasize again the importance of using non-compartmental methods in these algorithms as they are specifically designed for trials which use non-compartmental analysis.

We follow the same outline of procedure as previously, using a large number of simulations. Since we can now consider a larger range of time point choices, we refer to this set as the set of feasible time point choices.

Algorithm 3:

- (1) Define the set of feasible time choices $\{t_j\}$.
- (2) Assume an underlying PK model, correlation and variance structure.
- (3) Using the model in step 2, generate M complete datasets and extract data according to each time point choice.
- (4) Compute the optimality criterion for each time choice $\{t_j\}$.
- (5) Select the best choice of time points based on chosen optimality criterion.
- (6) Taking these time points as fixed, apply Algorithm 1.

To define the set of feasible time choices $\{t_j\}$ we must first define the set of feasible sampling time points \mathcal{T} . In order to ensure that we have enough sampling points in the appropriate times to approximate the time versus concentration curve well, we partition the sampling grid T into Z "zones" T_z for $z = 1, 2, \dots, Z$ such that

$$|\mathcal{T}| = \sum_{z=1}^Z |\mathcal{T}_z|.$$

Restrictions are placed on the number of time points that can be taken in each zone, with $\tau_z \leq |\mathcal{T}_z|$ denoting the number of sampling points to be picked from zone \mathcal{T}_z , with $\sum_{z=1}^Z \tau_z = J$ where J is the total number of time points

in the trial. We then generate all possible combinations that adhere to these restrictions, the number of which is $\Upsilon = \prod_{z=1}^Z \binom{|\mathcal{T}_z|}{\tau_z}$

In order to choose the best time points $\{t_j\}_v$ out of the possible Υ choices, we must choose an optimality criterion. Since the variance criterion used previously is not particularly useful for choosing time points independently of the sparse scheme, we instead use a measure of closeness to the population PK profile as optimality criterion. To measure the closeness to the population PK curve using non-compartmental methods, we essentially divide the time range into a very fine grid of G points. At each time point on the grid g , we measure the distance between the linearly interpolated population PK profile given by simulated data under the given scenario, and the underlying population PK profile. We call this distance $d_{g,i}$ for simulated dataset i . Figure 4.6 shows an illustration of measuring such distances for an example dataset and $\{t_j\}_v$.

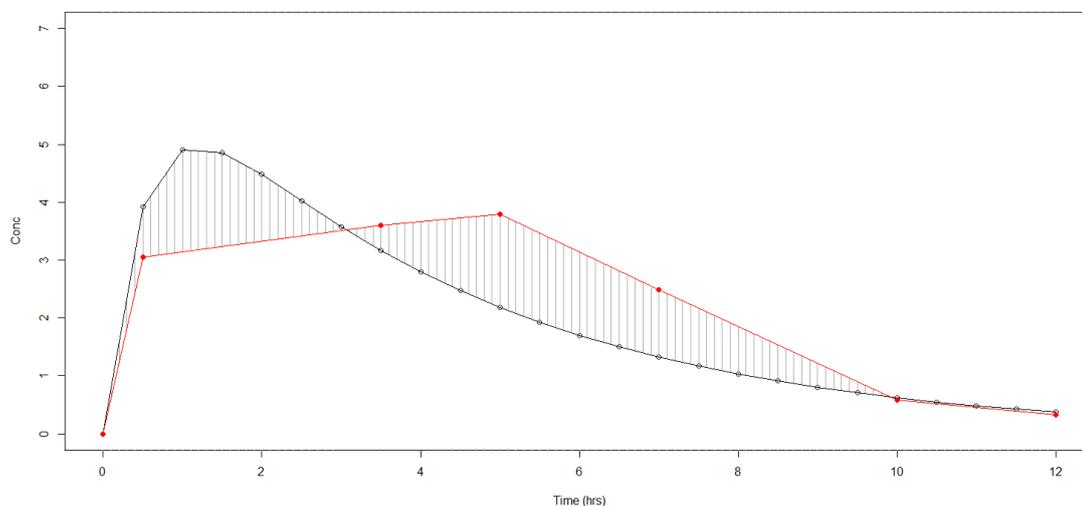


Figure 4.6: Measuring the difference between the true population curve and the simulated data at chosen time points.

For a given choice of time points $\{t_j\}_v$, we average the absolute values of

these distances

$$\Psi(\{t_j\}_v) = \frac{1}{M} \sum_{i=1}^M \sum_{g=1}^G |d_{g,i,v}|$$

The highest ranking of time point choice is therefore the choice that minimizes Ψ . For very large M , this will give us the choice of $\{t_j\}_v$ that is on average closest to the population curve.

For example, if we take \mathcal{T} to be times at half hourly intervals from 0.5 to 12 hours, then we may take the following zones and restrictions: $\mathcal{T}_1 = \{0.5\}$, $\mathcal{T}_2 = \{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$, $\mathcal{T}_3 = \{4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0, 9.5, 10.0, 10.5, 11.0, 11.5\}$ and $\mathcal{T}_4 = \{12.0\}$ with $\tau_1 = 1$, $\tau_2 = 3$, $\tau_3 = 1$ and $\tau_4 = 1$. This gives a total of 525 feasible choices of time points.

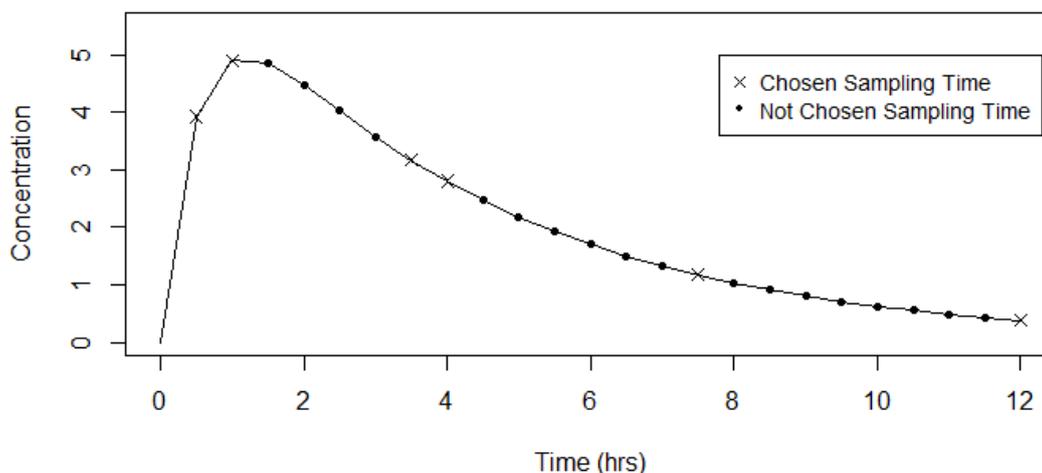


Figure 4.7: Optimal Sampling Time Points

Table G.3 in Appendix G gives the results for the application of this procedure in the same scenarios as previously considered. Given that there are roughly half as many choices as in the previous procedure, it is surprising that the ranks we see in the results are larger here. This suggests that the choice of time points is much less consistent over different scenarios than the choices of

sparse scheme. In terms of the $\{t_j\}_v$ themselves, the top five time point choices are all very similar:

(1) 225: (0.5, 1.0, 3.5, 4.0, 7.5, 12.0)

(2) 260: (0.5, 1.0, 3.5, 4.0, 8.0, 12.0)

(3) 235: (0.5, 1.5, 3.5, 4.0, 7.5, 12.0)

(4) 224: (0.5, 1.0, 6.0, 4.0, 7.5, 12.0)

(5) 234: (0.5, 1.5, 3.0, 4.0, 7.5, 12.0)

Figure 4.7 shows the graphical representation of these optimally chosen time points with respect to the underlying population profile.

Table G.4 in Appendix G and Figure 4.8 show the results for the full procedure, applying our method for choosing the sparse sampling scheme with our chosen optimal time points. Table 4.6 shows the final resulting optimal sparse sampling scheme which shows improvement, for example from the initial time points and optimal design of 6% for the variance of the AUC and 10% from the initial time points and scheme from Chapman et al.¹⁶

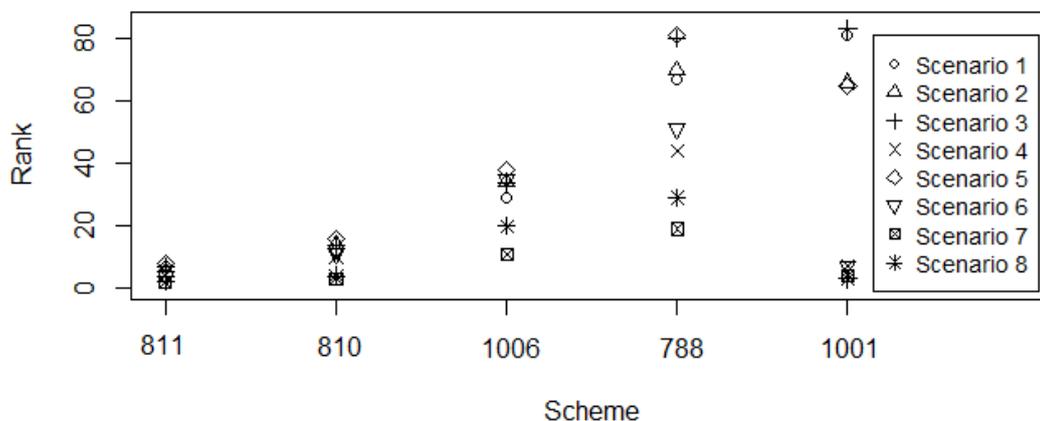


Figure 4.8: Top 5 overall schemes according to minimax criterion applied to equally weighted scaled sum for optimal time points. Ranks for each of the eight scenarios are plotted.

Table 4.6: Optimal Sparse Sampling Scheme with sampling time points (0.5, 1.0, 3.5, 4.0, 7.5, 12.0). \times indicates that the individual subject scheme is shared by the scheme from Chapman et al¹⁶ and \circ indicates that it is not.

Animal number	Sampling timepoint					
	#1	#2	#3	#4	#5	#6
1		\circ	\circ	\circ		
2	\circ	\circ			\circ	
3	\circ				\circ	\circ
4		\times	\times			\times
5			\circ	\circ		\circ
6		\times			\times	\times
7		\times		\times		\times
8	\times		\times	\times		
9	\times			\times	\times	
10	\times		\times		\times	
	n=5	n=5	n=5	n=5	n=5	n=5

4.5 Discussion

The proposed method of producing optimal designs for non-compartmental analysis of pharmacokinetic studies provides a solution to the problem of design for studies using microsampling. It offers a robust approach using non-compartmental methods to consider a worst case scenario situation covering multiple different settings. The obvious advantages to this are that it limits the assumptions made on the underlying model, error and variance structure as it takes into account the possibility of different scenarios for each. The minimax procedure also provides the robustness required, and the non-compartmental methods of analysis fall in line with the analysis methods used in such studies currently. It is critical to use the corresponding method of analysis for the optimality criterion to ensure that the chosen design is indeed optimal for purpose.

The importance of this is highlighted in the difference in results for the comparison of NCA optimality and model based optimality clear to see from Figure 4.4. The use of D -optimality may indeed be useful for trials that use PK modelling, but for those using non-compartmental analysis, it is clear to see that these choices would be sub-optimal. This importance continues to the choice of sampling time points using non-compartmental methods. The optimal time points chosen may seem counter intuitive from a modelling perspective, but are actually desirable for non-compartmental analysis. Without the assumptions that accompany model based methods, our interpretations are solely from the profile extracted from interpolating between measurements. This means that we miss any relevant information between the scheduled sampling times, information that can be filled using modelling assumptions if we were using that approach. For example in the scenarios explored previously, one might expect that in the optimal choice of time points, a sampling time at 2 hours may be more appropriate than at 3.5 hours for the second choice in zone 2. However

if we consider the linear interpolation between the measurements, we must be able to estimate the C_{max} as well as the elimination phase with some accuracy as we cannot rely on modelling assumptions such as that of exponential decay.

Although we implement the procedure based on the example sparse sampling scheme by Chapman et al,¹⁶ it is straightforward to apply to a number of other circumstances. For example, relaxing constraints on the number of subjects required to be sampled at each time point to be an interval rather than a fixed value is simple. In fact any alteration of these constraints is only limited by the computational power to generate and store the schemes and of course the physical feasibility of carrying out the designs. This may also be extended for instance to trials that implement multiple dosing schedules simply by altering the underlying model assumptions in the simulation stage. An advantage of this procedure is that it is able to incorporate multiple PK parameters in its optimality criterion, so even in the case of multiple dosing where there may be more PK parameters that must be estimated, the procedure can still be implemented successfully.

CHAPTER 5

**Paper C: Methods for Non-Compartmental Pharmacokinetic Analysis with
Observations below the Limit of Quantification**

5.1 Introduction

In pharmacokinetic (PK) studies, the objective is to learn about the absorption, distribution, metabolism and excretion (ADME) processes of an externally administered compound by measuring the concentration in bodily tissue such as blood or plasma at a number of time points after administration. However, some of these concentrations are reported as below the limit of quantification of the assay. The limit of quantification (LOQ) is defined as "The lowest concentration at which the analyte can not only be reliably detected, but at which some predefined goals for bias and imprecision are met".⁴ These observations are often referred to as "BLOQ" or "BQL" and require special attention in the data analysis.

Dealing with BLOQ observations when modelling is used has been vastly explored in the literature. The most notable of which is the contribution from Beal,¹⁰ describing seven key methods for fitting PK models when BLOQ observations are present. However, due to regulatory requirements,^{29,61} many PK studies use non-compartmental analysis (NCA) yet statistical methods for dealing with BLOQ observations in NCA are very much lacking.

The two strategies of PK analysis, modelling^{11,20} and NCA^{15,32,43} differ in their approach on a number of levels. Whilst the modelling approach offers the advantage of unstructured sampling schemes (i.e. fewer restrictions on the sparsity and structure of the sampling schedule), this comes at the cost of uncertainty over the model choice and potential technical difficulties in fitting the PK model. Many of the methods discussed by Beal¹⁰ involve either discarding BLOQ observations, replacing them with $LOQ/2$ or replacing them with 0 before proceeding to fit a PK model using such methods as maximum likelihood or least squares regression. Assumptions must be made about the underlying ADME process in order to use these approaches, and each of the approaches

is appropriate for different settings and require different assumptions. In NCA however, no such assumptions must be made, although some kind of approximation, usually linear, of the concentrations between observed time points must be made. The purpose of both approaches is to estimate PK parameters such as the area under the concentration versus time curve (AUC), a measure of total exposure of the compound, or the maximum concentration of the compound (C_{max}).

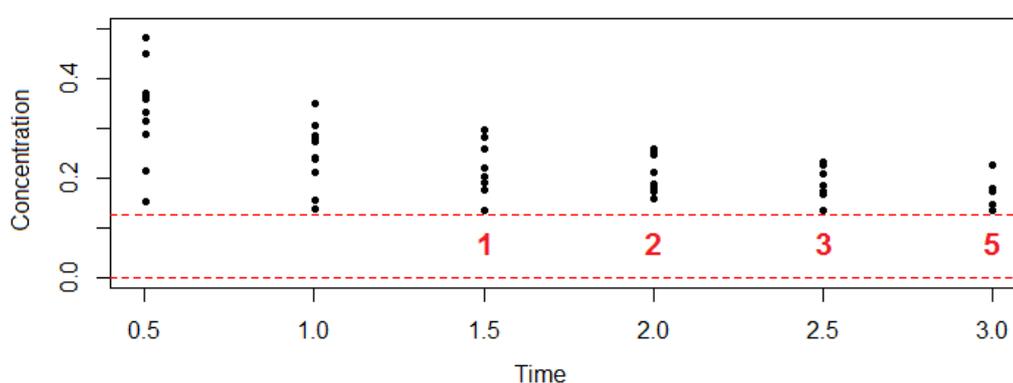


Figure 5.1: A motivating example, the red numbers indicate the number of observations that are BLOQ for that time point.

Figure 5.1 shows an illustration of a motivating example, with black points indicating observed responses, dashed red lines the levels in between which concentrations cannot be reliably detected and red number the number of responses in that region at a given time point. This data is generated from a one-compartmental IV bolus dose model with mixed effects, such as that described by Beal.¹⁰ With a large number of responses below the limit of quantification, and those responses that are above have low concentration values, the contribution of the BLOQ responses will vastly affect any estimate of the AUC and its variance; this is the PK parameter that is most affected by how BLOQ responses are dealt with. In the following, we introduce seven methods for including BLOQ observations in NCA for PK studies, and hence make as few as-

sumptions on the data as possible.

5.2 Methods

We introduce seven methods for comparison, focusing on how each of the methods impacts the \widehat{AUC} and its variance. We also focus on full sampling designs in our evaluations, although the methods may be extended to sparse sampling schemes without further complication. The first two methods are simple imputations, replacing the BLOQ observations with 0 and $LOQ/2$ and proceeding with the NCA approach. These are the current approaches applied in practice³⁵ and hence the benchmark upon which we wish to improve on. The remaining five methods use varying techniques to either impute values onto BLOQ responses or to approximate the summary statistics of the non-compartmental approximation of the concentration versus time curve.

For all methods, two different error structures, additive and multiplicative are considered. We assume concentrations from n subjects, labelled $i = 1, \dots, n$, are observed at J timepoints t_j for $j = 1, \dots, J$.

Additive Error Model:

The additive error model is defined as:

$$y_{ij} = \mu_j + \epsilon_{ij}$$

where y_{ij} is the observed response for subject i at the j th timepoint. The μ_j represents the population mean response at the j th timepoint. The ϵ_{ij} are the differences between the y_{ij} and μ_j and are assumed to be normally distributed. In this case we use the arithmetic mean $\bar{y}_j = \frac{1}{n} \sum_{i=1}^n y_{ij}$ as the basis for estimating the AUC. Assuming all observations were above the LOQ, the estimate of the population AUC is written as:

$$\widehat{AUC}^{(A)} = \sum_{j=1}^J \omega_j \bar{y}_j, \quad (5.1)$$

where \bar{y}_j is defined as above and ω_j are weights defined as:

$$\begin{aligned}\omega_j &= \frac{t_{j+1} - t_{j-1}}{2} \quad \text{for } j = 1, 2, \dots, (J-1), \\ &= \frac{t_J - t_{j-1}}{2} \quad \text{for } j = J.\end{aligned}$$

The variance approximation of the \widehat{AUC} for batch designs can be used:

$$\text{var} \left(\widehat{AUC}^{(A)} \right) = \sum_{b=1}^B \frac{s_b^2}{n_b}, \quad (5.2)$$

where

$$s_b^2 = \frac{1}{n_b - 1} \sum_{i=1}^{n_b} \left(\sum_{j \in J_b} \omega_j y_{ij} - \frac{1}{n_b} \sum_{k=1}^{n_b} \sum_{j \in J_b} \omega_j y_{kj} \right)^2.$$

with $J_b \subseteq \{1, \dots, J\}$ the indices of time points investigated in batch $b = 1, \dots, B$ and n_b the number of subjects in batch b . This generalized form of the variance approximation can then be used in the case of complete and serial sampling, which are special cases of batch sampling.

Multiplicative Error Model:

It is often more typical for the assumption on the errors to be multiplicative. In this scenario the arithmetic mean for the data is no longer a satisfactory measure of central tendency of the concentration at each time point, since it does not conform with the error model. Instead, we will use the geometric mean³¹ defined for time t_j with n observations as $(\prod_{i=1}^n x_{ij})^{\frac{1}{n}}$ or equivalently $e^{\frac{1}{n} \sum_{i=1}^n \log x_{ij}}$. The model is then:

$$y_{ij} = \mu_j e^{\epsilon_{ij}}$$

which we can rewrite as:

$$\log y_{ij} = \log \mu_j + \epsilon_{ij}$$

where y_{ij} is the observed response for subject i at the j th timepoint. The μ_j represent the population mean response at the j th timepoint. The ϵ_{ij} are the differences between the $\log y_{ij}$ and $\log \mu_j$ and the ϵ_{ij} are assumed to be normally distributed. Letting $c_{ij} = \log y_{ij}$, we have the geometric mean of the observations

per time point $e^{\bar{c}_j}$ where $\bar{c}_j = \frac{1}{n} \sum_{i=1}^n c_{ij}$ and use this as our mean estimate of the response per timepoint when estimating AUC . Assuming all observations are above the LOQ, the estimate of the population AUC is:

$$\widehat{AUC}^{(G)} = \sum_{j=1}^J \omega_j e^{\bar{c}_j}, \quad (5.3)$$

with ω_j and \bar{c}_j defined as previously. Using the variance approximation of the \widehat{AUC} for batch designs and a first order Taylor approximation results in:

$$\text{var} \left(\widehat{AUC}^{(G)} \right) = \sum_{b=1}^B s_b^{2(G)}, \quad (5.4)$$

where

$$s_b^{2(G)} = (\boldsymbol{\omega}_b e^{\bar{\mathbf{c}}_b})^T \widehat{\mathbf{V}}_b (\boldsymbol{\omega}_b e^{\bar{\mathbf{c}}_b}), \quad (5.5)$$

where $\boldsymbol{\omega}_b e^{\bar{\mathbf{c}}_b}$ is a vector of length $|J_b|$ with the j th element equaling $\omega_j e^{\bar{c}_j}$ and $\widehat{\mathbf{V}}_b$ is the variance-covariance matrix of observed log transformed data c_{ij} for $j \in J_b$. The denominator of n_b is not included in this form of (5.4) compared to (5.2) as the population estimate $s_b^{2(G)}$ (as opposed to the individual estimate s_b^2 previously) includes this multiplicative factor.

5.2.1 Method 1: Replace BLOQ values with 0

An easy strategy that is currently used is to replace any value below the limit of quantification by 0³⁵ and proceed with traditional NCA methods on the augmented data. When considering geometric means this method is infeasible for calculating any estimate of variance, as this involves estimating the variance of log-transformed data, therefore any log-transformed BLOQ values are undefined.

5.2.2 Method 2: Replace BLOQ values with $\frac{LOQ}{2}$

Similar to Method 1; any value below the limit of quantification is replaced by LOQ/2 and traditional analysis methods are used.

5.2.3 Method 3: Regression on Order Statistics (ROS) Imputation

Regression on Order Statistics³⁴ is a semi-parametric method of dealing with censored data and has its origins in environmental statistics. It involves replacing the censored values with different values, i.e. for a dataset with more than one BLOQ response, different values are imputed onto these multiple responses, as opposed to methods 1 and 2 which replace all BLOQ values with the same value. In order to apply this method, we consider each time point in turn, starting with the earliest time point where a BLOQ value is observed. If the error model is multiplicative, transform the data by $c_{ij} = \log y_{ij}$. The premise of this method is to calculate plotting positions for both observed and censored observations, similar to a QQ plot, then using a linear regression to impute values on the BLOQ observations. The method proceeds in detail as follows:

- (1) Identify BLOQ values
- (2) Start at the earliest time point for which a BLOQ value is observed, labelled the j th timepoint. Define $n - m$ as the number of detected responses above or equal to the previously defined LOQ, and m the number of BLOQ values at this time point. From this, we estimate the empirical exceedance probability by the proportion of the sample greater than or equal to the LOQ:

$$pe = \frac{m}{n}.$$

- (3) We then calculate the plotting positions for each of the $n - m$ (ordered from lowest to highest) detected values:

$$pd_j = (1 - pe) + \left(\frac{k_d}{n - m + 1} \right) pe \quad \text{for } k_d = 1, \dots, n - m.$$

- (4) We also calculate the plotting positions for each of the m censored *BLOQ* values:

$$pc_j = \left(\frac{k_c}{m+1} \right) (1 - pe) \quad \text{for } k_c = 1, \dots, m.$$

- (5) Compute a normal quantile for each value of pd_{k_d} and pc_{k_c} as:

$$z_{k_d}^d = \Phi^{-1}(pd_{k_d}), \quad z_{k_c}^c = \Phi^{-1}(pc_{k_c})$$

- (6) Construct a simple linear regression on the $z_{k_d}^d$ and the (ordered from lowest to highest) y_{k_d} .

$$E(y_{k_d}) = \hat{\alpha} + \hat{\beta} z_{k_d}^d.$$

- (7) Then we calculate imputed values for the BLOQ values:

$$y_{k_c}^c = \hat{\alpha} + \hat{\beta} z_{k_c}^c.$$

These are ordered by the ordering at the previous timepoint, i.e. the highest imputed value is assigned to the subject that has the highest response value on the previous time point.

- (8) This continues to the next time point until all BLOQ values are imputed.
- (9) We transform back onto the observation scale if necessary and compute the PK parameter on the augmented data.

Figure 5.2.3 shows an example of the application of this method for a single time point. Proceed with NCA methods using equations 5.1 - 5.4.

5.2.4 Method 4: Maximum Likelihood per timepoint (Summary)

This method does not impute values for the BLOQ observations, but instead provides summary statistics for the concentration at each timepoint under the assumption that each time point is independent. For the j th timepoint, we

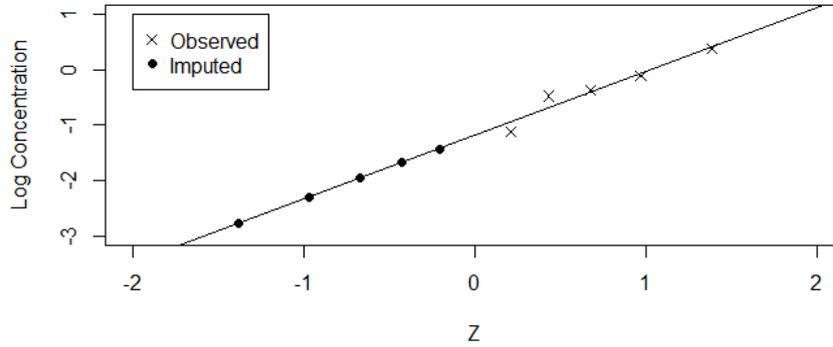


Figure 5.2: Regression on Order Statistics example illustrates how imputed values are calculated.

consider a censored likelihood of:

$$L(\mu_j^{(A)}, \sigma_j^{(A)2}) = \left(\Phi \left(\frac{LOQ - \mu_j^{(A)}}{\sigma_j^{(A)}} \right) \right)^m \prod_{i=1}^{n-m} \frac{1}{\sigma_j^{(A)} \sqrt{2\pi}} \exp \left(\frac{-(y_{ij} - \mu_j^{(A)})^2}{2\sigma_j^{(A)2}} \right),$$

for the assumption of additive errors, and

$$L(\mu_j^{(G)}, \sigma_j^{(G)2}) = \left(\Phi \left(\frac{\log LOQ - \mu_j^{(G)}}{\sigma_j^{(G)}} \right) \right)^m \prod_{i=1}^{n-m} \frac{1}{e^{c_{ij}} \sigma_j^{(G)} \sqrt{2\pi}} \exp \left(\frac{-(c_{ij} - \mu_j^{(G)})^2}{2\sigma_j^{(G)2}} \right),$$

for the assumption of multiplicative error. We maximize over μ_j and σ_j^2 to obtain estimates $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ for each timepoint t_j . These estimates are then used in the calculation of the point estimate of the AUC and its variance.

5.2.5 Method 5: Maximum Likelihood per timepoint (Imputation)

This method is, in essence, a hybrid of methods 3 and 4. It combines the superior estimation of the mean and variance per time point that censored maximum likelihood brings,¹⁴ and retains the structure of the between time point relationship that the imputation methods uses. It begins with using maximum likelihood to obtain values of $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ for each timepoint in the same manner

as Method 4 and subsequently uses these estimates to impute values onto the BLOQ responses.

Additive Error:

Estimate the probability of a response being BLOQ :

$$p_{BLOQ} = P(Y_j < LOQ) = \Phi \left(\frac{LOQ - \hat{\mu}_j^{(A)}}{\hat{\sigma}_j^{(A)}} \right),$$

where Φ is the cumulative distribution function of the standard Normal distribution.

$$p_{k_c} = \frac{k_c}{m+1} p_{BLOQ} \text{ for } k_c = 1, \dots, m.$$

Transform to response scale using the inverse Normal cumulative distribution function (Φ^{-1}).

$$y_{k_c}^c = \hat{\mu}_j^{(A)} + \hat{\sigma}_j^{(A)} \Phi^{-1}(p_{k_c}).$$

These imputed values are ordered in the same way as in M3.

Multiplicative Error:

Estimate the probability of a response being BLOQ:

$$p_{BLOQ} = P(Y_j < \log LOQ) = \Phi \left(\frac{\log LOQ - \hat{\mu}_j^{(G)}}{\hat{\sigma}_j^{(G)}} \right),$$

where F_j is the cumulative distribution function of the Normal distribution with parameters $\hat{\mu}_j^{(G)}$ and $\hat{\sigma}_j^{(G)^2}$. We then equally space the probabilities for the censored observations

$$p_{k_c} = \frac{k_c}{m+1} p_{BLOQ} \text{ for } k_c = 1, \dots, m.$$

Transform to response scale using the inverse Normal cumulative distribution function (Φ^{-1}) and exponentiate.

$$y_{k_c}^c = e^{\hat{\mu}_j^{(G)} + \hat{\sigma}_j^{(G)} \Phi^{-1}(p_{k_c})}.$$

These imputed values are ordered in the same way as in M3. Proceed with NCA methods using equations 5.1 - 5.4.

5.2.6 Method 6: Full Likelihood

This method takes into account correlation between responses at different timepoints by considering all timepoints together. In this approach we estimate the covariance matrix of observations assuming a multivariate Normal or Log-normal distribution.

We consider that the data are n independent and identically distributed observations from a $MVN_J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Our objective is to estimate $\hat{\boldsymbol{\mu}}$ & $\hat{\boldsymbol{\Sigma}}$, the maximum likelihood estimates for the mean and variance of the observations assuming a multivariate normal distribution. From these we can then calculate the \widehat{AUC} and its variance. For each subject $i = 1, \dots, n$, we partition into censored and non-censored observations:

$$\mathbf{y}_i^{(c)} = \mathbf{y}_{i, \{j: \gamma_{ij}=1\}}$$

$$\mathbf{y}_i^{(-c)} = \mathbf{y}_{i, \{j: \gamma_{ij}=0\}}$$

where γ_{ij} is an indicator, taking the value 1 if the observation is censored and 0 otherwise.

We then partition our parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, with superscripts (c) for censored parameters and $(-c)$ for uncensored, with $\boldsymbol{\Sigma}^{(c)(-c)}$ the censored/uncensored covariance matrix.

Then the conditional (on the uncensored values) distribution of the censored observations following Eaton²² is:

$$MVN(\boldsymbol{\mu}^{(c)} + \boldsymbol{\Sigma}^{(c)(c)}(\boldsymbol{\Sigma}^{(-c)(-c)})^{-1}(\mathbf{y}^{(-c)} - \boldsymbol{\mu}^{(-c)}), \boldsymbol{\Sigma}^{(c)(c)} - \boldsymbol{\Sigma}^{(c)(-c)}\boldsymbol{\Sigma}^{(-c)(-c)^{-1}}\boldsymbol{\Sigma}^{(-c)(c)})$$

and the log likelihood can be found as:

$$\sum_{i=1}^n \left(\log(F(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}, LOQ)) + \log(f(\boldsymbol{\mu}^{(-c)}, \boldsymbol{\Sigma}^{(-c)}, \mathbf{y}_i^{(-c)})) \right),$$

where F and f are the *cdf* and *pdf* of the multivariate normal distribution respectively. This is maximized over $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to give MLE $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. The point estimate and variance of the *AUC* can be estimated from this as follows:

$$\widehat{AUC}^{(A)} = \boldsymbol{\omega}^T \hat{\boldsymbol{\mu}}$$

$$\text{var} \left(\widehat{AUC}^{(A)} \right) = \boldsymbol{\omega}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\omega}$$

Similarly, one may construct exactly the same log likelihood but with the log-normal distribution. Here the parameter estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ represent the mean vector and covariance matrix of the log-transformed data. The point estimate and variance of the *AUC* can be estimated from this as follows:

$$\widehat{AUC}^{(G)} = \boldsymbol{\omega}^T e^{\hat{\boldsymbol{\mu}}}$$

$$\text{var} \left(\widehat{AUC}^{(G)} \right) = (\boldsymbol{\omega} e^{\hat{\boldsymbol{\mu}}})^T \hat{\boldsymbol{\Sigma}} (\boldsymbol{\omega} e^{\hat{\boldsymbol{\mu}}})$$

5.2.7 Method 7: Kernel Density Imputation

This methods differs from the previous ones in that it does not assume a specific error distribution, but estimates it from the data. We do still however consider the two cases, using arithmetic (comparable to additive error assumptions) and geometric (comparable to multiplicative error assumptions) means of the responses to estimate the *AUC* and its variance. The basis of this method uses kernel density estimation,⁶⁷ which, given Y_i and a kernel function K , estimates the density of the Y_i as follows:

$$\hat{f}(y) = \frac{1}{nh} \sum_{i=1}^n K \left(\frac{y - Y_i}{h} \right),$$

where h is a parameter known as the bandwidth, which can be prespecified or optimized over. In the following, we use a Gaussian Kernel, $K(t) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)t^2}$,

and use Silverman's 'Rule of Thumb'⁶⁷ to calculate the bandwidth parameter, $h = 1.06\hat{\sigma}n^{\frac{1}{5}}$. This h is calculated each time the density \hat{f} is re-estimated, so that each \hat{f} has a recalculated bandwidth h .

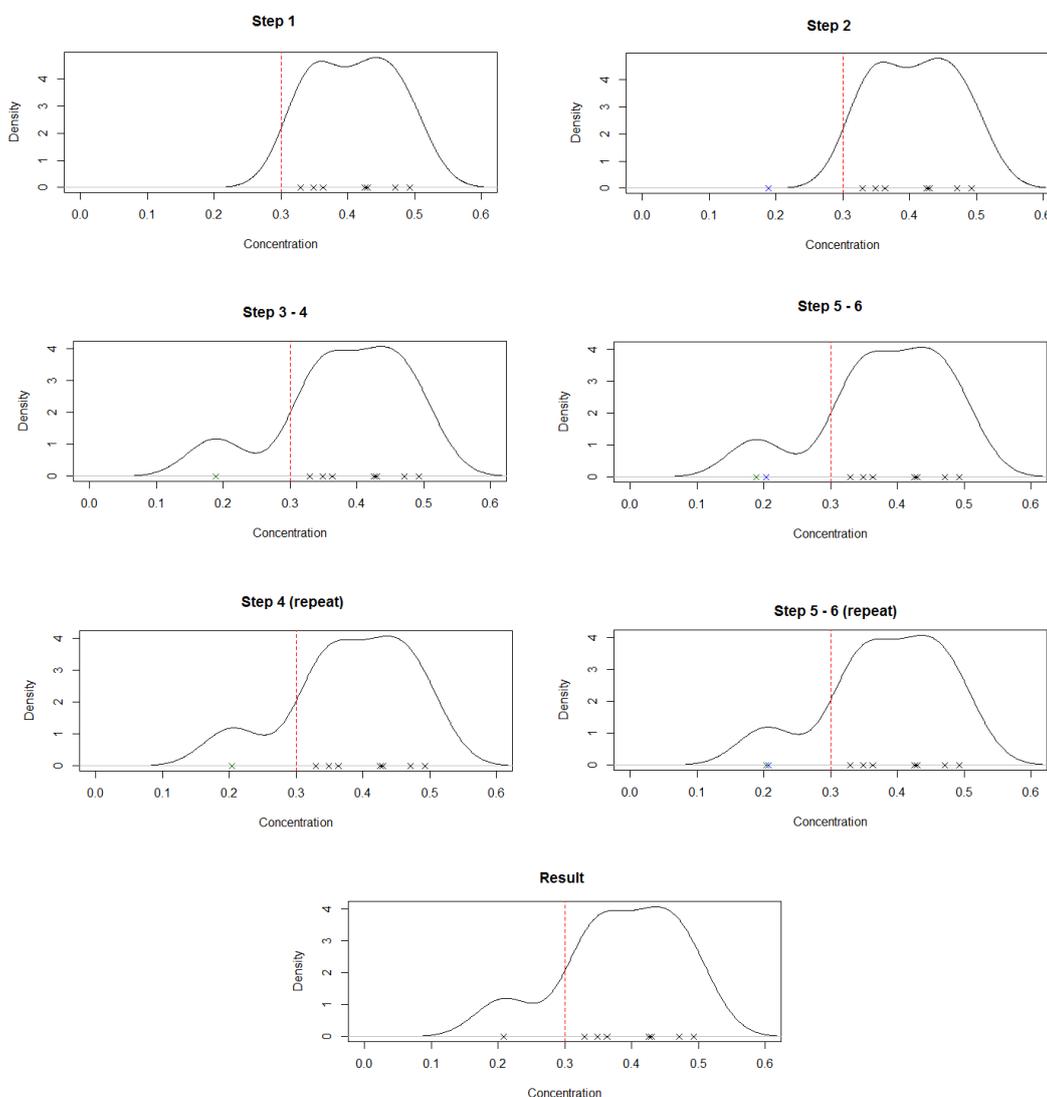


Figure 5.3: A graphical illustration of the KD algorithm. Blue crosses indicate a new k_i calculated based on the current \hat{f}_i . Green crosses indicate previous k_i values. The red line indicates the LOQ.

For each timepoint with BLOQ observations:

- (1) Calculate \hat{f}_0 based on uncensored data $\mathbf{y}^{(-c)}$.
- (2) Compute $k_0 = E_{\hat{f}_0}(Y|Y < LOQ)$.
- (3) Initialize $i = 1$.
- (4) Append k_{i-1} to the uncensored data $\mathbf{y}^{(-c)}$ and recalculate \hat{f}_i .
- (5) Compute $k_i = E_{\hat{f}_i}(Y|Y < LOQ)$
- (6) Let $i = i + 1$
- (7) Repeat steps 4-6 until $|k_i - k_{i-1}| < \epsilon$ for some very small prespecified value ϵ . This k_i is the value to be imputed.

This process is repeated as many times as needed to get m imputed values for this timepoint. Figure 5.3 shows an example of the application of this method. Proceed with NCA methods using equations 5.1 - 5.4.

5.2.8 Example Application

Table 5.1 shows how the \widehat{AUC} and its variance are affected by the choice of method of dealing with BLOQ values for the motivating example introduced in Figure 5.1.

There is a wide range of values for \widehat{AUC} and its variance, all for the same dataset. In both the cases of assumptions of arithmetic and geometric mean, Method 1 unsurprisingly has the lowest \widehat{AUC} . In the case of using geometric means, this is severely lower than all other methods. Method 4 has the lowest variance in each case, as expected. When assuming an additive error model, Method 6 failed to produce any results as the assumptions deviated too far from the characteristics of the dataset.

Table 5.1: The application of the seven methods to the motivating example illustrated in Figure 5.1.

	Additive Error Model / Arithmetic Means		Multiplicative Error Model / Geometric Means	
	\widehat{AUC}	$\text{var}(\widehat{AUC})$	\widehat{AUC}	$\text{var}(\widehat{AUC})$
Truth	0.6008	0.0275	0.5704	0.0321
Method 1	0.5623	0.0453	0.2778	NA
Method 2	0.5889	0.0322	0.5481	0.0399
Method 3	0.6012	0.0276	0.5861	0.0267
Method 4	0.6008	0.0062	0.5800	0.0067
Method 5	0.6027	0.0272	0.5817	0.0286
Method 6	NA	NA	0.6011	0.0090
Method 7	0.5963	0.0288	0.5636	0.0337

5.3 Results

In order to evaluate the performance of the seven methods previously discussed, they are all applied to simulated data. Following an example from Beal,¹⁰ the following model is used to generate data at time t :

$$y(t) = C(t) \exp(e(t)),$$

where $C(t)$ is the PK model and $e(t)$ is the error model. The PK model is a one compartmental IV dose model:

$$C(t) = \frac{\text{dose}}{V_d} \exp(CL \cdot t).$$

The error model is Normally distributed $e(t) \sim N(0, h(t))$ with

$$h(t) = 0.03 + 0.165 \frac{C(t)^{-1}}{C(1.5)^{-1} + C(t)^{-1}}.$$

Data are generated at times 0.5, 1, 1.5, 2, 2.5, 3 hours. Two scenarios are considered, using fixed effects and using mixed effects. For fixed effects, the parameters take values $CL = 0.693$, $V_d = 1$ and $\text{dose} = 1$ while for mixed effects, $CL = \widetilde{CL} \exp(\eta_1)$ and $V_d = \widetilde{V}_d \exp(\eta_2)$, with $\widetilde{CL} = 0.693$, $\widetilde{V}_d = 1$, $\eta_1 \sim N(0, \omega_1^2)$ and $\eta_2 \sim N(0, \omega_2^2)$, $\text{corr}(\eta_1, \eta_2) = 0$ and $\omega_1 = \omega_2 = 0.2$. An example of such a dataset is illustrated in Figure 5.4 A case with a smaller clearance and dose

($CL = 0.231$, $dose = 0.25$) is also considered, where the contributions to the AUC will be more affected by how the BLOQ observations are dealt with, an example of this is illustrated in Figure 5.5. Many more responses are BLOQ for this second example, with one subject having all observed responses BLOQ.

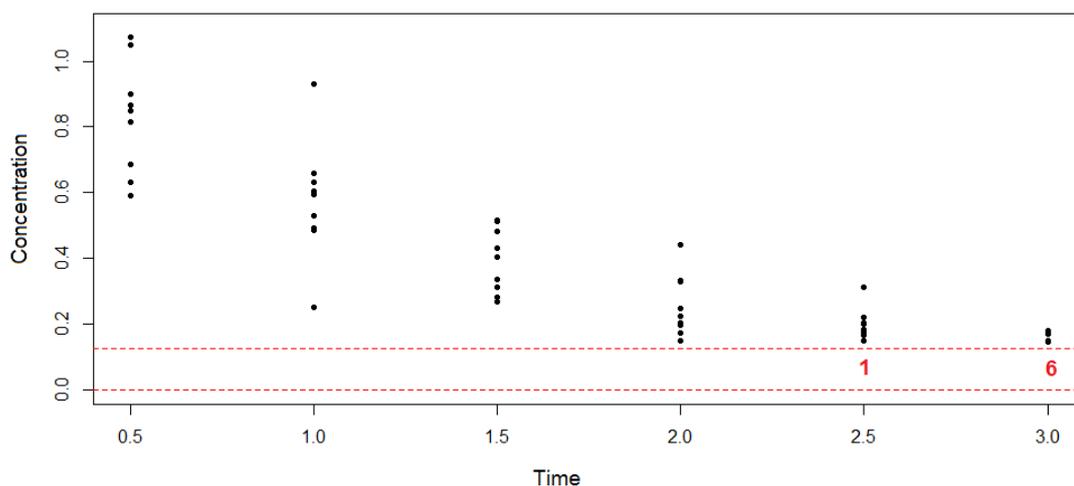


Figure 5.4: Illustration of Example from Beal,¹⁰ the red numbers indicate the number of observations that are BLOQ for that time point.

When applying Method 6, we must constrict the dimensionality of the parameter space in order for it to be feasible to realistically used. For example, for this particular set up with six time points considered, there are 27 free parameters to optimize over. This means that optimizing over the multivariate normal log likelihood takes a very long time and is often unsuccessful or unstable. When performing the maximum likelihood procedure, we therefore now restrict the elements of the variance matrix Σ to be 0 for all of the covariances

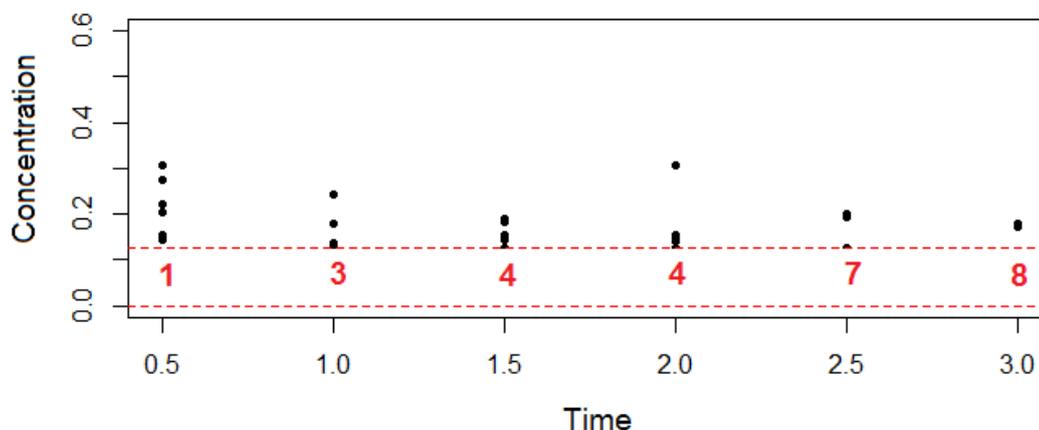


Figure 5.5: Illustration of Example extended from Beal,¹⁰ with smaller clearance and dose. The red numbers indicate the number of observations that are BLOQ for that time point.

that are not for consecutive time points.

$$\Sigma = \begin{pmatrix} var_1 & cov_{12} & 0 & 0 & 0 & 0 \\ cov_{12} & var_2 & cov_{23} & 0 & 0 & 0 \\ 0 & cov_{23} & var_3 & cov_{34} & 0 & 0 \\ 0 & 0 & cov_{34} & var_4 & cov_{45} & 0 \\ 0 & 0 & 0 & cov_{45} & var_5 & cov_{56} \\ 0 & 0 & 0 & 0 & cov_{56} & var_6 \end{pmatrix}$$

In terms of suitability, Methods 1, 3 and 6 are not applicable to all datasets and hence are less useful. Method 1 cannot compute any estimates when using geometric means as summaries due to non-finite values resulting from taking logs. Method 3 requires fitting a linear regression on responses per timepoint and hence is infeasible in scenarios with high levels of censoring that can result in only one response above the LOQ for a given time point. In the simulations, around 1% of the simulated datasets did not result in an estimate. Method 6 is even more unstable, not giving results at all when assumptions on the distri-

bution of the responses is incorrectly specified, and even when the distribution is correctly assumed, up to 3% of simulated datasets do not give results for the fixed effects model and as much as 24% when using mixed effects.

It is desirable for a method to approximate the non-compartmental point estimate and variance of the \widehat{AUC} as closely as possible to that which we would estimate were we to know the true concentration for every response, regardless of a limit of quantification. Therefore as a measure of performance of the methods, we use closeness to the full non-compartmental estimates of these values, $\widehat{AUC}_{BLOQ} - \widehat{AUC}_T$ and $\text{var}(\widehat{AUC}_{BLOQ}) - \text{var}(\widehat{AUC}_T)$, where subscript *BLOQ* indicates one of the methods of dealing with BLOQ values has been used, and subscript *T* indicates the estimate has been calculated as if we knew all observed values, regardless of LOQ.

The results for comparison of the seven methods are presented graphically in two ways: box plots and colour plots. Summary tables of more detailed results can be seen in Appendix H. The performance measures are plotted over 1000 simulations, with standard deviations plotted as opposed to variance for consistency in units. The box plots (Figures 5.6 and 5.7) show the spread of these measures over the simulations. For the most appropriate methods, the colour plots (Figure 5.8) then show the relationship between the deviation from the point estimate and the deviation from the standard deviation over all simulations.

For many methods, Figures 5.6 and 5.7 show the estimate of the \widehat{AUC} underestimates the true non-compartmental estimate. Although Method 5 shows promising performance averaging over the simulations, there is a wider spread over the simulations reaching above and below the truth. Methods 2, 3, 5 and 7 show the most promising results, and are therefore compared using the colour plots. The colour plots in Figure 5.8 are a clear and direct comparison between the four best performing of seven methods, with Method 7 a clear front runner.

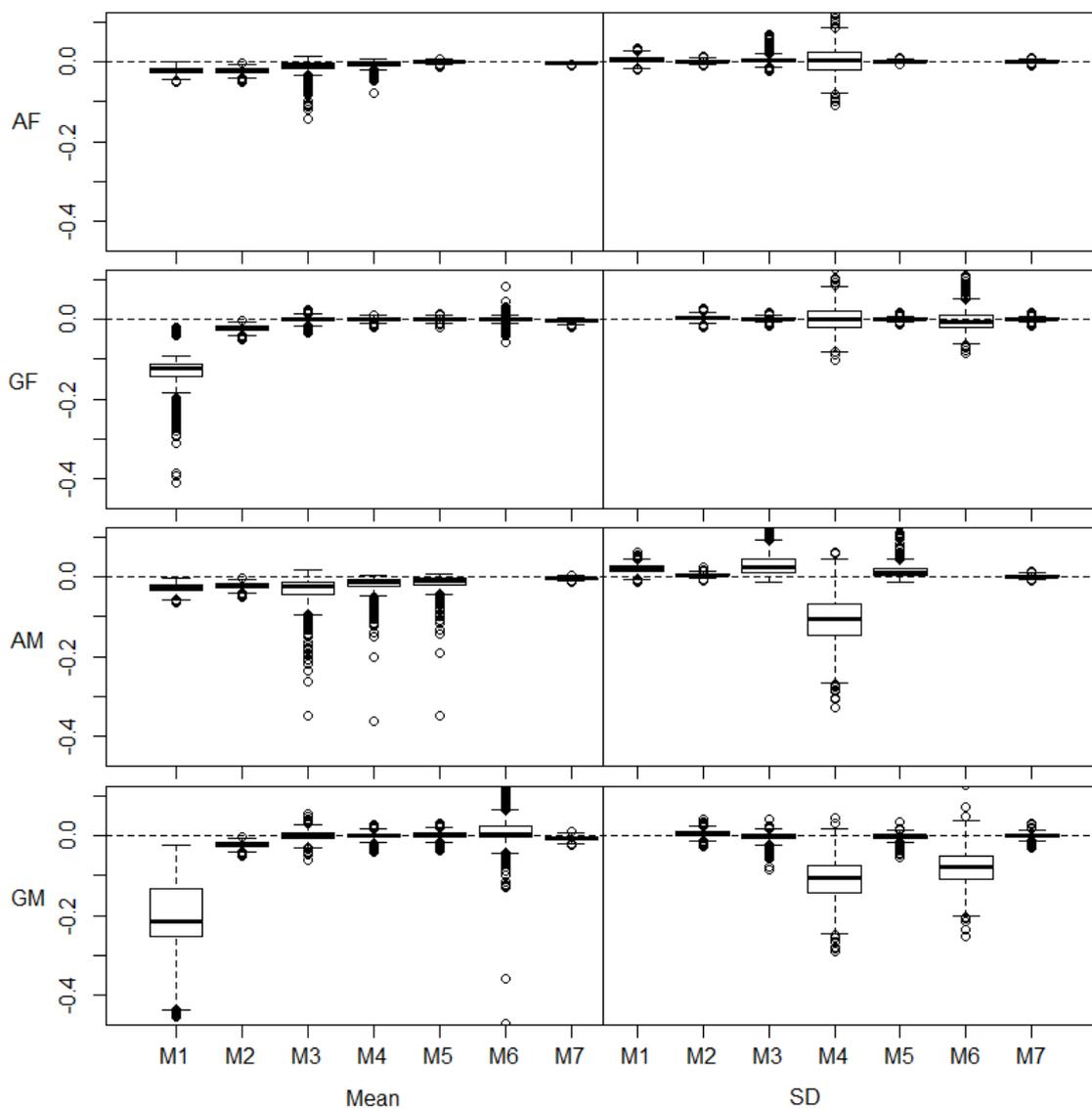


Figure 5.6: Results showing deviation from the non-compartmental \widehat{AUC} and its variance with data generated from models with higher dose and clearance. Results over 1000 simulations. (F)=Data generated using fixed effects model. (M)=Data generated using mixed effects model. (A)=Analysed using arithmetic means. (G)=Analysed using geometric means. M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation.

It gives the most consistent results across all simulations and these differences are closest to zero.

As expected, Method 1 underestimates the non-compartmental \widehat{AUC} . One

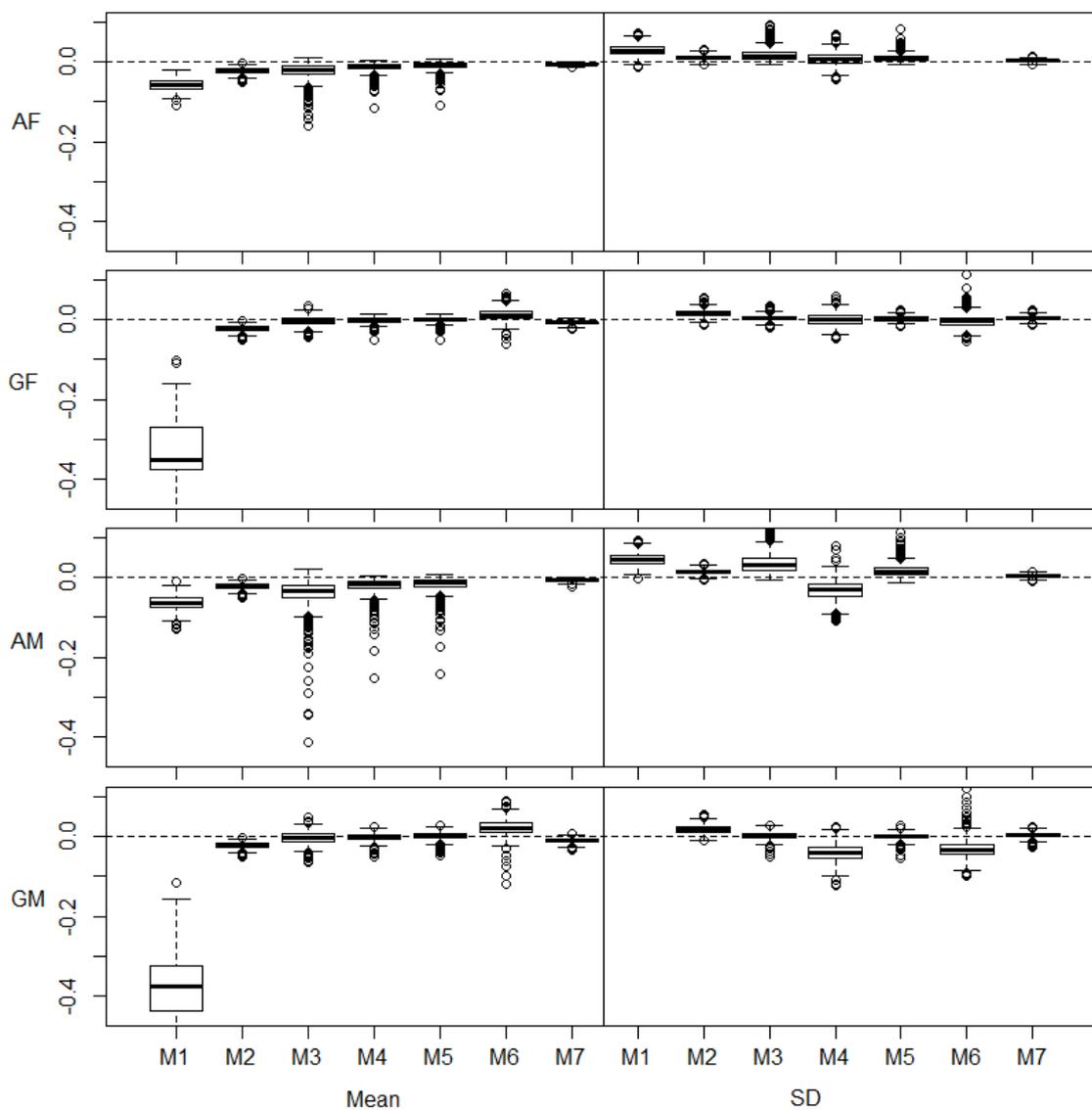


Figure 5.7: Results showing deviation from the non-compartmental \widehat{AUC} and its variance with data generated from models with lower dose and clearance. Results over 1000 simulations. (10 subjects, 6 timepoints) (F)=Data generated using fixed effects model. (M)=Data generated using mixed effects model. (A)=Analysed using arithmetic means. (G)=Analysed using geometric means. M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation.

may expect that by imputing the same value onto all BLOQ responses, the estimate of the variance would be underestimated. However, when applied to datasets generated using the mixed effects models, this method overestimates

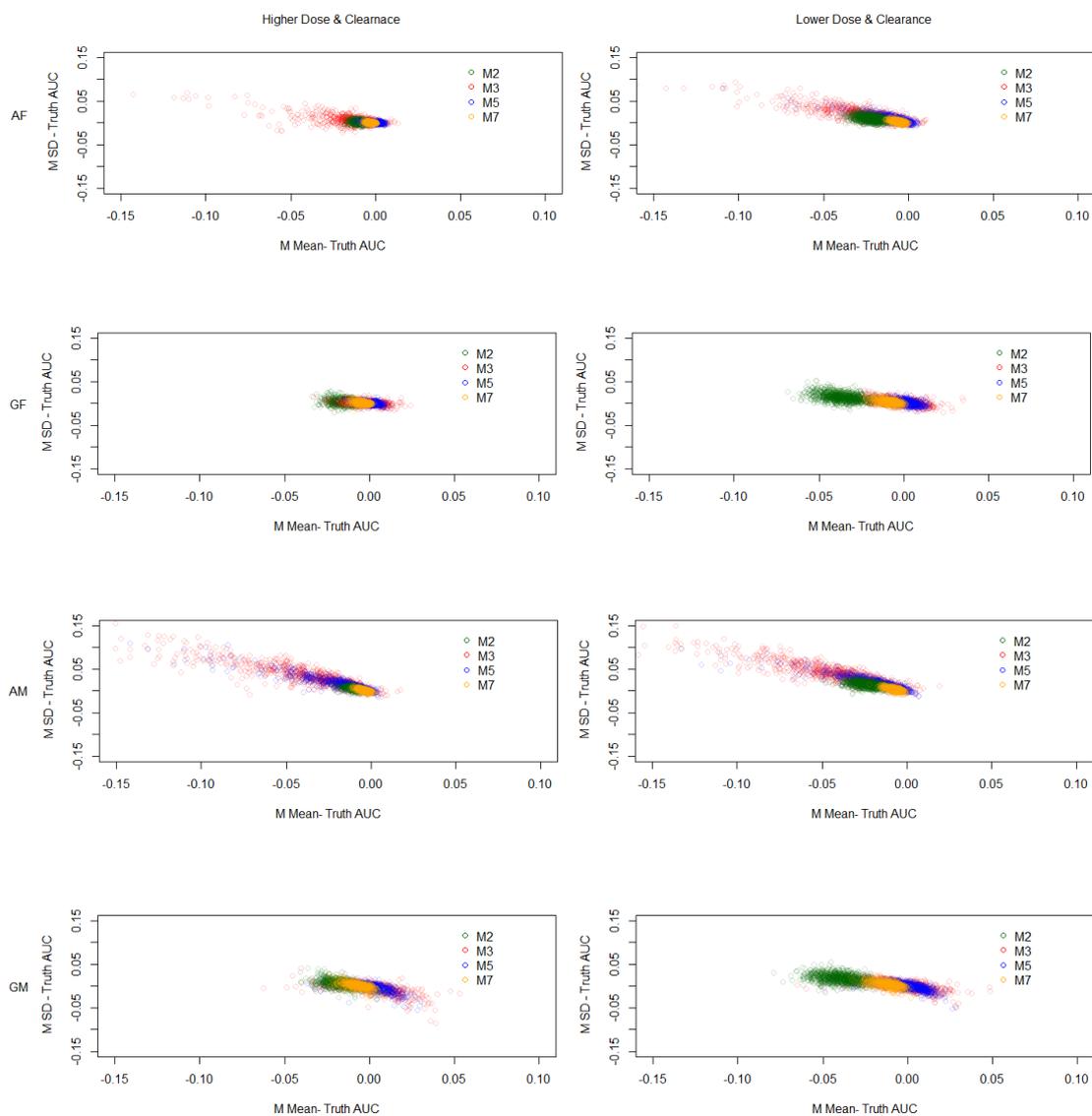


Figure 5.8: Results showing deviation from the non-compartmental \widehat{AUC} and its variance with data generated from both models. Results over 1000 simulations. (10 subjects, 6 timepoints) (F)=Data generated using fixed effects model. (M)=Data generated using mixed effects model. (A)=Analysed using arithmetic means. (G)=Analysed using geometric means. M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M5: ML per timepoint Imputation, M7: Kernel Density Imputation.

the variance of the \widehat{AUC} . The more the point estimate is underestimated, the more the variance is overestimated. This is because as the imputed values draw

the point estimate of the \widehat{AUC} towards 0, they also increase the deviations of the individual observations from the mean value. This method, as pointed out earlier, is unsuitable when considering the geometric means as methods of summary.

Method 2 shows reasonable results, however still underestimating the \widehat{AUC} and overestimating the variance, especially when the data is analysed using geometric means. This is to be expected as the imputation of $LOQ/2$ is based on assumptions consistent with analysis using arithmetic means.

We see a similar trend as with Method 1 with Method 3 in the case of using arithmetic means, but in an even more extreme way. This method underestimates the \widehat{AUC} and overestimates the variance even more, especially when the data is generated using the mixed effects model. This method assumes a normal distribution of concentrations per timepoint when analysing using arithmetic means. This is a significant deviation from the assumptions from the model used in data generation. Although this method performs well for data generated using a fixed effects model and analysed using geometric means - that is when the assumptions of the method are valid. However, when we deviate from these assumptions, the method performs poorly.

When applied to datasets generated using fixed effects models, Method 4 generally performs well. However, when applied to datasets generated using mixed effects, this method severely underestimates the variance of the \widehat{AUC} . This is because this method assumes independence between timepoints, hence the covariance of responses across timepoints is assumed to be zero. Since these are in fact positively correlated, treating the responses at timepoints as independent will underestimate the variance of the \widehat{AUC} . This gives this method very poor coverage.

Method 5 performs reasonably well when data is generated using the fixed effects model, and using the original parameter values from Beal,¹⁰ even when

deviating from assumptions on the distribution of concentrations per timepoint. However, for datasets generated using mixed effects models, this method does not perform as well, underestimating the \widehat{AUC} and overestimating the variance - with a similar performance for the adjusted model dose and clearance.

Although on a theoretical level Method 6 should perform well, it shows sporadic results. As well as being significantly more computationally intensive than all of the other methods, it does not always give an output and when it does the values are often questionable. Even with the restrictions imposed on the covariance matrix, the high number of free parameters makes the optimization unstable, and when there are deviations from assumptions, the optimization often fails. This method is therefore unsuitable for any realistic use.

Method 7 consistently performs well across all scenarios. The estimates of \widehat{AUC} and its variance are very close to those which they would be without censoring on BLOQ values. This is because this method does not make assumptions on the distributions, however does use information from the uncensored observations to impute different values onto the BLOQ responses. It is computationally efficient and can be applied to datasets even with high levels of censoring. We therefore recommend this method as the most appropriate non-compartmental analysis method of dealing with BLOQ responses.

5.4 Discussion

In this paper we have evaluated the performance of seven different non-compartmental approaches to dealing with some data below the limit of quantification. The simple imputations of Methods 1 and 2, those currently used in practice, are compared to five alternative methods in a number of different scenarios. Using the models used by Beal¹⁰ and that with lower dose and clearance.

The results have shown that the simple imputation methods perform poorly, especially in scenarios with a large proportion of BLOQ responses. Methods that

use maximum likelihood also fail to estimate the \widehat{AUC} and its variance well. It is clear that the method of kernel density imputation is the best performing out of all the methods considered and is hence is the preferred method for dealing with BLOQ responses in NCA.

The limitations of the method include the issue from which all methods apart from simple imputation of a single value suffer - the scenario where all responses at a given time point are reported as BLOQ. In this case, since the only information on responses at that time point is that they all lie between 0 and LOQ, nothing is known about the distribution of the responses and hence no kernel density estimation can be calculated.

Although in this paper we have only looked at studies that have full sampling schemes, the kernel density imputation method can easily be applied to sparse sampling schemes. The imputation values are calculated in exactly the same way, it is only the process of ordering of these values that will differ.

For serial sampling, where one blood sample is taken per subject, the ordering at any given time point is of no concern and hence can be assigned randomly. In batch designs, the subjects are divided into batches with each batch of subjects all sampled at the same time points with no other subjects sampled at any of these time points (hence the time points for each batch form a partition of the set of all time points). Here the ordering can be based on the previous time point that the specific batch was sampled at, instead of the previous time point. For more flexible designs, the subjects are not separated into disjoint batches by their time points, but for each individual set of sampling times, there must be at least two subjects following this in order for the variance of the non-compartmental estimate of the \widehat{AUC} to be calculated. If all subjects with BLOQ responses are on the same set of sampling times, the ordering will be based on the responses at previous time point from those sampling times. If all subjects are not on the same set of sampling times then one may choose the allocation will be random or

by some other rule, for example one based on gradients of linear interpolations between responses at different time points.

This method is by no means limited to PK studies. It can also be further extended to other scenarios where left-censored values are present but no model fitting takes place. The scope of the application of kernel density imputation is wide, and the potential to extend to further more complicated settings is promising.

CHAPTER 6

Thesis Conclusions, Limitations and Further Work

6.1 Overview

This thesis has covered three separate albeit related research topics in the area of pharmacokinetics. Although mostly motivated by the blood sampling method of microsampling, these topics also stand alone in their methodologies to be applied to numerous different settings. This chapter provides an overview of the conclusions obtained from each of the topics for research, the limitations in the approaches and further work that can stem from these.

6.2 Conclusions

Firstly, the multiple comparison procedure discussed in Chapter 3 is a real step forward in the process of providing evidence for the use of microsampling in GLP studies. The modelling approach is used in this instance and hence the correlation between the PK parameters can be estimated using the Fisher Information Matrix as a result from the model fitting process. Using model averaging over a set of candidate models ensures that the uncertainty in the model selection process is included in the estimate of the PK parameters and their variance, hence eliminating some of the disadvantages of PK modelling compared to NCA. The extension to equivalence testing for longitudinal data makes this procedure an excellent choice for the application of comparing microsampling to traditional sampling in PK studies. The type I error rate and power of the procedure indicate a promising opportunity for opening up the application of microsampling.

Secondly, the algorithm for selecting an optimal sparse sampling scheme for PK studies using NCA introduced in Chapter 4 is not only extremely useful in the context of PK studies using microsampling, but can also be extended to other scenarios where sparse sampling is required, e.g. in studies in infants. This method allows for a very high level of flexibility in the design of the sparse

sampling scheme as there is no specification on the individual schemes that requires an analytic form of the variance of PK parameter estimates. This is as well as the flexibility from allowing for deviations from any assumed model via use of the minimax procedure. The scope for inclusion of multiple PK parameters with user defined importance weights is a necessary aspect to ensure that the chosen designs are suitable for purpose. The two stages of the procedure give the opportunity for minimizing both the bias and the variance of the PK parameter estimates over the conditions that are most appropriate to the criteria. Simulation results have shown that use of this algorithm can considerably reduce the variance of PK parameter estimates, a very promising outcome. This methodology is soon to be available in the R package 'PKdesign', also with an R shiny app so that scientists may use this in practice.

Finally, the kernel density imputation approach to dealing with BLOQ observations proposed in Chapter 5 is a novel method for non-parametric censoring that has possibility for application extending above and beyond PK studies. The obvious advantage of not needing to specify not only a PK model but also an error model places this method on an assumptional par with the simple imputation methods currently used. It does however rise above these methods in the use of uncensored data to calculate the imputed values and of course the superiority of imputing different values onto censored responses for different subjects. The results from simulations reflect these notable improvements over other methods, even when the stricter assumptions of other methods are indeed valid. The kernel density imputation method gives point estimates and variance of the \widehat{AUC} close to that which we would see were there to be no censoring, the main objective of any imputation method. There is no doubt that this method would be extremely useful in situations where both high and low levels of censoring occur and distributional assumptions are to be kept to a minimum.

6.3 Limitations

As with any methodology, the approaches introduced in this thesis have limitations in their application, both in a practical sense and in the assumptions that they require.

Considering the method introduced in Chapter 3 for testing the equivalence of blood sampling methods in PK studies using model averaged derived PK parameters, as with any approach that requires model fitting, there may be setbacks and complexities when fitting the model. This is especially relevant in the case of using the more complex PK models with mixed effects. It is also worth noting that the assumptions that are required for this method e.g. asymptotic normality of the derived PK parameter estimates place limitations on the appropriateness of the approach, potentially restricting the validity of results if applied to settings where these assumptions are violated.

Although the two stage algorithm proposed in Chapter 4 to find optimal sparse sampling schemes for PK studies offers the advantages of using non-compartmental methods, it still requires some prior knowledge of the assumed underlying PK model. This limitation is partly managed by the use of the mini-max criteria to include uncertainty over these assumptions. However, the use of NCA then requires simulations to be used, and hence the computational cost of the algorithm is more than that of a model based approach to optimal designs for PK studies. The use of simulations also comes at the expense that due to simulation error, the optimization is only an approximation, albeit a good approximation due to the large number of simulations.

The novel kernel density imputation approach to dealing with BLOQ observations in non-compartmental analysis of PK studies that is introduced in Chapter 5 has much fewer limitations than other methods available for such an application. One limitation is however that it offers no solution to the possible

scenario where all responses at a single time point are below the limit of quantification, as no kernel density estimation may be calculated to approximate the distribution of observations if there are no numerical responses observed. A further limitation is in the choice of bandwidth. Currently, the bandwidth is recalculated at each iteration of the algorithm where kernel density estimation is applied using Silverman's 'Rule of Thumb' for simplicity. There are of course potentially other more complex approaches to calculating the bandwidth parameter, which may improve the accuracy of the density estimation and hence the performance of the algorithm. A final limitation comes in the way that the values are imputed, by using the ordering of responses at previous timepoints. This is possibly inducing further correlation between responses at time points which in turn may be affecting the variance of the PK parameter estimates.

6.4 Further Work

A possible extension to the work discussed in Chapter 3 is to consider the specification of equivalence margins. Stricter margins that are more harmonious with tests for individual bioequivalence may be used instead of those in line with testing for population equivalence. Further investigation may be needed to fine tune these margins for any practical application.

There are many possibilities in terms of expanding on the methodology introduced in Chapter 4. The main opportunity for this is in the application to a wider variety of settings, for example multiple dosing schemes. Here, the specification of the underlying model would be altered to accommodate this change, and consideration would have to be given to the first stage of the procedure for choosing the timepoints. One may also wish to relax the restrictions on the sampling schemes in order to extend the approach to allow for unbalanced designs. These extensions are mainly straightforward and require no changes to the processes of the methodology itself, only to the application. Another consideration

for extension is in the choice of the optimality criterion itself. One approach could be to use likelihood based criteria, indeed related to the maximum likelihood methods in Chapter 5. As already established, the non-compartmental estimate of the *AUC* and its variance can be written as a function of the mean and variance of responses across timepoints, for which we can construct a likelihood based on certain distributional assumptions, such as a Lognormal distribution of responses at a given timepoint. One may then use this to minimize some function of the Fisher Information Matrix, similar to the approach in model based optimality, however still without the specification of the PK model in the optimality criterion. It would be interesting to see how this approach compares to that of Chapter 4 as it requires more distributional assumptions than this approach, but fewer than using traditional model based optimality.

For the kernel density imputation method for dealing with BLOQ values, only full sampling schemes are evaluated in the results in Chapter 5, however it would be an interesting extension to also apply this to sparse sampling schemes. The algorithm for calculating the imputed can be applied in the same way, however further exploration into the ordering of the allocation of the imputed values to different subjects would be required. In order to address the limitation of this method in terms of correlation between responses at multiple time points, a possible extension would be to use multivariate kernel density estimation to account for this correlation. This would need further attention to detail in the process of imputing values so that the increase in dimensionality does not adversely affect the convergence or indeed the performance of the algorithm.

Bibliography

- ¹ J. E. AHN, M. O. KARLSSON, A. DUNNE, AND T. M. LUDDEN, *Likelihood based approaches to handling data below the quantification limit using NONMEM VI*, Journal of Pharmacokinetics and Pharmacodynamics, 35 (2008), pp. 401–421.
- ² H. AKAIKE, *Information theory and an extension of the maximum likelihood principle*, in Second International Symposium on Information Theory, Budapest: Akademiai Kiado, 1973, pp. 267–281.
- ³ D. G. ALTMAN, *Practical Statistics for Medical Research*, CRC Press, 1991.
- ⁴ D. A. ARMBRUSTER AND T. PRY, *Limit of Blank , Limit of Detection and Limit of Quantitation*, The Clinical Biochemist Reviews, 29 (Sup 1) (2008), pp. 49–52.
- ⁵ A. C. ATKINSON AND A. DONEV, *Optimum Experimental Design*, Clarendon Press, Oxford, 1992.
- ⁶ A. J. BAILER, *Testing for the equality of area under the curves when using destructive measurement techniques.*, Journal of pharmacokinetics and biopharmaceutics, 16 (1988), pp. 303–9.
- ⁷ A. J. BAILER AND W. W. PIEGORSCH, *Estimating Integrals Using Quadrature Methods with an Application in Pharmacokinetics*, Biometrics, 46 (1990), pp. 1201–1211.
- ⁸ R. BANKS, *The fourth R of research*, American Association for Laboratory Animal Science, 34 (1995), pp. 50–51.
- ⁹ C. BAZZOLI, S. RETOUT, AND F. MENTRÉ, *Design evaluation and optimisation in multiple response nonlinear mixed effect models: PFIM 3.0*, Computer Methods and Programs in Biomedicine, 98 (2010), pp. 55–65.
- ¹⁰ S. L. BEAL, *Ways to fit a PK model with some data below the quantification limit*, Journal of Pharmacokinetics and Pharmacodynamics, 28 (2001), pp. 481–504.
- ¹¹ P. L. BONATE, *Pharmacokinetic-Pharmacodynamic Modeling and Simulation*, Springer, New York, 2006.
- ¹² S. T. BUCKLAND, K. P. BURNHAM, AND N. H. AUGUSTIN, *Model Selection : An Integral Part of Inference*, Biometrics, 53 (1997), pp. 603–618.
- ¹³ K. P. BURNHAM AND D. R. ANDERSON, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer-Verlag, 2nd ed., 2002.
- ¹⁴ W. BYON, C. V. FLETCHER, AND R. C. BRUNDAGE, *Impact of censoring data below an arbitrary quantification limit on structural model misspecification*, Journal of Pharmacokinetics and Pharmacodynamics, 35 (2008), pp. 101–116.
- ¹⁵ W. CAWELLO, *Parameters for Compartment-free Pharmacokinetics - Standardisation of Study Design, Data Analysis and Reporting*, Shaker Verlag, Aachen, 2003.

- ¹⁶ K. CHAPMAN, S. CHIVERS, D. GLIDDON, D. MITCHELL, S. ROBINSON, T. SANGSTER, S. SPARROW, N. SPOONER, AND A. WILSON, *Overcoming the barriers to the uptake of nonclinical microsampling in regulatory safety studies.*, *Drug discovery today*, 19 (2014), pp. 528–532.
- ¹⁷ C. CHATFIELD, *Model Uncertainty*, *Data Mining and Statistical Inference*, *Journal of the Royal Statistical Society. Series A*, 158 (1995), pp. 419–466.
- ¹⁸ S.-C. CHOW AND J.-P. LIU, *Design and Analysis of Bioavailability and Bioequivalence Studies*, CRC Press, 3rd ed., 2008.
- ¹⁹ CLINICAL AND LABORATORY STANDARDS INSTITUTE, *Protocols for Determination of Limits of Detection and Limits of Quantitation ; Approved Guideline*. CLSI Document EP17-A, vol. 24, 2004.
- ²⁰ M. DAVIDIAN AND D. M. GILTINAN, *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall/CRC, 1995.
- ²¹ L. DILLEN, T. LOOMANS, G. VAN DER PERRE, D. VERSWETVELD, K. WUYTS, AND L. DE ZWART, *Blood microsampling using capillaries for drug exposure determination in early preclinical studies: a beneficial strategy to reduce blood sample volumes*, *Bioanalysis*, 6 (2014), pp. 1–14.
- ²² M. L. EATON, *Multivariate Statistics: A Vector Space Approach*, John Wiley & Sons, Inc., 1983.
- ²³ B. EFRON, *Bootstrap Methods: Another Look at the Jackknife*, *The Annals of Statistics*, 7 (1979), pp. 1–26.
- ²⁴ G. EMMONS AND M. ROWLAND, *Pharmacokinetic considerations as to when to use dried blood spot sampling.*, *Bioanalysis*, 2 (2010), pp. 1791–1796.
- ²⁵ M. EVANS AND T. SWARTZ, *Approximating Intervals via Monte Carlo and Deterministic Methods*, Oxford University Press, Oxford, 2000.
- ²⁶ V. FEDOROV AND S. LEONOV, *Population pharmacokinetic measures, their estimation and selection of sampling times.*, *Journal of biopharmaceutical statistics*, 17 (2007), pp. 919–41.
- ²⁷ V. V. FEDOROV AND P. HACKL, *Model-Oriented Design of Experiments*, Springer, New York, 1997.
- ²⁸ V. V. FEDOROV AND S. L. LEONOV, *Optimal Design for Nonlinear Response Models*, Chapman & Hall/CRC, 2014.
- ²⁹ J. GABRIELSSON AND D. WEINER, *Pharmacokinetic and pharmacodynamic data analysis: concepts and applications*, vol. 1, CRC Press, 2001.
- ³⁰ R. GAGNON AND S. LEONOV, *Optimal population designs for PK models with serial sampling*, *Journal of biopharmaceutical statistics*, 15 (2005), pp. 143–163.

- ³¹ R. C. GAGNON AND J. J. PETERSON, *Estimation of confidence intervals for area under the curve from destructively obtained pharmacokinetic data.*, *Journal of pharmacokinetics and biopharmaceutics*, 26 (1998), pp. 87–102.
- ³² M. GIBALDI AND D. PERRIER, *Pharmacokinetics*, Marcel Dekker, New York, 1982.
- ³³ F. A. GRAYBILL, *Matrices with Applications in Statistics*, Wadsworth International Group, 2nd ed., 1983.
- ³⁴ D. R. HELSEL, *Statistics for Censored Environmental Data Using Minitab and R*, Wiley, 2nd ed., 2012.
- ³⁵ J. P. HING, S. G. WOOLFREY, D. GREENSLADE, AND P. M. C. WRIGHT, *Analysis of toxicokinetic data using NONMEM: Impact of quantification limit and replacement strategies for censored data*, *Journal of Pharmacokinetics and Pharmacodynamics*, 28 (2001), pp. 465–479.
- ³⁶ Y. HOCHBERG, *A sharper Bonferroni procedure for multiple tests of significance*, *Biometrika*, 75 (1988), pp. 800–802.
- ³⁷ D. J. HOLDER, F. HSUAN, R. DIXIT, AND K. SOPER, *a Method for Estimating and Testing Area Under the Curve in Serial Sacrifice, Batch, and Complete Data Designs*, *Journal of Biopharmaceutical Statistics*, 9 (1999), pp. 451–464.
- ³⁸ S. HOLM, *A Simple Sequentially Rejective Multiple Test Procedure*, *Scandinavian Journal of Statistics*, 6 (1979), pp. 65–70.
- ³⁹ G. HOMMEL, *Multiple test procedures for arbitrary dependence structures*, *Metrika*, 33 (1986), pp. 321–336.
- ⁴⁰ —, *A stagewise rejective multiple test procedure based on a modified Bonferroni test*, *Biometrika*, 75 (1988), pp. 383–386.
- ⁴¹ T. JAKI AND M. J. WOLFSEGGGER, *A Theoretical Framework for Estimation of AUCs in Complete and Incomplete Sampling Designs*, *Statistics in Biopharmaceutical Research*, 1 (2009), pp. 176–184.
- ⁴² —, *Estimation of pharmacokinetic parameters with the R package PK*, *Pharmaceutical Statistics*, 10 (2011), pp. 284–288.
- ⁴³ —, *Non-compartmental estimation of pharmacokinetic parameters for flexible sampling designs*, *Statistics in Medicine*, 31 (2012), pp. 1059–1073.
- ⁴⁴ T. JAKI, M. J. WOLFSEGGGER, AND M. PLONER, *Confidence intervals for ratios of AUCs in the case of serial sampling: a comparison of seven methods*, *Pharmaceutical Statistics*, 8 (2009), pp. 12–24.
- ⁴⁵ S. S. JAMBHEKAR AND P. J. BREEN, *Basic Pharmacokinetics*, Pharmaceutical Press, London, 2009.
- ⁴⁶ S. M. JENSEN, C. B. PIPPER, AND C. RITZ, *Evaluation of multi-outcome longitudinal studies*, *Statistics in Medicine*, 34 (2015), pp. 1993–2003.

- ⁴⁷ S. M. JENSEN AND C. RITZ, *Simultaneous Inference for Model Averaging of Derived Parameters*, *Risk Analysis*, 35 (2015), pp. 68–76.
- ⁴⁸ O. JONSSON, R. P. VILLAR, L. B. NILSSON, C. NORSTEN-HÖÖG, J. BROGREN, M. ERIKSSON, K. KÖNIGSSON, AND A. SAMUELSSON, *Capillary microsampling of 25 micro-litre blood for the determination of toxicokinetic parameters in regulatory studies in animals*, *Bioanalysis*, 4 (2012), pp. 661–674.
- ⁴⁹ A. KÄLLÉN, *Computational Pharmacokinetics*, Chapman & Hall/CRC, 2008.
- ⁵⁰ A. R. KROMMER AND C. W. UEBERHUBER, *Computational Integration*, SIAM, Philadelphia, 1998.
- ⁵¹ D. J. LUNN AND L. J. AARONS, *Markov Chain Monte Carlo Techniques for Studying Interoccasion and Intersubject Variability : Application to Pharmacokinetic Data*, *Journal of the Royal Statistical Society. Series C*, 46 (1997), pp. 73–91.
- ⁵² R. MARCUS, E. PERITZ, AND K. GABRIEL, *On closed testing procedures with special reference to ordered analysis of variance*, *Biometrika*, 63 (1976), pp. 655–660.
- ⁵³ NC₃Rs, *Rat: Decision Tree for Blood Sampling*. <https://www.nc3rs.org.uk/rat-decision-tree-blood-sampling>. (accessed August 8, 2017).
- ⁵⁴ J. R. NEDELMAN AND E. GIBIANSKY, *The variance of a better AUC estimator for sparse, destructive sampling toxicokinetics*, *Journal of Pharmaceutical Sciences*, 85 (1996), pp. 884–886.
- ⁵⁵ J. R. NEDELMAN, E. GIBIANSKY, AND D. T. W. LAU, *Applying Bailer’s method for AUC confidence intervals to sparse sampling*, *Pharmaceutical research*, 12 (1995), pp. 124–128.
- ⁵⁶ L. B. NILSSON AND O. JONSSON, *Capillary microsampling in the regulatory environment : validation and use of bioanalytical capillary microsampling methods*, *Bioanalysis*, 5 (2013), pp. 731–738.
- ⁵⁷ C. B. PIPPER, C. RITZ, AND H. BISGAARD, *A versatile method for confirmatory evaluation of the effects of a covariate in multiple models*, *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 61 (2012), pp. 315–326.
- ⁵⁸ N. POWLES-GLOVER, S. KIRK, C. WILKINSON, S. ROBINSON, AND J. STEWART, *Assessment of toxicological effects of blood microsampling in the vehicle dosed adult rat*, *Regulatory Toxicology and Pharmacology*, 68 (2014), pp. 325–331.
- ⁵⁹ F. PUKELSHEIM, *Optimal Design of Experiments*, Wiley, New York, 1993.
- ⁶⁰ H. QUAN, J. BOLOGNESE, AND W. YUAN, *Assessment of equivalence on multiple endpoints*, *Statistics in Medicine*, 20 (2001), pp. 3159–3173.
- ⁶¹ B. REISFIELD AND A. N. MAYENO, eds., *Computational Toxicology*, vol. 1, Springer, 2013.

- ⁶² S. RETOUT, S. DUFFULL, AND F. MENTRÉ, *Development and implementation of the population Fisher information matrix for the evaluation of population pharmacokinetic designs*, *Computer Methods and Programs in Biomedicine*, 65 (2001), pp. 141–151.
- ⁶³ S. RETOUT, F. MENTRÉ, AND R. BRUNO, *Fisher information matrix for non-linear mixed-effects models: Evaluation and application for optimal design of enoxaparin population pharmacokinetics*, *Statistics in Medicine*, 21 (2002), pp. 2623–2639.
- ⁶⁴ C. RITZ, R. PILMANN LAURSEN, AND C. TRAB DAMSGAARD, *Simultaneous inference for multilevel linear mixed models—with an application to a large-scale school meal study*, *Journal of the Royal Statistical Society. Series C: Applied Statistics*, (2016), pp. 295–311.
- ⁶⁵ W. RUSSELL AND R. BURCH, *The Principles of Humane Experimental Technique*, Methuen, London, 1959.
- ⁶⁶ D. J. SCHUIRMANN, *A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability.*, *Journal of pharmacokinetics and biopharmaceutics*, 15 (1987), pp. 657–680.
- ⁶⁷ B. W. SILVERMAN, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, 1986.
- ⁶⁸ R. J. SIMES, *An improved Bonferroni procedure for multiple tests of significance*, *Biometrika*, 73 (1986), pp. 751–754.
- ⁶⁹ M. TOD, F. MENTRÉ, Y. MERLÉ, AND A. MALLET, *Robust optimal design for the estimation of hyperparameters in population pharmacokinetics*, *Journal of Pharmacokinetics and Biopharmaceutics*, 26 (1998), pp. 689–716.
- ⁷⁰ A. VAN DER VAART, *Asymptotic Statistics*, Cambridge University Press, 2000.
- ⁷¹ M. J. WOLFSEGGER AND T. JAKI, *Estimation of AUC from 0 to infinity in serial sacrifice designs*, *Journal of Pharmacokinetics and Pharmacodynamics*, 32 (2005), pp. 757–766.
- ⁷² M. J. WOLFSEGGER AND T. JAKI, *Non-compartmental estimation of pharmacokinetic parameters in serial sampling designs*, *Journal of Pharmacokinetics and Pharmacodynamics*, 36 (2009), pp. 479–494.
- ⁷³ S. P. WRIGHT, *Adjusted P-Values for Simultaneous Inference*, *Biometrics*, 48 (1992), pp. 1005–1013.
- ⁷⁴ Y. WU, W. PIEGORSCH, R. WEST, D. TANG, M. PETKEWICH, AND W. PAN, *Multiplicity-adjusted inferences in risk assessment: Benchmark analysis with continuous response data.*, *Environmental and Ecological Statistics*, 13 (2006), pp. 125–141.
- ⁷⁵ C. YEH, *Estimation and significant tests of area under the curve derived from incomplete blood sampling*, in *Statistical Association Proceedings of the Biopharmaceutical Section*, 1990, pp. 74–81.

SUPPLEMENTARY MATERIAL

APPENDIX A

PK Parameters as Functions of Model Parameters

A.1 For Candidate Model 3.1

$$t_{\frac{1}{2}} = -\frac{\log(2)}{B_1}$$

$$C_{max} = B_0$$

A.2 For Candidate Model 3.2

$$t_{\frac{1}{2}} = \frac{1}{B_1} \log\left(\frac{B_0 - \log(2)}{B_0}\right)$$

$$C_{max} = \exp(B_0)$$

A.3 For Candidate Models 3.6 and 3.7

$$AUC_{24} = \frac{k_a F D}{V(k_a - k_e)} \left(\left(\frac{\exp(-24k_a) - 1}{k_a} \right) - \left(\frac{\exp(-24k_e) - 1}{k_e} \right) \right)$$

$$t_{max} = \frac{\log(k_e) - \log(k_a)}{k_e - k_a}$$

$$C_{max} = \frac{k_a F D}{V(k_a - k_e)} \left(\left(\frac{k_a}{k_e} \right)^{\left(\frac{k_e}{k_e - k_a} \right)} - \left(\frac{k_a}{k_e} \right)^{\left(\frac{k_a}{k_e - k_a} \right)} \right)$$

APPENDIX B

Sampling Time Points for Simulations

Table B.1: Sampling Time Points Used in Simulation Studies

Number of Time Points	Time Points
3	(1, 10, 24)
4	(1, 8, 18, 36)
5	(1, 7, 14, 21, 36)
6	(1, 6, 12, 18, 24, 36)
7	(1, 4, 8, 12, 18, 24, 36)
8	(1, 3, 6, 8, 12, 18, 24, 36)
9	(1, 2, 4, 6, 8, 12, 18, 24, 36)
10	(1, 2, 3, 4, 6, 8, 12, 18, 24, 36)

APPENDIX c

Derivation of Second Order Approximation

Using the notation defined in Section 3.2.2, a first order Taylor approximation of f is given by:

$$f(\hat{\tau}) \approx f(\mu) + \mathbf{D}^T(\hat{\tau} - \mu)$$

Hence we may approximate the variance of f by:

$$\begin{aligned} \text{Var}[f(\hat{\tau})] &\approx \text{Var}[f(\mu) + \mathbf{D}^T(\hat{\tau} - \mu)] \\ &= \text{Var}[\mathbf{D}^T(\hat{\tau} - \mu)] \\ &= \mathbf{D}^T \text{Var}[(\hat{\tau} - \mu)] \mathbf{D} \\ &= \mathbf{D}^T \hat{\Sigma}(\hat{\tau}) \mathbf{D} \end{aligned}$$

A second order Taylor approximation of f is given by:

$$f(\hat{\tau}) \approx f(\mu) + \mathbf{D}^T(\hat{\tau} - \mu) + \frac{1}{2}(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu)$$

Since taking the variance of this is not as straightforward as in the first order case, we use that $\text{Var}[f(\hat{\tau})] = E[f^2(\hat{\tau})] - E^2[f(\hat{\tau})]$ so need to calculate $f^2(\hat{\tau})$ and $E[f(\hat{\tau})]$. Firstly, $f^2(\hat{\tau})$:

$$\begin{aligned} f^2(\hat{\tau}) &\approx \left(f(\mu) + \mathbf{D}^T(\hat{\tau} - \mu) + \frac{1}{2}(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu) \right)^2 \\ &= f^2(\mu) + \mathbf{D}^T(\hat{\tau} - \mu)(\hat{\tau} - \mu)^T \mathbf{D} + \frac{1}{4}(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu)(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu) \\ &\quad + 2f(\mu)\mathbf{D}^T(\hat{\tau} - \mu) + f(\mu)(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu) + \mathbf{D}^T(\hat{\tau} - \mu)(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu). \end{aligned}$$

Hence

$$E^2[f(\hat{\tau})] \approx E \left[\begin{array}{l} f^2(\mu) + \mathbf{D}^T(\hat{\tau} - \mu)(\hat{\tau} - \mu)^T \mathbf{D} + \frac{1}{4}(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu)(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu) \\ + 2f(\mu)\mathbf{D}^T(\hat{\tau} - \mu) + f(\mu)(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu) + \mathbf{D}^T(\hat{\tau} - \mu)(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu). \end{array} \right] \quad (\text{C.1})$$

$$= f^2(\mu) + \mathbf{D}^T \hat{\Sigma}(\hat{\tau}) \mathbf{D} + \frac{1}{4} E[(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu)(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu)] + f(\mu) \text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\} \quad (\text{C.2})$$

$$= f^2(\mu) + \mathbf{D}^T \hat{\Sigma}(\hat{\tau}) \mathbf{D} + \frac{1}{4} \left(\text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\} \right)^2 + \frac{1}{2} \text{tr}\{(\mathbf{H} \hat{\Sigma}(\hat{\tau}))^2\} + f(\mu) \text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\}. \quad (\text{C.3})$$

Where (C.2) follows from (C.1) since the expectations of first and third order moments of normal random variables with mean $\mathbf{0}$ is $\mathbf{0}$ (the fourth and last terms disappear). By Theorem 10.9.10 from Graybill,³³ which states:

Theorem. Let \mathbf{x} be an $n \times 1$ vector with distribution $N(\mathbf{x} : \mathbf{0}, \mathbf{V})$; let \mathbf{A} , \mathbf{B} and \mathbf{C} be symmetric matrices of constants. Then

$$(1) E[(\mathbf{x}^T \mathbf{A} \mathbf{x})(\mathbf{x}^T \mathbf{B} \mathbf{x})] = [\text{tr}(\mathbf{A} \mathbf{V})][\text{tr}(\mathbf{B} \mathbf{V})] + 2\text{tr}(\mathbf{A} \mathbf{V} \mathbf{B} \mathbf{V}),$$

$$(2) \text{Cov}[\mathbf{x}^T \mathbf{A} \mathbf{x}, \mathbf{x}^T \mathbf{B} \mathbf{x}] = 2\text{tr}(\mathbf{A} \mathbf{V} \mathbf{B} \mathbf{V}),$$

$$(3) \text{Var}[\mathbf{x}^T \mathbf{A} \mathbf{x}] = 2\text{tr}(\mathbf{A} \mathbf{V})^2.$$

along with the fact that for these conditions, $E[\mathbf{x}^T \mathbf{A} \mathbf{x}] = \text{tr}(\mathbf{A} \mathbf{V})$, the fifth term from (C.1) to (C.2) follows. The above theorem provides reasoning for (C.2) to (C.3). Now we calculate $E[f(\hat{\tau})]$:

$$\begin{aligned} E[f(\hat{\tau})] &\approx E[f(\mu) + \mathbf{D}^T(\hat{\tau} - \mu) + \frac{1}{2}(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu)] \\ &= f(\mu) + \mathbf{D}^T E[(\hat{\tau} - \mu)] + \frac{1}{2} E[(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu)] \\ &= f(\mu) + \frac{1}{2} \text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\}. \end{aligned}$$

Now combining the above:

$$\begin{aligned} \text{Var}[f(\hat{\tau})] &= E[f^2(\hat{\tau})] - E^2[f(\hat{\tau})] \\ &\approx f^2(\mu) + \mathbf{D}^T \hat{\Sigma}(\hat{\tau}) \mathbf{D} + \frac{1}{4} \left(\text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\} \right)^2 + \frac{1}{2} \text{tr}\{(\mathbf{H} \hat{\Sigma}(\hat{\tau}))^2\} \\ &\quad + f(\mu) \text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\} - (f(\mu) + \frac{1}{2} \text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\})^2 \\ &= f^2(\mu) + \mathbf{D}^T \hat{\Sigma}(\hat{\tau}) \mathbf{D} + \frac{1}{4} \left(\text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\} \right)^2 + \frac{1}{2} \text{tr}\{(\mathbf{H} \hat{\Sigma}(\hat{\tau}))^2\} \\ &\quad + f(\mu) \text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\} - f^2(\mu) + f^2(\mu) \text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\} - \frac{1}{4} \left(\text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\} \right)^2 \\ &= \mathbf{D}^T \hat{\Sigma}(\hat{\tau}) \mathbf{D} + \frac{1}{2} \text{tr}\{(\mathbf{H} \hat{\Sigma}(\hat{\tau}))^2\}. \end{aligned}$$

The first part is recognizable as the estimate of the variance for the first order Taylor approximation, and the second part is therefore the second order part of the approximation.

APPENDIX D

Average Bias of Parameter Estimates

Table D.1: The average bias of the estimate of the PK parameters. The true values are $t_{\frac{1}{2}} = 42.80264$ and $C_{max} = 110.9412$

		# Subjects							
		5		10		100		1000	
# Time Points		$t_{\frac{1}{2}}$	C_{max}	$t_{\frac{1}{2}}$	C_{max}	$t_{\frac{1}{2}}$	C_{max}	$t_{\frac{1}{2}}$	C_{max}
3		0.913	-0.317	0.132	-0.203	-0.591	-0.149	-0.573	-0.0937
4		0.348	-0.211	-0.0657	-0.218	-0.291	-0.192	-0.141	-0.0987
5		0.481	-0.338	0.0809	-0.326	-0.268	-0.233	-0.136	-0.0933
6		0.398	-0.341	-0.0164	-0.279	-0.296	-0.239	-0.150	-0.0856
7		0.157	-0.270	-0.0918	-0.290	-0.327	-0.216	-0.138	-0.0847
8		0.128	-0.272	-0.121	-0.291	-0.333	-0.202	-0.141	-0.0789
9		0.0625	-0.235	-0.143	-0.261	-0.354	-0.182	-0.138	-0.0671
10		0.0128	-0.212	-0.263	-0.196	-0.374	-0.168	-0.149	-0.058

APPENDIX E

Additional Type I Error Rate Results

Table E.1: Type I error rate for varying numbers of time points and combinations of PK parameters for an oral administration of a compound. Error bounds for 10,000 simulations are 2.214 and 2.806 for equivalence testing.

Parameter Combination / # Subjects								
	AUC_{24} & C_{max}				AUC_{24} & t_{max}			
# Time Points	5	10	100	1000	5	10	100	1000
4	3.01	2.55	2.28	2.48	2.73	2.75	2.57	2.66
5	2.71	2.59	2.61	2.68	2.76	2.63	2.63	2.65
6	2.93	2.49	2.36	2.51	2.96	2.47	2.47	2.28
7	2.47	2.50	2.50	2.30	2.65	2.63	2.63	2.31
8	2.53	2.33	2.80	2.67	2.87	2.60	2.60	2.54
9	2.88	2.59	2.51	2.64	2.41	2.39	2.39	2.45
10	2.98	2.58	2.36	2.51	2.50	2.65	2.65	2.27

Table E.2: Type I error rate for varying numbers of time points and combinations of PK parameters for an oral administration of a compound. Error bounds for 10,000 simulations are 2.214 and 2.806 for equivalence testing.

Parameter Combination / # Subjects								
	t_{max} & C_{max}				AUC_{24} , C_{max} & t_{max}			
# Time Points	5	10	100	1000	5	10	100	1000
4	2.53	2.51	2.59	2.15	2.78	2.66	2.32	2.60
5	2.81	2.38	2.53	2.65	2.48	2.67	2.43	2.55
6	2.85	2.59	2.55	2.35	2.67	2.58	2.58	2.68
7	2.60	2.50	2.69	2.84	2.44	2.39	2.60	2.25
8	2.57	2.57	2.23	2.53	2.58	2.51	2.48	2.92
9	2.37	2.40	2.26	2.54	2.34	2.54	2.65	2.47
10	2.73	2.30	2.80	2.60	2.40	2.40	2.24	2.16

APPENDIX F

Minimax Scenarios

The 8 scenarios with combinations of high (H) and low (L) variance for $(\text{var}(V), \text{var}(k_e), \text{var}(k_a))$ are as follows:

- (1) *LLL*;
- (2) *LLH*;
- (3) *LHL*;
- (4) *HLL*;
- (5) *LHH*;
- (6) *HHL*;
- (7) *HLH*;
- (8) *HHH*.

APPENDIX G

Results Tables I

Table G.1: Top 5 overall schemes using time points (0.5, 1.0, 2.0, 4.0, 9.0, 12.0) according to minimax criteria applied to equally weighted scaled sum of AUC and C_{max} variance: Ranks in the 8 scenarios, maximum rank, and total sum rank. (* indicates the maximum rank for that scheme)

		Top 5 Schemes				
		1001	1003	1026	993	1039
8 Scenarios	1 (LLL)	5	1	2	11	4
	2 (LLH)	7*	1	2	17*	3
	3 (LHL)	3	1	2	7	5
	4 (HLL)	3	9*	11	4	15
	5 (LHH)	5	1	2	12	3
	6 (HHL)	2	8	16*	3	17*
	7 (HLH)	3	5	8	1	7
	8 (HHH)	2	5	9	1	8
	Max	7	9	16	17	17
Total	30	31	52	56	62	

Table G.2: Top 5 overall schemes using time points (0.5, 1.0, 2.0, 4.0, 9.0, 12.0) according to minimax criteria applied to equally weighted scaled sum of AUC and C_{max} variance: Efficiency measure in the 8 scenarios, maximum efficiency measure, and total sum efficiency measure. (* indicates the maximum efficiency measure for that scheme)

		Top 5 Schemes				
		1003	1001	1026	1039	746
8 Scenarios	1 (LLL)	1.000	1.008	1.005	1.007	1.008
	2 (LLH)	1.000	1.009*	1.006	1.007	1.008
	3 (LHL)	1.00	1.006	1.005	1.007	1.009
	4 (HLL)	1.007	1.001	1.008	1.009	1.012
	5 (LHH)	1.000	1.007	1.006	1.007	1.008
	6 (HHL)	1.007*	1.001	1.010*	1.010*	1.012*
	7 (HLH)	1.003	1.001	1.005	1.005	1.005
	8 (HHH)	1.003	1.001	1.007	1.006	1.006
	Max	1.007	1.009	1.010	1.010	1.012
Total	8.021	8.033	8.051	8.058	8.068	

Table G.3: Optimal Time Point Choices Top 5 overall time point choices according to minimax criteria: Ranks in the 8 scenarios, maximum rank, and total sum rank. (* indicates the maximum rank for that time point choice)

		Top 5 Time Choices				
		225	260	235	224	234
8 Scenarios	1 (LLL)	21	24	13	16	18
	2 (LLH)	13	12	22	20	27
	3 (LHL)	39	35	22	28	21
	4 (HLL)	32	36	45	42	52
	5 (LHH)	40*	30	29	36	41
	6 (HHL)	26	31	40	38	42
	7 (HLH)	34	40*	50*	51*	55*
	8 (HHH)	27	33	40	41	44
	Max	40	40	50	51	55
	Total	232	241	261	272	300

Table G.4: Top 5 overall schemes using optimal time points (0.5, 1.0, 3.5, 4.0, 7.5, 12.0) according to minimax criteria applied to equally weighted scaled sum of AUC and C_{max} variance: Ranks in the 8 scenarios, maximum rank, and total sum rank. (* indicates the maximum rank for that scheme)

		Top 5 Schemes				
		811	810	1006	788	1001
8 Scenarios	1 (LLL)	7	12	29	67	81
	2 (LLH)	5	14	34	70	66
	3 (LHL)	7	14	34	80	83*
	4 (HLL)	4	10	20	44	7
	5 (LHH)	8*	16*	38*	81*	65
	6 (HHL)	4	11	35	51	7
	7 (HLH)	2	3	11	19	4
	8 (HHH)	2	4	20	29	3
	Max	8	16	38	81	83
	Total	39	84	221	441	316

APPENDIX H

Results Tables II

Table H.1: Results showing average deviation from the non-compartmental \widehat{AUC} and its variance with data generated from the fixed effects model with higher dose and clearance. Results over 1000 simulations. (10 subjects, 6 timepoints) (A) = Analysed using arithmetic means. (G) = Analysed using geometric means. (N) = Confidence interval calculated using normal distribution on the \widehat{AUC} . (L) = Confidence interval calculated using log-normal distribution on the \widehat{AUC} . M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation. * indicates not all analyses were successful.

	Mean (A)	Var (A)	Cvg 95% (A, N)	Cvg 95% (A, L)	Mean (G)	Var (G)	Cvg 95% (G, N)	Cvg 95% (G, L)
(Truth)	1.0796	0.0274	96.1	96.5	1.0300	0.0242	96.4	95.9
M1	-0.0223	0.0020	92.9	94.4	-0.1373	NA	NA	NA
M2	-0.0082	0.0006	96.2	96.2	-0.0139	0.0011	96.3	96.2
M3	-0.0120*	0.0018*	95.2*	95.8*	-0.0010*	0.0003*	95.8*	95.8*
M4	-0.0063	0.0000	96.8	97.0	-0.0008	-0.0006	97.8	97.6
M5	0.0000	0.0002	96.4	96.7	0.0002	0.0002	96.2	95.9
M6	NA	NA	NA	NA	-0.0004*	-0.0012*	96.1*	96.0*
M7	-0.0023	0.0002	96.4	96.4	-0.0038	0.0004	96.4	96.0

Table H.2: Results showing average deviation from the non-compartmental \widehat{AUC} and its variance with data generated from the mixed effects model with higher dose and clearance. Results over 1000 simulations. (10 subjects, 6 timepoints) (A) = Analysed using arithmetic means. (G) = Analysed using geometric means. (N) = Confidence interval calculated using normal distribution on the \widehat{AUC} . (L) = Confidence interval calculated using log-normal distribution on the \widehat{AUC} . M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation. * indicates not all analyses were successful.

	Mean (A)	Var (A)	Cvg 95% (A, N)	Cvg 95% (A, L)	Mean (G)	Var (G)	Cvg 95% (G, N)	Cvg 95% (G, L)
(Truth)	1.1115	0.1178	94.9	95.2	1.0229	0.0974	95.1	95.0
M1	-0.0281	0.0138	93.5	94.80	-0.2089	NA	NA	NA
M2	-0.0082	0.0032	94.8	95.3	-0.0155	0.0033	94.9	94.7
M3	-0.0327*	0.0249*	94.8*	95.4*	0.0021*	0.0033*	95.2*	95.0*
M4	-0.0194	-0.0667	83.0	83.8	-0.0005	-0.0585	85.8	85.2
M5	-0.0163	0.0108	94.5	95.1	0.0016	-0.0018	95.1	94.9
M6	NA	NA	NA	NA	0.0183*	4.3121*	89.4*	89.2*
M7	-0.0038	0.0012	94.9	95.2	-0.0065	0.0007	94.8	95.1

Table H.3: Results showing average deviation from the non-compartmental \widehat{AUC} and its variance with data generated from the fixed effects model with lower dose and clearance. Results over 1000 simulations. (10 subjects, 6 timepoints) (A) = Analysed using arithmetic means. (G) = Analysed using geometric means. (N) = Confidence interval calculated using normal distribution on the \widehat{AUC} . (L) = Confidence interval calculated using log-normal distribution on the \widehat{AUC} . M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation. * indicates not all analyses were successful.

	Mean (A)	Var (A)	Cvg 95% (A, N)	Cvg 95% (A, L)	Mean (G)	Var (G)	Cvg 95% (G, N)	Cvg 95% (G, L)
(Truth)	0.5065	0.0056	96.6	96.8	0.4816	0.0048	96.4	96.1
M1	-0.0576	0.0052	63.8	70.1	-0.3302	NA	NA	NA
M2	-0.0221	0.0016	88.2	90.5	-0.0367	0.0023	79.2	83.6
M3	-0.0237*	0.0033*	90.7*	92.9*	-0.0029*	0.0005*	95.2	95.2*
M4	-0.0128	-0.0007	95.0	95.6	-0.0019	-0.0003	96.8	96.6
M5	-0.0100	0.0015	96.6	97.2	0.0004	0.0003	95.8	95.6
M6	NA	NA	NA	NA	0.0115*	0.0018*	92.4*	91.5*
M7	-0.0057	0.0004	96.1	96.7	-0.0084	0.0006	95.6	96.1

Table H.4: Results showing average deviation from the non-compartmental \widehat{AUC} and its variance with data generated from the mixed effects model with lower dose and clearance. Results over 1000 simulations. (10 subjects, 6 timepoints) ((A) = Analysed using arithmetic means. (G) = Analysed using geometric means. (N) = Confidence interval calculated using normal distribution on the \widehat{AUC} . (L) = Confidence interval calculated using log-normal distribution on the \widehat{AUC} . M1: Replace BLOQ values with 0, M2: Replace BLOQ values with LOQ/2, M3: ROS Imputation, M4: ML per timepoint Means, M5: ML per timepoint Imputation, M7: Kernel Density Imputation. * indicates not all analyses were successful.

	Mean (A)	Var (A)	Cvg 95% (A, N)	Cvg 95% (A, L)	Mean (G)	Var (G)	Cvg 95% (G, N)	Cvg 95% (G, L)
(Truth)	0.5173	0.0182	94.8	95.7	0.4814	0.0152	96.2	96.1
M1	-0.0640	0.0141	81.9	87.7	0.3735	NA	NA	NA
M2	-0.0229	0.0040	91.4	93.1	-0.0393	0.0047	88.4	91.6
M3	-0.0397*	0.0128*	91.2*	95.1*	-0.0034*	0.0005*	95.3*	95.6*
M4	-0.0217	-0.0082	82.0	84.2	-0.0026	-0.0089	85.0	85.6
M5	-0.0176	0.0053	94.2	95.8	0.0005	-0.0001	96.5	95.9
M6	NA	NA	NA	NA	0.0317*	2.2850*	79.9*	78.2*
M7	-0.0073	0.0011	94.1	95.0	0.0113	0.0010	95.3	95.6