

# Comparing sampling methods for pharmacokinetic studies using model averaged derived parameters

H. Barnett<sup>[1]</sup>, H. Geys<sup>[2]</sup>, T. Jacobs<sup>[2]</sup> and T. Jaki<sup>[1]</sup>

July 2017

<sup>[1]</sup> Department of Mathematics and Statistics, Lancaster University,  
Lancaster, United Kingdom

<sup>[2]</sup> Janssen Pharmaceutica, Beerse, Belgium

## Abstract

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.

## 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.<sup>1</sup> 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<sup>2,3</sup> or using compartmental methods such as fitting non-linear mixed effects models.<sup>4</sup> 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 & Ritz<sup>5</sup> 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.

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 1). Blood sampling and analysis was undertaken by the same analyst and all animals entered the study on the same day.

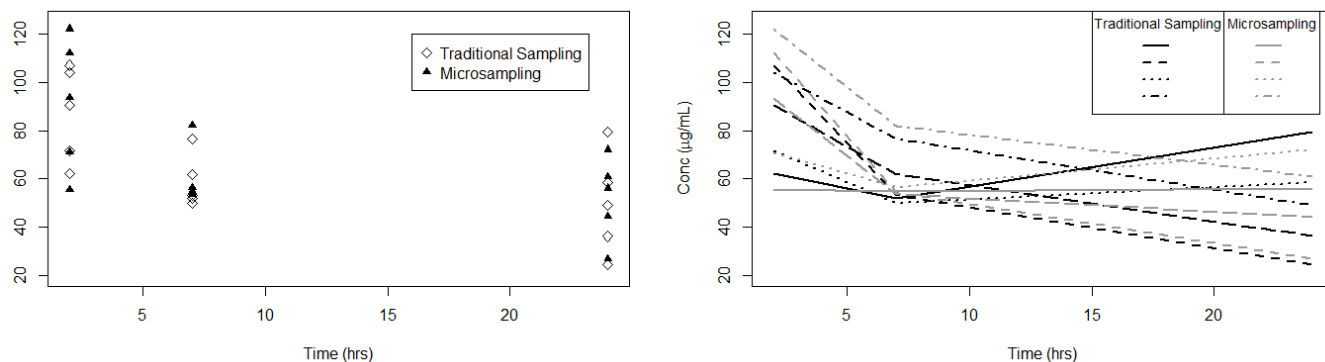


Figure 1: Example dataset with individual concentrations (left) and spaghetti plot (right).

## 2 Superiority Testing

### 2.1 Baseline Method

Jensen & Ritz<sup>5</sup> 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 averaging<sup>6</sup> in order to account for uncertainty in the choice of model and simultaneous confidence intervals are calculated.<sup>7</sup> 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 used<sup>5</sup> :

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

where the  $w_k$ 's are the model specific weights such that  $\sum_{k=1}^K w_k = 1$ . Buckland et al.<sup>6</sup> 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,<sup>8</sup> 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  $\kappa_l$  is in each corresponding confidence interval with simultaneous error rate  $\alpha$ , we fail to reject the null hypothesis at simultaneous level  $\alpha$ . If at least one  $\kappa_l$  is not in its corresponding interval, then the null hypothesis can be rejected.

In order to calculate simultaneous confidence intervals for the multiple derived parameters with simultaneous coverage  $1 - \alpha$ , the methods of Pipper et al.,<sup>7</sup> 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 as<sup>5</sup> :

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

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  $\tau_k$ .

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 & Ritz<sup>5</sup> 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.

### 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 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,  $\tau_S$  and  $\tau_M$  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 1:** Log Linear

$$g(\tau, t) = \beta'_0 \exp(\beta_1 t) + \epsilon. \tag{1}$$

**Candidate Model 2:** Log Log Linear

$$g(\tau, t) = \exp(\beta'_0 \exp(\beta_1 t)) + \epsilon. \tag{2}$$

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 & Ritz<sup>5</sup> 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 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.

### 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 AICc<sup>8</sup> which corresponds to model 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 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 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. In terms of bias of the model averaged derived parameters, 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 2 in Appendix D for the full range of values. All things considered, this amount of bias is not too concerning for the procedure.

## 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:  $\tau$  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

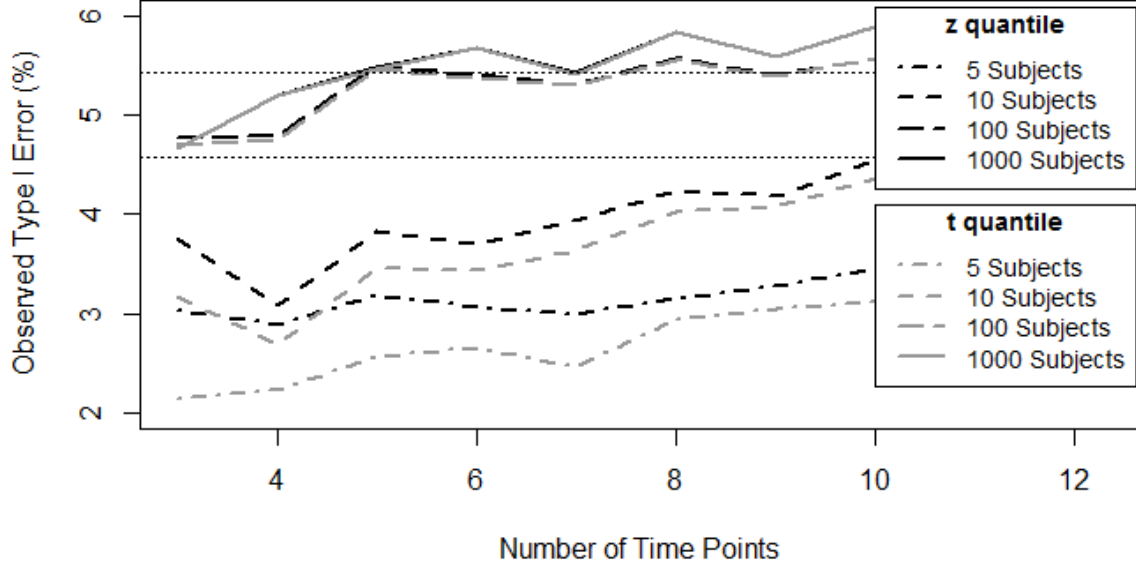


Figure 2: Comparison of observed type  $I$  error rate for varying number of time points and subjects for use of  $z$  and  $t$  quantile. Horizontal dotted lines show error bounds for 10,000 simulations

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  $\mu$  and  $\mathbf{H} = \partial_{\tau}^2 f(\hat{\tau})$  is the Hessian of  $f(\tau)$  evaluated at  $\mu$ .

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  $f$ :

$$E[f(\hat{\tau})] \approx f(\mu) + \frac{1}{2} \text{tr}\{\mathbf{H}\hat{\Sigma}(\hat{\tau})\}. \quad (3)$$

and

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

See Appendix C for the full derivations.

### 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.

### 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).

The implications of these changes can be seen in Figure 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 is closer to the target level for small sample sizes 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.

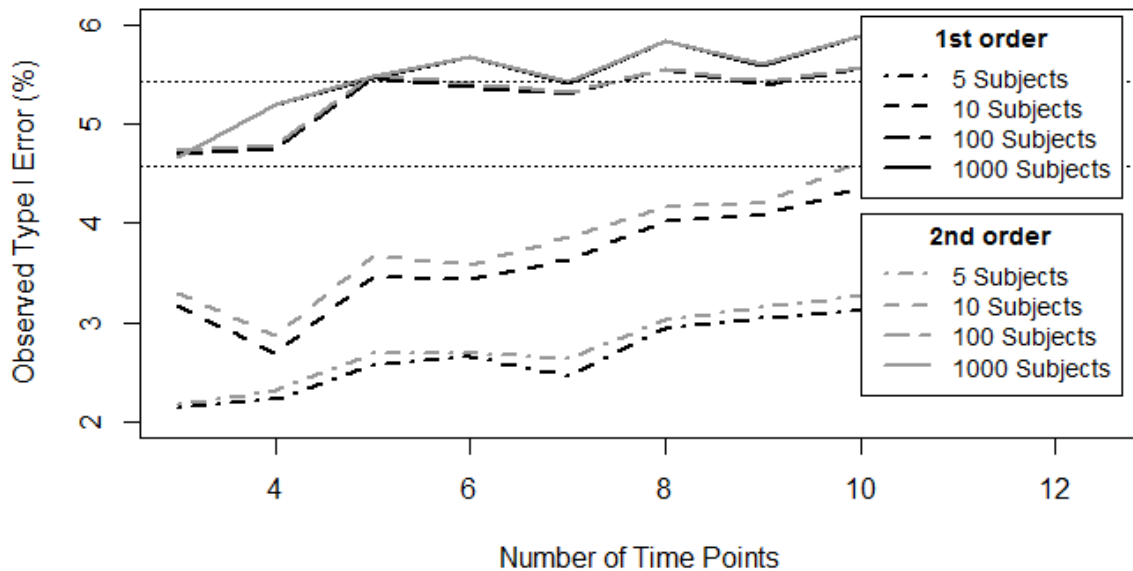


Figure 3: Comparison of observed type *I* error for varying number of time points and subjects for use of 1st and 2nd order approximation. Horizontal dotted lines show error bounds for 10,000 simulations.

### 3 Equivalence Testing

After this initial exploration of the properties of the procedure by Jensen & Ritz,<sup>5</sup> 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.<sup>9</sup> 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 familywise type  $I$  error rate is controlled. We will use the method proposed by Quan et al.<sup>10</sup> 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  $\theta$  and variance  $\Sigma$ , with  $\sigma_l = \sqrt{\Sigma_{ll}}$ , where  $\Sigma_{ll}$  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  $\Sigma$  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.



### 3.1 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.<sup>11</sup> 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.

### 3.2 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 1 and 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 4: Log Linear with Multiplicative Error

$$g(\tau, t) = \exp(\beta'_0 \exp(\beta_1 t))\epsilon, \tag{4}$$

#### Candidate Model 5: Log Linear with Multiplicative Error

$$g(\tau, t) = \beta'_0 \exp(\beta_1 t)\epsilon, \tag{5}$$

where  $\log(\epsilon) \sim N(0, \sigma^2)$ . Model 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%.

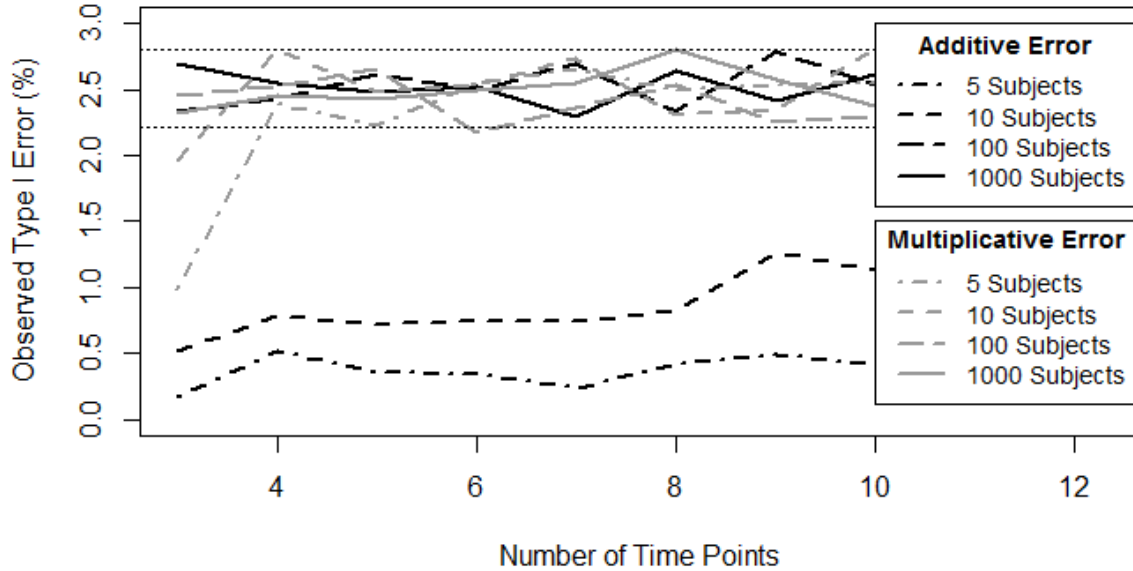


Figure 4: Comparison of observed type I error for varying number of time points and subjects for equivalence testing. Horizontal dotted lines show error bounds for 10,000 simulations.

Figure 4 show 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 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.

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.

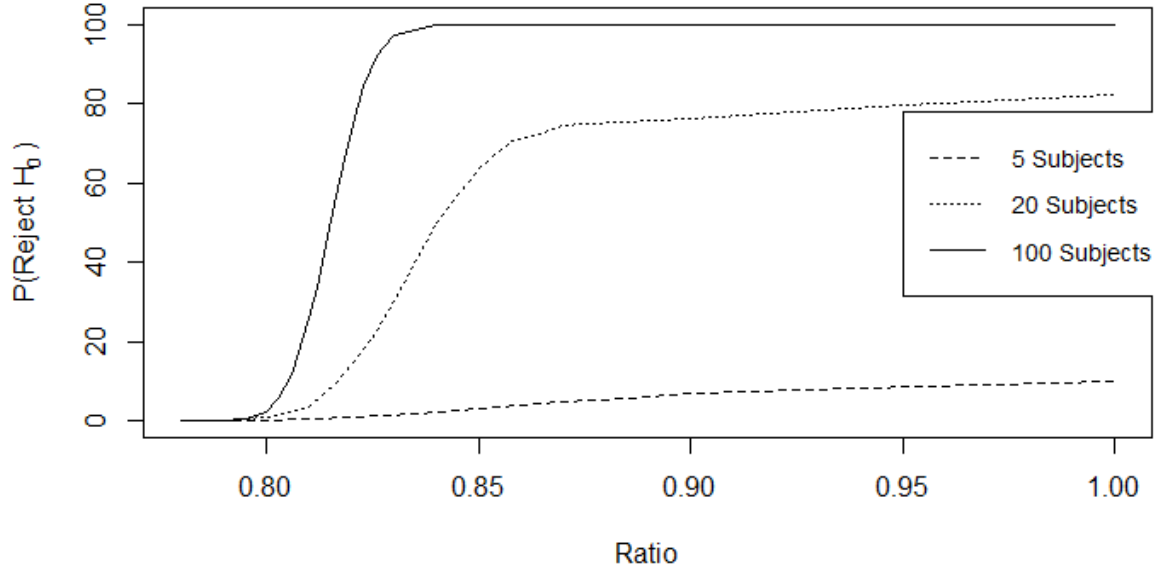


Figure 5: Power of procedure for equivalence testing with 5, 20 and 100 subjects at 3 time points.

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

**Candidate Model 6: Oral Dose with Multiplicative Error**

$$g(\tau, t) = \left( \frac{k_a F D}{V(k_a - k_e)} (e^{-k_e t} - e^{-k_a t}) \right) \epsilon, \quad (6)$$

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

**Candidate Model 7: Oral Dose with Additive Error**

$$g(\tau, t) = \frac{k_a F D}{V(k_a - k_e)} (e^{-k_e t} - e^{-k_a t}) + \epsilon, \quad (7)$$

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

Candidate model 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,<sup>12</sup> assuming the  $\log(\epsilon) \sim N(0, \sigma^2)$  with  $\sigma = 0.05$ .  $\sigma$  is chosen larger than that considered by Lunn & Aarons<sup>13</sup> and identical to one example evaluated by Tod et al.<sup>14</sup> 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 3 and 4 in Appendix E. The power is also sufficient for a reasonable sample size. Figure 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. Although these weights are quite different, 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.

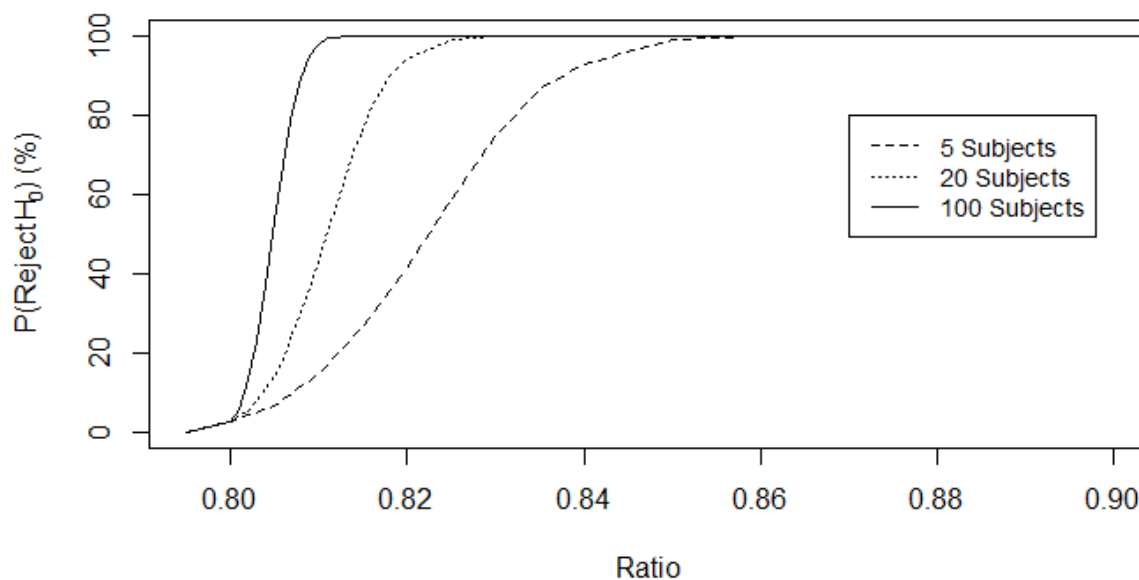


Figure 6: Power of procedure for equivalence testing with 5, 20 and 100 subjects at 7 time points for an oral administration of a compound.

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 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.<sup>15,16</sup> 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.

It is clear from Figure 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 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.

## 4 Discussion

The method by Jensen & Ritz<sup>5</sup> 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 methods<sup>5</sup> do need adjustment in order to be applied to our desired setting. However since our choice of designs are different to those discussed<sup>5</sup> 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 order

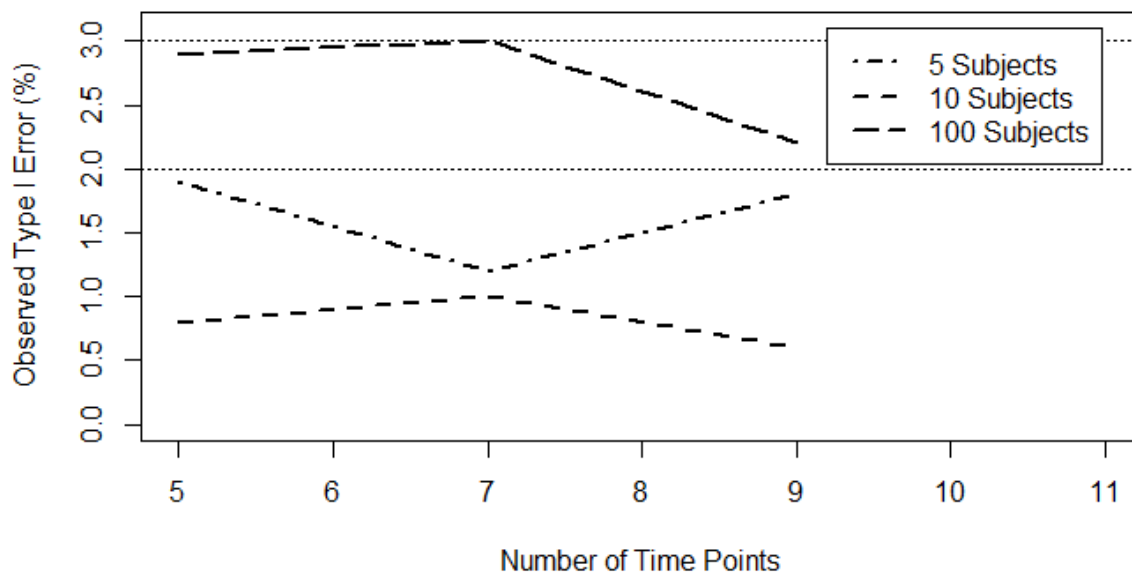


Figure 7: Type I error rate for varying numbers of time points for 5, 10 and 100 total subjects considering  $AUC$  and  $C_{max}$  as PK parameters for an oral administration of a compound. Horizontal dotted lines show error bounds for 1000 simulations. (Equivalence testing)

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. 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

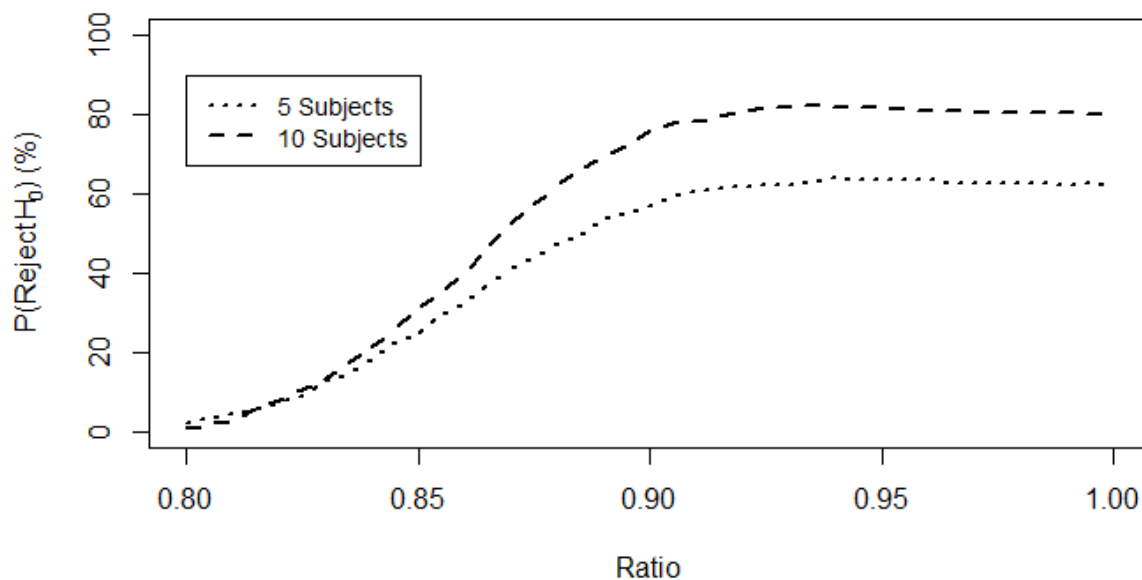


Figure 8: Power of procedure for equivalence testing for 5 time points for 5 and 10 total subjects considering  $AUC$  and  $C_{max}$  as PK parameters for an oral administration of a compound.

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.

## 5 Acknowledgments and Funding

We gratefully acknowledge the funding and support of the EPSRC funded EP/H023151/1 STOR-i centre for doctoral training and Janssen Pharmaceutica. Prof Jaki's contribution to this report is independent research supported by the National Institute for Health Research (NIHR Senior Research Fellowship, Dr Thomas Jaki, SRF-2015-08-001). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health

The authors wish to thank Loeckie de Zwart and Bianca Feyen for their help in providing data and background information on the use of microsampling.

## References

- <sup>1</sup> Chapman K, Chivers S, Gliddon D, Mitchell D, Robinson S, Sangster T, Sparrow S, Spooner N, Wilson A. Overcoming the barriers to the uptake of nonclinical microsampling in regulatory safety studies. *Drug discovery today* may 2014; **19**(5):528–532, doi:10.1016/j.drudis.2014.01.002.
- <sup>2</sup> Jaki T, Wolfsegger MJ. Non-compartmental estimation of pharmacokinetic parameters for flexible sampling designs. *Statistics in Medicine* may 2012; **31**(11-12):1059–73, doi:10.1002/sim.4386.
- <sup>3</sup> Jaki T, Wolfsegger MJ. Estimation of pharmacokinetic parameters with the R package PK. *Pharmaceutical Statistics* 2011; **10**(3):284–288, doi:10.1002/pst.449.
- <sup>4</sup> Davidian M, Giltinan DM. *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall/CRC, 1995.
- <sup>5</sup> Jensen SM, Ritz C. Simultaneous Inference for Model Averaging of Derived Parameters. *Risk Analysis* 2015; **35**(1):68–76, doi:10.1111/risa.12242.
- <sup>6</sup> Buckland ST, Burnham KP, Augustin NH. Model Selection : An Integral Part of Inference. *Biometrics* 1997; **53**(2):603–618.
- <sup>7</sup> Pippier CB, Ritz C, Bisgaard H. 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* 2012; **61**(2):315–326.
- <sup>8</sup> Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd edn., Springer-Verlag, 2002.
- <sup>9</sup> Schuirmann DJ. 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* 1987; **15**(6):657–680, doi:10.1007/BF01068419.
- <sup>10</sup> Quan H, Bolognese J, Yuan W. Assessment of equivalence on multiple endpoints. *Statistics in Medicine* 2001; **20**(21):3159–3173, doi:10.1002/sim.985.
- <sup>11</sup> Chow SC, Liu Jp. *Design and Analysis of Bioavailability and Bioequivalence Studies*. 3rd edn., CRC Press, 2008.
- <sup>12</sup> Gibaldi M, Perrier D. *Pharmacokinetics*. Marcel Dekker: New York, 1982.
- <sup>13</sup> Lunn DJ, Aarons LJ. Markov Chain Monte Carlo Techniques for Studying Interoccasion and Intersubject Variability : Application to Pharmacokinetic Data. *Journal of the Royal Statistical Society. Series C* 1997; **46**(1):73–91.
- <sup>14</sup> Tod M, Mentré F, Merlé Y, Mallet A. Robust optimal design for the estimation of hyperparameters in population pharmacokinetics. *Journal of Pharmacokinetics and Biopharmaceutics* 1998; **26**(6):689–716, doi:10.1023/A:1020703007613.
- <sup>15</sup> Jensen SM, Pippier CB, Ritz C. Evaluation of multi-outcome longitudinal studies. *Statistics in Medicine* 2015; **34**(12):1993–2003, doi:10.1002/sim.6461.



<sup>16</sup> Ritz C, Pilmann Laursen R, Trab Damsgaard C. 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; :295–311doi:10.1111/rssc.12161.

<sup>17</sup> Graybill FA. *Matrices with Applications in Statistics*. 2nd edn., Wadsworth International Group, 1983.

## A PK Parameters as Functions of Model Parameters

### A.1 For Candidate Model 1

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

$$C_{max} = B_0$$

### A.2 For Candidate Model 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 6 and 7

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

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

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

## B Sampling Time Points for Simulations

## C Derivation of Second Order Approximation

Using the notation defined in Section 2.2, a first order Taylor approximation of  $f$  is given by:

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

Table 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)

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

$$E^2[f(\hat{\tau})] \approx E \left[ \begin{array}{l} f^2(\mu) + \mathbf{D}^T(\hat{\tau} - \mu)(\hat{\tau} - \mu)^T \mathbf{D} + \frac{1}{4}(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu)(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu) \\ + 2f(\mu)\mathbf{D}^T(\hat{\tau} - \mu) + f(\mu)(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu) + \mathbf{D}^T(\hat{\tau} - \mu)(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu). \end{array} \right] \quad (8)$$

$$= f^2(\mu) + \mathbf{D}^T \hat{\Sigma}(\hat{\tau}) \mathbf{D} + \frac{1}{4} E[(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu)(\hat{\tau} - \mu)^T \mathbf{H}(\hat{\tau} - \mu)] + f(\mu) \text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\} \quad (9)$$

$$= f^2(\mu) + \mathbf{D}^T \hat{\Sigma}(\hat{\tau}) \mathbf{D} + \frac{1}{4} \left( \text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\} \right)^2 + \frac{1}{2} \text{tr}\{(\mathbf{H} \hat{\Sigma}(\hat{\tau}))^2\} + f(\mu) \text{tr}\{\mathbf{H} \hat{\Sigma}(\hat{\tau})\}. \quad (10)$$

Where (9) follows from (8) 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,<sup>17</sup> 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 (8) to (9) follows. The above theorem provides reasoning for (9) to (10). 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:

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

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.

## D Average Bias of Parameter Estimates

Table 2: 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

## E Additional Type I Error Rate Results

Table 3: 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								
# Time Points	$AUC_{24}$ & $C_{max}$				$AUC_{24}$ & $t_{max}$			
	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 4: 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								
# Time Points	$t_{max}$ & $C_{max}$				$AUC_{24}$ , $C_{max}$ & $t_{max}$			
	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