Journal of Pharmacokinetics and Pharmacodynamics (2021) 48:623–638
https://doi.org/10.1007/s10928-021-09760-1

ORIGINAL PAPER

Pharmacometric estimation methods for aggregate data, including data simulated from other pharmacometric models

Pyry Antti Juhana Välitalo[1,2]

Received: 20 December 2020 / Accepted: 3 May 2021 / Published online: 22 June 2021

Abstract

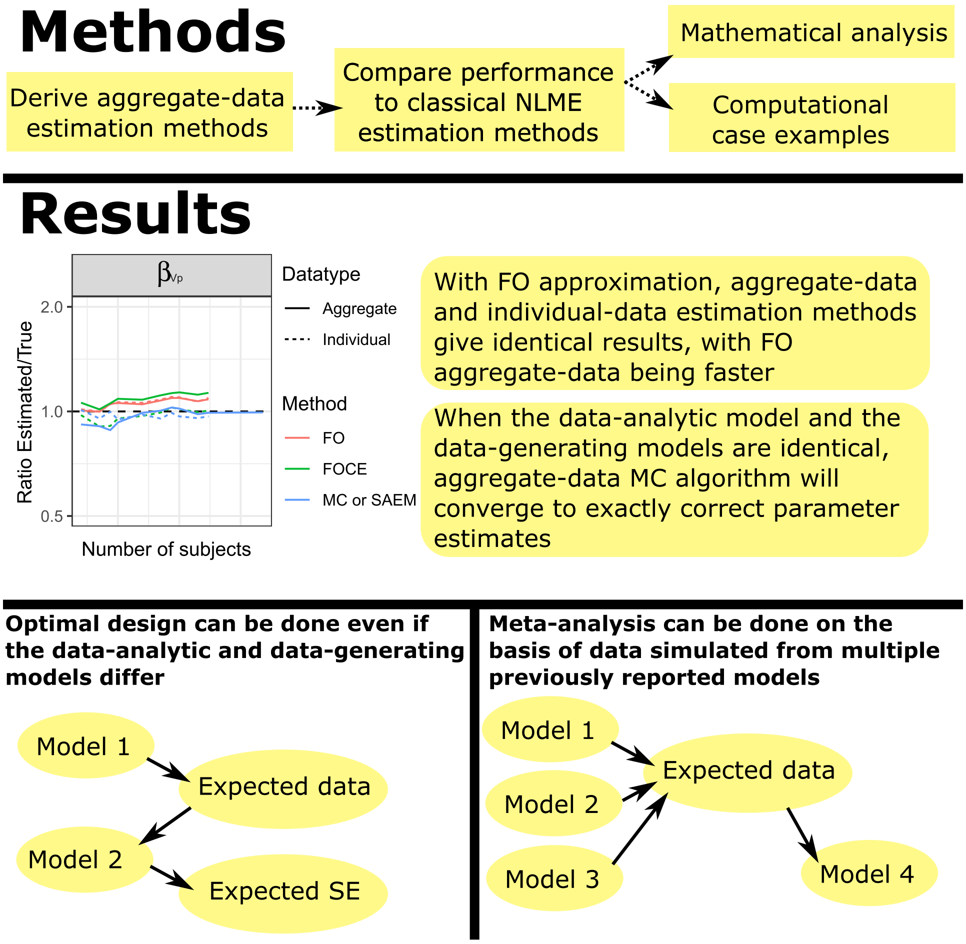
Lack of data is an obvious limitation to what can be modelled. However, aggregate data in the form of means and possibly (co)variances, as well as previously published pharmacometric models, are often available. Being able to use all available data is desirable, and therefore this paper will outline several methods for using aggregate data as the basis of parameter estimation. The presented methods can be used for estimation of parameters from aggregate data, and as a computationally efficient alternative for the stochastic simulation and estimation procedure. They also allow for population PK/PD optimal design in the case when the data-generating model is different from the data-analytic model, a scenario for which no solutions have previously been available. Mathematical analysis and computational results confirm that the aggregate-data FO algorithm converges to the same estimates as the individual-data FO and yields near-identical standard errors when used in optimal design. The aggregate-data MC algorithm will asymptotically converge to the exactly correct parameter estimates if the data-generating model is the same as the data-analytic model. The performance of the aggregate-data methods were also compared to stochastic simulations and estimations (SSEs) when the data-generating model is different from the data-analytic model. The aggregate-data FO optimal design correctly predicted the sampling distributions of 200 models fitted to simulated datasets with the individual-data FO method.

✉ Pyry Antti Juhana Välitalo
pyry.valitalo@gmail.com

[1] School of Pharmacy, University of Eastern Finland, Yliopistonranta 1 C, 70210 Kuopio, Finland

[2] Finnish Medicines Agency, Microkatu 1, 70210 Kuopio, Finland

Graphic abstract



Methods

Derive aggregate-data estimation methods

Compare performance to classical NLME estimation methods

Mathematical analysis

Computational case examples

Results

With FO approximation, aggregate-data and individual-data estimation methods give identical results, with FO aggregate-data being faster

When the data-analytic model and the data-generating models are identical, aggregate-data MC algorithm will converge to exactly correct parameter estimates

Optimal design can be done even if the data-analytic and data-generating models differ

Meta-analysis can be done on the basis of data simulated from multiple previously reported models

Keywords Pharmacometrics · Population pharmacokinetics · Aggregate data · Model-based meta-analysis

Introduction

Population pharmacokinetic estimation by nonlinear mixed-effects models was introduced by Sheiner and Beal [1–4]. In their seminal work, the authors outlined how the nonlinear mixed-effects model can be linearized around the expected values of the random effects. In practice, this enabled sparse data from individuals to be included in mathematical models of drug concentration and effect. The original estimation algorithm was named the first-order method, or the FO method for short. Several other, more accurate, estimation methods were later introduced. These estimation methods require individual-level data from subjects.

In this paper, the term "aggregate data" refers to data such as mean observed concentrations and the variance–covariance matrix of the observed concentrations. The need to analyze aggregate data may arise e.g. when performing a model-based meta-analysis. A recent example of model-based meta-analysis was published by Weber et al. [5]. In their example, a Bayesian approach for the joint analysis of individual-level data and aggregate mean data was developed. The mean aggregate data, extracted from literature, were included in the calculation of overall likelihood as data points. The observed means were contrasted with model-predicted means and covariances of simulated datapoints. The authors discussed that the mean aggregate data were generally informative for fixed-effect parameters but not for random-effects parameters. As such, the mean data alone are not enough to inform pharmacometric models. An estimation method which would allow estimation of all parameters including fixed effects, random effects and residual variability on the basis of aggregate data alone would be desirable.

Nonlinear mixed-effects models can be complex, and thus it can be complex to design experiments that aim to utilize these models. Therefore, tools that help in designing these experiments are relevant. Optimal design refers to calculation of expected Fisher Information Matrix (FIM), given a study design and some assumed model. Calculation of expected FIM for nonlinear mixed-effects models was originally reported by Mentre and coworkers using the first-order linearization [6]. One currently present limitation of optimal design is that the same model must be assumed for both data generation and data analysis.

The purpose of this manuscript is to outline an approach for fitting full pharmacometric models to aggregate data consisting of mean vector and variance–covariance matrix. The maximum likelihood estimators for aggregate data are presented. It is shown that the mean vector and variance–covariance matrix can be simulated from a priori defined models. This enables the model-based meta-analysis of data which are extracted from multiple previously reported models. It also allows optimal design in the case when the data-analytic and the data-generating models differ.

Theoretical

Definitions and notation

This manuscript uses a column vector notation and "log" refers to natural logarithm. The word "design" refers to the independent variables of the data. Usually, the most obvious design aspect is the PK/PD sampling schedule.

A design for a multi-response mixed-effects model is composed of N subjects, each with an associated elementary design $\boldsymbol{\xi}_i (i = 1, ..., N)$; hence, a design for a population of N subjects can be described as follows:

$$\boldsymbol{\Xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N) \tag{1}$$

Each elementary design $\boldsymbol{\xi}_i$ can be further divided into subdesigns:

$$\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK}) \tag{2}$$

With $\boldsymbol{\xi}_{ik}$, $k = 1, ..., K$ being the design associated with the kth response such as drug concentration or drug effect. In this manuscript dealing with aggregate data, an assumption is made that the elementary designs can also be grouped across individuals. However, the framework also allows elementary designs unique to an individual.

The individual response y_{ik} is modelled as follows:

$$y_{ik} = f_k(\boldsymbol{\theta}_i, \xi_{ik}) + h_k(\boldsymbol{\theta}_i, \xi_{ik}, \varepsilon_{ik}) \tag{3}$$

where $f_k(.)$ is the structural model for the kth response, $\boldsymbol{\theta}_i$ is the ith subject's parameter vector, $h_k(.)$ is the residual error model for response k, (often additive, proportional or a combination of additive and proportional), and ε_{ik} is the residual error for response k in subject i. The general prediction and residual error functions for all responses are denoted by $f()$ and $h()$, respectively. The residual error ε_{ik} are distributed with a mean of zero and additive and proportional variance terms as elements of $\boldsymbol{\Sigma}$. A matrix $\boldsymbol{Y}$ of responses is defined as:

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1K} \\ y_{21} & y_{22} & & \\ \vdots & & \ddots & \\ y_{N1} & & & y_{NK} \end{bmatrix} \tag{4}$$

The individual parameter vector $\boldsymbol{\theta}_i$, with parameter(s) that might be shared between responses, is described as follows:

$$\boldsymbol{\theta}_i = g(\boldsymbol{\beta}, \boldsymbol{b}_i) \tag{5}$$

where $\boldsymbol{\beta}$ is the vector of fixed effects parameters, or typical subject parameter and $\boldsymbol{b}_i$ is the vector of v random effects for subject i. The random effects $\boldsymbol{b}_i$ are assumed to be normally distributed with a mean of zero and a covariance matrix $\boldsymbol{\Omega}$ of size $v \times v$ with off-diagonal elements either as parameters to be estimated, or fixed to zero. The vector of population parameters is thus defined as follows:

$$\boldsymbol{\Psi} = [\boldsymbol{\beta}, \boldsymbol{vec}(\boldsymbol{\Omega}), \boldsymbol{vec}(\boldsymbol{\Sigma})] = \left[\boldsymbol{\beta}, \omega_1^2, \omega_{12}, \omega_2^2, \ldots, \omega_v^2, \sigma_1^2, \sigma_2^2\right] \tag{6}$$

where $\boldsymbol{vec}(.)$ refers to vectorization operation.

The aggregate mean response vector $\overline{\mathrm{y}}_{\mathrm{k}}$ at subdesign ξ_k and the aggregate (co)variance response V_{kl} between subdesigns $\boldsymbol{\xi}_{*k}$ and $\boldsymbol{\xi}_{*l}$ are defined as

$$\overline{\mathrm{y}}_{\mathrm{k}} = \frac{1}{\mathrm{N}} \sum_{i=1}^{n} y_{ik} \tag{7}$$

$$V_{kl} = \frac{1}{\mathrm{N}} \sum_{i=1}^{n} (y_{ik} - \overline{y}_k)(y_{il} - \overline{y}_l) \tag{8}$$

The mean vector $\overline{\boldsymbol{y}}$ consists of $(\overline{y}_1, \overline{y}_2, ..., \overline{y}_K)$ and the variance–covariance matrix $\boldsymbol{V}$ of the observed data is a K-K matrix, the elements of which correspond to the elements V_{kl}.

Parameter estimation when individual-level data are available

This subsection presents already known and published results about parameter estimation when individual-level data are available. It is included as an introductory, relevant background information.

The likelihood $L(\boldsymbol{y}_i|\boldsymbol{\Psi})$ of observed data $\boldsymbol{y}_i$ for individual i given parameters $\boldsymbol{\Psi}$, is defined as

$$L(\boldsymbol{y}_i|\boldsymbol{\Psi}) = \int l(\boldsymbol{y}_i|\boldsymbol{\theta}_i, \xi_i)p(\boldsymbol{\theta}_i|\boldsymbol{\Psi})d\boldsymbol{\theta}_i \tag{9}$$

where $l(\boldsymbol{y}_i|\boldsymbol{\theta}_i, \xi_i)$ is the conditional likelihood of $\boldsymbol{y}_i$ given individual parameters $\boldsymbol{\theta}_i$ and design factors ξ_i.

The overall likelihood function $L(\boldsymbol{Y}|\boldsymbol{\Psi})$ for the data is the product of individual likelihood functions, or mathematically $L(\boldsymbol{Y}|\boldsymbol{\Psi}) = \prod L(\boldsymbol{y}_i|\boldsymbol{\Psi})$. To avoid floating point errors associated with very high or low numerical values, it is common to maximize a log-likelihood function, equal to the sum of individual log-likelihoods.

In FO and FOCE approximations, the individual data are treated as multi-variate normally distributed. The log-likelihood can then be expressed as

$$\begin{aligned}\log(L(\boldsymbol{y}_i|\boldsymbol{\Psi})) &= -\frac{1}{2}\left(\boldsymbol{y}_{res,i}^T\tilde{\boldsymbol{V}}_i^{-1}\boldsymbol{y}_{res,i} + \log\left|\tilde{\boldsymbol{V}}_i\right|\right) \\ &= -\frac{1}{2}\left(tr\left(\boldsymbol{R}_i\tilde{\boldsymbol{V}}_i^{-1}\right) + \log\left|\tilde{\boldsymbol{V}}_i\right|\right)\end{aligned} \tag{10}$$

where $\boldsymbol{y}_{res,i}$ is the vector of residuals, $\boldsymbol{R}_i$ is the outer product of the residuals $\boldsymbol{R}_i = \boldsymbol{y}_{res,i}^T \times \boldsymbol{y}_{res,i}$ and $\tilde{\boldsymbol{V}}_i$ is the individual predicted variance–covariance matrix. To clarify the above expressions, we note that

$$\boldsymbol{y}_{res,i}^T\tilde{\boldsymbol{V}}_i^{-1}\boldsymbol{y}_{res,i} = \sum_k y_{res,ik}\left(\sum_l y_{res,il}\left(\tilde{\boldsymbol{V}}_i^{-1}\right)_{kl}\right) = tr\left(\boldsymbol{R}_i\tilde{\boldsymbol{V}}_i^{-1}\right) \tag{11}$$

With nonlinear mixed-effects models, there is no closed-form solution to $L(\boldsymbol{y}_i)$. Therefore, various approximations have been developed. The FO approximation linearizes the model around the expected average value of the random effect at zero, and FOCE approximation linearizes the model around the conditional maximum a posteriori estimates of $\boldsymbol{b}_i$.

$$\boldsymbol{y}_{res,FOCE,i} = \boldsymbol{y}_i - \left(f\left(\tilde{\boldsymbol{\theta}}_i, \xi_i\right) - \frac{\delta f\left(\tilde{\boldsymbol{\theta}}_i, \xi_i\right)}{\delta \boldsymbol{b}_i}\tilde{\boldsymbol{b}}_i\right) \tag{12}$$

$$\begin{aligned}\tilde{\boldsymbol{V}}_{FOCE,i} = &\left(\frac{\delta f\left(\tilde{\boldsymbol{\theta}}_i, \xi_i\right)}{\delta \boldsymbol{b}_i}\right)\boldsymbol{\Omega}\left(\frac{\delta f\left(\tilde{\boldsymbol{\theta}}_i, \xi_i\right)}{\delta \boldsymbol{b}_i}\right)^T \\ &+ diag\left(\left(\frac{\delta h\left(\tilde{\boldsymbol{\theta}}_i, \xi_i, \boldsymbol{\varepsilon}_i\right)}{\delta \boldsymbol{\varepsilon}_i}\right)\boldsymbol{\Sigma}\left(\frac{\delta h\left(\tilde{\boldsymbol{\theta}}_i, \xi_i, \boldsymbol{\varepsilon}_i\right)}{\delta \boldsymbol{\varepsilon}_i}\right)^T\right)\end{aligned} \tag{13}$$

In the above equations, expression $\frac{\delta f(\boldsymbol{\theta}_i, \xi_i)}{\delta \boldsymbol{b}_i}$ is the Jacobian matrix of model predictions with regard to random effects, with K number of rows and v number of columns, evaluated at $\tilde{\boldsymbol{b}}_i$. Similarly, expression $\frac{\delta h(\boldsymbol{\theta}_i, \xi_i, \epsilon_i = 0)}{\delta \epsilon_i}$ is the Jacobian matrix of residual variability with regard to the residual variability terms, evaluated at $\epsilon_i = \vec{0}$, with K number of rows and number of columns equal to the number of residual variability terms. In FOCE method, $\tilde{\boldsymbol{b}}_i$ is the vector of maximum a posteriori estimates of the random effects for individual i, given individual data $\boldsymbol{y}_i$ and population parameters $\boldsymbol{\Psi}$. Similarly, $\tilde{\boldsymbol{\theta}}_i$ is the vector of individual parameters that result from substituting $\tilde{\boldsymbol{b}}_i$ to equation 5. In FO method, $\tilde{\boldsymbol{b}}_i$ is a vector of zeros, therefore the above equations also cover FO approximation as a special case.

It is relevant to note that maximizing the log-likelihood with regard to the FO and FOCE approximations is not guaranteed to lead to exactly correct parameter estimates, even if the amount of subjects in dataset would approach infinity. The expected mismatch between the true values and the parameter estimates becomes more pronounced as either the variance of random effects increases, or as the number of data points per individual decreases. Newer, EM-based estimation algorithms such as the importance sampling EM algorithm [7] and the SAEM algorithm [8–10] do not have this problem. While these EM-based algorithms are still "approximations" because of Monte-Carlo sampling, they will converge to the exact maximum likelihood parameter estimates, given a sufficiently high number of iterations and Monte-Carlo samples of random effects. For reasons of conciseness, the mathematical details of importance sampling and SAEM estimation algorithms will not be covered here.

Estimation based on aggregate data

For N individuals sharing the same study design, the FO log-likelihood expression introduced in equation 10 can be simplified to equation 14. For detailed steps, please refer to Appendix 1.

$$\log(L(\bar{\boldsymbol{y}}, \boldsymbol{V}|\boldsymbol{\Psi})) = -\frac{N}{2}\left(tr\left(\boldsymbol{V}\cdot\tilde{\boldsymbol{V}}^{-1}\right) + (\bar{\boldsymbol{y}} - \tilde{\boldsymbol{y}})^T\tilde{\boldsymbol{V}}^{-1}(\bar{\boldsymbol{y}} - \tilde{\boldsymbol{y}}) + \log\left|\tilde{\boldsymbol{V}}\right|\right) \tag{14}$$

In the above equation, $\tilde{\boldsymbol{y}}$ is the vector of mean model predictions. This log-likelihood expression is highly similar to the expression used for fitting structural equation models to variance terms. As shown by Jöreskog [11], the observed variance–covariance matrix of observations is Wishart distributed and therefore the log-likelihood expression for $\tilde{\boldsymbol{V}}_i$ is identical to equation 14 with the $(\bar{\boldsymbol{y}} - \tilde{\boldsymbol{y}})^T\tilde{\boldsymbol{V}}^{-1}(\bar{\boldsymbol{y}} - \tilde{\boldsymbol{y}})$ term omitted.

The proposed expressions for the log-likelihood of aggregate data are highly similar to the expression for the log-likelihood of individual data. Both expressions involve taking the log-determinant of the predicted variance–covariance matrix, the only difference is in use of $\tilde{\boldsymbol{V}}$ versus $\tilde{\boldsymbol{V}}_i$. Further, both expressions involve taking the trace of residual variance–covariance matrix (either $\boldsymbol{V}_i$ or $\mathbf{V}$) multiplied by the inverse of the predicted variance–covariance matrix. A major difference between the two methods is that the individual likelihood $L(\boldsymbol{y}_i|\boldsymbol{\Psi})$ is calculated by integrating over the random effect values while considering the individual data $\boldsymbol{y}_i$. For the analysis of aggregate data, such a thing is not possible, and instead it is only possible to integrate over the (unknown) random effect values while considering the aggregate data; the mean vector and variance–covariance matrix of observations.

Although equation 14 was transformed from the log-likelihood of FO method, it applies generally to aggregate data consisting of means and variance–covariance matrices (please see Appendix 2).

FO, FOCE and Monte–Carlo approximations of the predicted aggregate data

In this manuscript, three methods for integrating over the random effect values $\boldsymbol{b}_i$ will be proposed: The already mentioned aggregate-data FO method, the aggregate-data FOCE method, and the aggregate-data MC method. These three methods are fed into computer optimization algorithms.

The FO aggregate method is identical to the FO method for non-aggregate data. For the FOCE aggregate method, it is not possible to estimate maximum a posteriori values of $\tilde{\boldsymbol{b}}_i$ because individual data are not available. Therefore, the FOCE aggregate method consists of Monte Carlo integration over a set of quasi-randomly sampled values of $\tilde{\boldsymbol{b}}_i$, similar but not identical to the optimal design FOCE-like approximation described by Retout and Mentré [12].

$$\tilde{\boldsymbol{y}}_{FOCE} = \frac{1}{N_{sim}} \sum_{i}^{N_{sim}} \left(f(\boldsymbol{\theta}_i, \boldsymbol{\xi}_i) - \frac{\delta f(\boldsymbol{\theta}_i, \boldsymbol{\xi}_i)}{\delta \boldsymbol{b}_{i,sim}} \boldsymbol{b}_{i,sim} \right) \tag{15}$$

$$\begin{aligned} \tilde{\boldsymbol{V}}_{FOCE} = {} & \frac{1}{N_{sim}} \sum_{i}^{N_{sim}} \left(\frac{\delta f(\boldsymbol{\theta}_i, \boldsymbol{\xi}_i)}{\delta \boldsymbol{b}_{i,sim}} \right) \boldsymbol{\Omega} \left(\frac{\delta f(\boldsymbol{\theta}_i, \boldsymbol{\xi}_i)}{\delta \boldsymbol{b}_{i,sim}} \right)^T \\ & + diag\left(\left(\frac{\delta h(\boldsymbol{\theta}_i, \boldsymbol{\xi}_i, \boldsymbol{\varepsilon}_i)}{\delta \boldsymbol{\varepsilon}_i} \right) \boldsymbol{\Sigma} \left(\frac{\delta h(\boldsymbol{\theta}_i, \boldsymbol{\xi}_i, \boldsymbol{\varepsilon}_i)}{\delta \boldsymbol{\varepsilon}_i} \right)^T \right) \end{aligned} \tag{16}$$

where $\boldsymbol{b}_{i,sim}$ is one simulated vector of random effect values out of the total of N_{sim} simulated random effect vectors. The differences between equations 15–16 and the FOCE-like approximation described by Retout and Mentré [12] are explored in the Discussion: Limitations subsection.

At this point it may be observed that numerical integration could be done directly on the raw simulated data instead of doing numerical integration over a first-order approximation. For this reason, the aggregate-data MC method is introduced.

$$\tilde{\boldsymbol{y}}_{MC} = \frac{1}{N_{sim}} \left(\vec{1}_N^T \boldsymbol{Y}_{sim} \right)^T \tag{17}$$

$$\begin{aligned} \tilde{\boldsymbol{V}}_{MC} = {} & \frac{1}{N_{sim}} \left(\boldsymbol{Y}_{sim} - \vec{1}_N^T \otimes \tilde{\boldsymbol{y}}_{MC} \right)^T \left(\boldsymbol{Y}_{sim} - \vec{1}_N^T \otimes \tilde{\boldsymbol{y}}_{MC} \right) \\ & + \frac{1}{N_{sim}} \sum_{i}^{N_{sim}} diag\left(\left(\frac{\delta h(\boldsymbol{\theta}_i, \boldsymbol{\xi}_i, \boldsymbol{\varepsilon}_i = 0)}{\delta \boldsymbol{\varepsilon}_i} \right) \boldsymbol{\Sigma} \left(\frac{\delta h(\boldsymbol{\theta}_i, \boldsymbol{\xi}_i, \boldsymbol{\varepsilon}_i = 0)}{\delta \boldsymbol{\varepsilon}_i} \right)^T \right) \end{aligned} \tag{18}$$

where $\vec{1}_N^T$ is a vector of ones of length N_{sim}, symbol $\otimes$ refers to outer product, and $\boldsymbol{Y}_{sim}$ is the $N_{sim} \times K$ matrix of simulated responses, based on function f and a set of Sobol-sampled values of $\boldsymbol{b}_{sim}$. This method allows interaction between $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$ values. It is noted that if residual error distribution is symmetric (additive, proportional or additive + proportional), then the MC aggregate method, at sufficiently high N_{sim}, will produce exactly correct predictions of mean vector and variance–covariance matrix. Therefore, if the above conditions are fulfilled, the MC aggregate method will converge to the exact maximum likelihood estimates of the aggregate data.

Applications

The aggregate data log-likelihood expressions can be directly applied in estimation, optimal design and as a replacement to stochastic simulation and estimation.

The expressions can be readily used to fit models to observed aggregate data, possibly side to side with individual-level data. For FO method, the aggregate-data estimation is guaranteed to give results identical to individual-data estimation (Appendix 1). For the aggregate-data MC method, there is a guarantee that given a sufficiently high N_{sim} and sufficiently high number of subjects in the dataset, the parameter estimates will converge to the same ones as with individual-data SAEM algorithm, if the data-generating model and the data-analytic model are the same. However, if the data-generating model and the data-analytic model are different, then individual-data SAEM algorithm will allow the random effects distributions to differ from normality to some degree; the aggregate-data MC algorithm is not able to do this. Therefore, when data-generating and data-analytic models are different, the aggregate-data MC results are likely to differ from individual-data SAEM results. Finally, for FOCE algorithm, there is no guarantee that the aggregate-data and individual-data methods give identical results. This is explored in detail in section Discussion: Limitations.

The newly developed expressions are also readily applicable to optimal design. One can simply simulate the expected aggregate data based on equations 15–18 and take the numerical Hessian of log-likelihood for the simulated aggregate data with respect to model parameters. Appendix

3 proves that taking the Hessian of the aggregate data log-likelihood directly results in the published optimal design expressions [12] for the FO approximation of population Fisher Information Matrix.

Finally, the newly developed expressions for aggregate data log-likelihood can be used as a replacement to procedures known as stochastic simulation and estimation (SSE). In these procedures, multiple datasets are simulated under the evaluated design, one or more models are fitted to each of the simulated datasets, and summaries of the parameter estimates across all fitted models are calculated. The newly developed expressions can be used to first simulate expected aggregate data from the data-generating model according to equations 15–18, and then to fit the data-analytic model to the simulated aggregate data. This needs to be done only once with sufficiently high N_{sim}, and not for multiple simulated datasets.

Methods

Four case studies are presented. The first one compares the aggregate data log-likelihood to previously published log-likelihood calculations in a paper containing the derivation of NONMEM estimation methods [13]. The second case example examines the estimation properties of the aggregate data estimation methods and compares them to the classical estimation methods based on individual data. A simulated dataset with an increasing number of simulated subjects is used, and convergence to the correct parameter values is monitored. Finally, the third and fourth case examples examine the potential of the newly derived expressions to be used in optimal design, and as a replacement for stochastic simulations and estimations.

We emphasize that these simulation case examples function as sanity checks to demonstrate some specific features of the aggregate data estimation methods. The case examples alone do not prove anything. The actual proofs regarding the aggregate data estimation methods are mathematical.

Case 1: log-likelihood value comparison

This case example shows that the aggregate-data FO approximation results in exactly identical objective function value as the individual-data FO approximation. For this case example, the model and dataset described and used by Wang [13] were utilized due to their public availability. The data table can be observed in the original publication, and will not be repeated here. As described in the original publication [13], the model for kth measurement of ith subject is specified as:

$$y_{ik} = 10 \cdot exp\left(-\beta_1 \times exp\left(b_{i,1}\right) \cdot t_{ik}\right) + \varepsilon_{ik}$$

where the distribution of ε depends on the residual error model. The model objective function was calculated at the same parameter estimates as used in the original publication, namely $\beta_1 = 0.5$, $\omega_1^2 = 0.04$, $\sigma_1^2 = 0.1$. Objective function values with additive and proportional residual error models were evaluated. The FO and FOCE algorithms were used both for individual-data and aggregate-data estimation.

Case 2: parameter estimation accuracy

This case example demonstrates that the aggregate-data MC approximation converges to the correct parameter estimates when the size of the dataset is sufficiently high. It also further demonstrates that the individual-data and aggregate-data FO approximations result in identical parameter estimates. A two-compartment mammillary model with first-order absorption and elimination is used. Parameter values of 5 L/h clearance, 10 L central volume of distribution, 30 L peripheral volume of distribution, 10 L/h inter-compartmental clearance and 1/h absorption rate constant are used for fixed effects. Log-normally distributed inter-individual variability with log-standard deviation of 0.3 is used for all parameters. No correlations between random effects were defined. Additive residual error of 0.2 mg/L is used. Pharmacokinetic sampling is performed at 0.1, 0.25, 0.5, 1, 2, 3, 5, 8 and 12 h after 100 mg study drug dosing.

Data for 3000 subjects were simulated, and models were fitted to datasets of 25, 50, 75, 100, 250, 500, 750, 1000, 2000 and 3000 subjects. Estimation methods were FO-aggregate, FO-individual, FOCE-aggregate, FOCE-individual, MC-aggregate and SAEM-individual. For FOCE-aggregate and MC-aggregate methods, the number of Monte Carlo generated random effect values was either 300 or the number of subjects in the dataset, whichever was higher.

To verify that the minimum number of 300 Monte Carlo generated random effect vectors was adequate, the Monte Carlo approximation standard error of the log-likelihood was calculated via leave-one-out cross-validation. Briefly, each of the 300 random effects vectors was sequentially left out of the log-likelihood calculation, resulting in 300 leave-one-out log-likelihood values. The standard deviation of these 300 log-likelihood values was calculated to get the Monte Carlo approximation error of log-likelihood.

Case 3: optimal design

This case example shows that aggregate-data and individual-data FO approximations result in identical predicted relative standard errors. It also explores differences in relative standard errors predicted by aggregate-data FOCE versus individual-data FOCE approximation, and standard errors predicted by aggregate-data MC versus Monte Carlo approximation [14]. The model defined in Case 2 was used, and the number of subjects was set to 100. Expected aggregate data were simulated with FO, FOCE and MC approximations of aggregate data (equations 15–18). Then, the Hessian of aggregate data log-likelihood as a function of parameters was calculated using each of the approximations. The consistency of the expected standard errors calculated this way were compared with the expected standard errors calculated with published expressions of population FIM. The FO and FOCE approximated population FIM was calculated with the established R library *PopED*, with 300 random effect Monte Carlo samples for the FOCE approximation. The Monte Carlo approximated population FIM, as detailed by Riviere et al. [14], was calculated with the R library *MIXFIM* with 5000 MC samples and 500 MCMC samples.

Case 4: stochastic simulation and estimation

This case example demonstrates how the aggregate-data expressions can be used to replace stochastic simulations and estimations. The data-generating model was defined as a transit compartmental model with mean transit time of 1 h and 2 transit compartments, both having log-normally distributed inter-individual variability with log-standard deviation of 0.3. The data-analytic model remained the same as defined in Case 2. The number of subjects was set to 100. Expected aggregate data were simulated using equations 17–18. The data-analytic model was fitted to the expected aggregate data using aggregate-data FO, FOCE and MC methods, and parameter estimates together with expected standard errors were calculated for each method. For conciseness, this procedure is from now on referred as "aggregate-data OD". In this case, the data-analytic model was different from the data-generating model and therefore it would not have been possible to use the previously published optimal design expressions.

The standard errors, predicted by the aggregate-data estimation, were evaluated by comparing them to SSE results. A total of 200 datasets were first simulated as individual data using the data-generating model, and the data-analytic model was fitted to each of the datasets using the individual-data FO, FOCE and SAEM algorithms; this procedure is henceforth referred as "individual-data SSE". Then, from each of the generated datasets, an aggregate dataset was calculated using equations 7 and 8 and the data-analytic model was fitted using the aggregate-data FO, FOCE and MC methods; this procedure is henceforth referred as "aggregate-data SSE". The means and standard deviations of the parameter estimates were calculated for both individual-data FOCE estimations and aggregate-data FOCE estimations.

Software and algorithms

R version 3.6.0 was used for computations and visualizing the results. For numerical integration across random effect values, Sobol-sequenced random numbers were used via *randtoolbox* R library, because these low discrepancy sequences are superior to pseudo-random numbers for numerical integration [15]. The *nlmixr* R library was used for fitting the FO, FOCE and SAEM models to individual-level data [16]. The *PopED* R library was used for generating the expected FIM for FO and FOCE approximations [17]. Although *PopED* uses different expressions for population FIM than the one featured in Appendix 1 [12], the two sets of expressions yield identical results [18] and thus the use of *PopED* is justified. The *MIXFIM* R library was used to calculate the Monte Carlo exact population FIM [14]. The R library *numDeriv* was used for calculation of accurate numerical derivatives. In addition, several other R libraries (*tidyverse*, *Rcpp*) were used to speed up the computations and to keep the code concise. The source code for all computations is available as an electronic supplement.

Results

Case 1: log-likelihood value comparison

This case example demonstrates that the individual-data and aggregate-data FO approximations result in identical log-likelihood values. As shown in Table 1 for FO method, there is a perfect match between the OFV values calculated with the aggregate data method versus the OFV values

Table 1 Objective function values for the Wang 2007 model and dataset [13], comparing the aggregate data estimation method to results from nlmixr and the original published calculations

Approximation	Residual error type	Aggregate data	Reference
FO	**Additive**	**0.0258**	**0.0258**
FOCE	Additive	− 0.0659	− 2.0588
FO	**Proportional**	**39.2132**	**39.2132**
FOCE	Proportional	39.2008	39.2067
FOCEI	Proportional	39.2027	39.4576

The reference values are those originally reported for NONMEM [13], and subsequently replicated with nlmixr. The rows where the aggregate data estimation matches perfectly the individual data estimation are highlighted in bold calculated with the individual data method. For FOCE and FOCEI, the two methods give similar but not identical results. This indicates that the FOCE(I) aggregate data method is not equivalent to the FOCE individual data method.

Case 2: parameter estimation accuracy

Figure 1 shows that the FO method gives nearly identical results for aggregate and individual data. This further confirms that the aggregate data are a sufficient statistic of the individual data in the case of FO method. The FO and FOCE methods, either aggregate or individual-data based, do not converge to the correct parameter values. This is expected, since the FO and FOCE likelihood functions are approximations, not guaranteed to lead to the exactly correct parameter estimates.

The SAEM method converges to the correct parameter estimates roughly when the dataset has 3000 thousand subjects. There seems to be some remaining bias in parameters β_{ka}, β_{Vc} and ω^2_Q, however it can safely be assumed that this remaining bias would also disappear with further increasing dataset size, as the SAEM has been proven to converge to the exact maximum likelihood estimates [9]. The aggregate-data MC method also converges to the correct parameter estimates albeit slower, in this case with a dataset size up to 23,000 subjects. Again, the actual proof is mathematical (see Appendix 2), and the simulation example shown here serves as a sanity check.

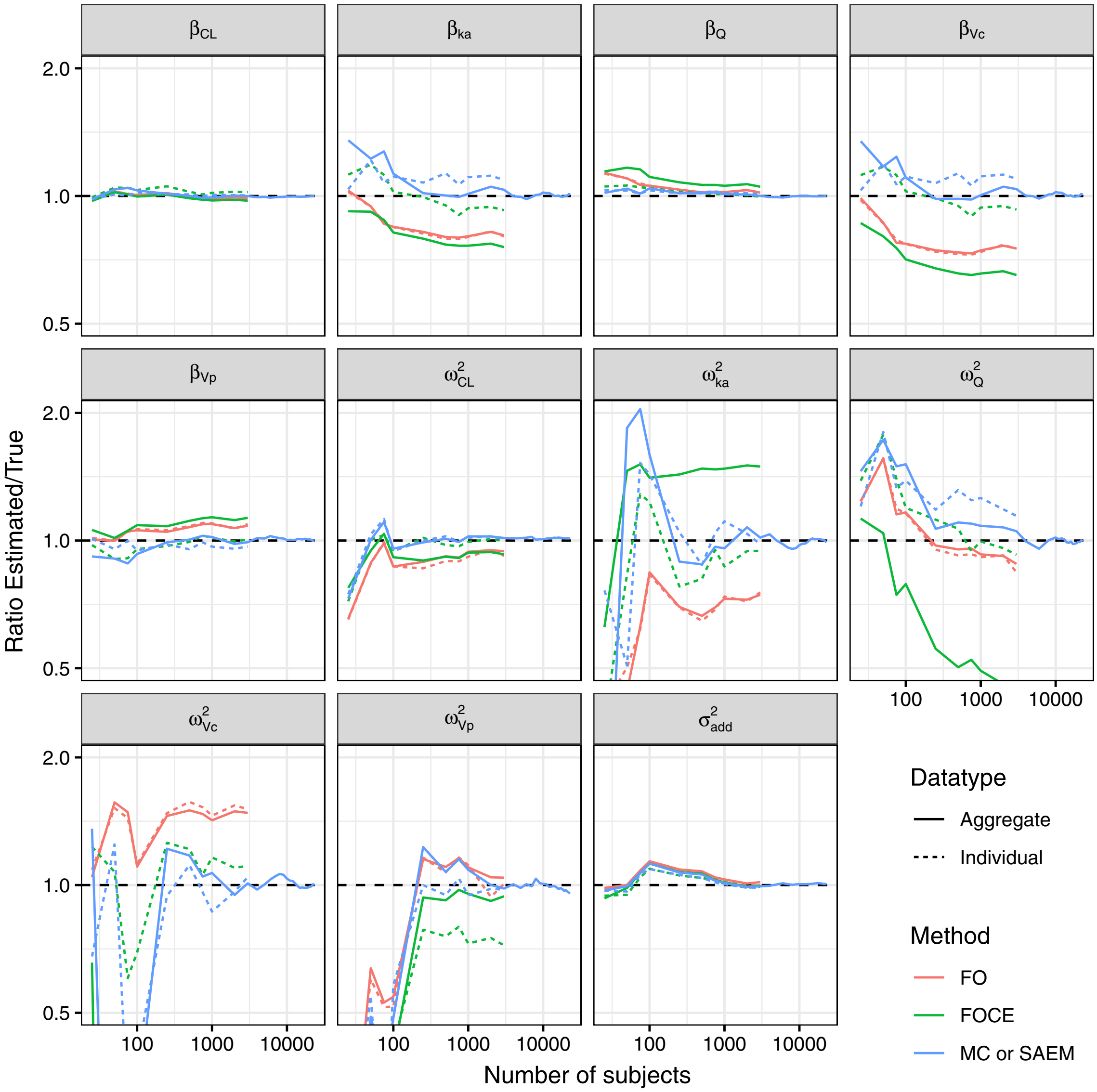


Fig. 1 Parameter estimate ratios for different estimation algorithms as a function of number of simulated subjects. The exact likelihood estimation method for individual data was SAEM, and for aggregate data the aggregate-data MC estimation method

For the aggregate-data MC estimation method with 300 MC simulated subjects, the Monte Carlo approximation standard error of the log-likelihood function was calculated via leave-one-out cross-validation. The approximation standard error was 0.0011 per one subject in the dataset, and would e.g. correspondingly be 1.1 for a dataset of 1000 thousand subjects. This approximation error was considered to be both acceptable, and to have a sufficient safety margin. Thus, it was concluded that 300 MC simulated subjects is an acceptable minimum to use in Case examples 2, 3 and 4.

Case 3: optimal design

Figure 2 shows that the FO method gives identical results for aggregate and individual data, as expected (see Appendices 1 and 3). The highest relative difference between the two methods was a 0.045% greater RSE predicted by PopED, as compared to the aggregate data log-likelihood, and this difference is likely to result from numerical differences in computations.

For FOCE the method, the RSE predictions differ between PopED and the aggregate data log-likelihood method, as is expected based on the implementation details. In general, the standard errors predicted for the

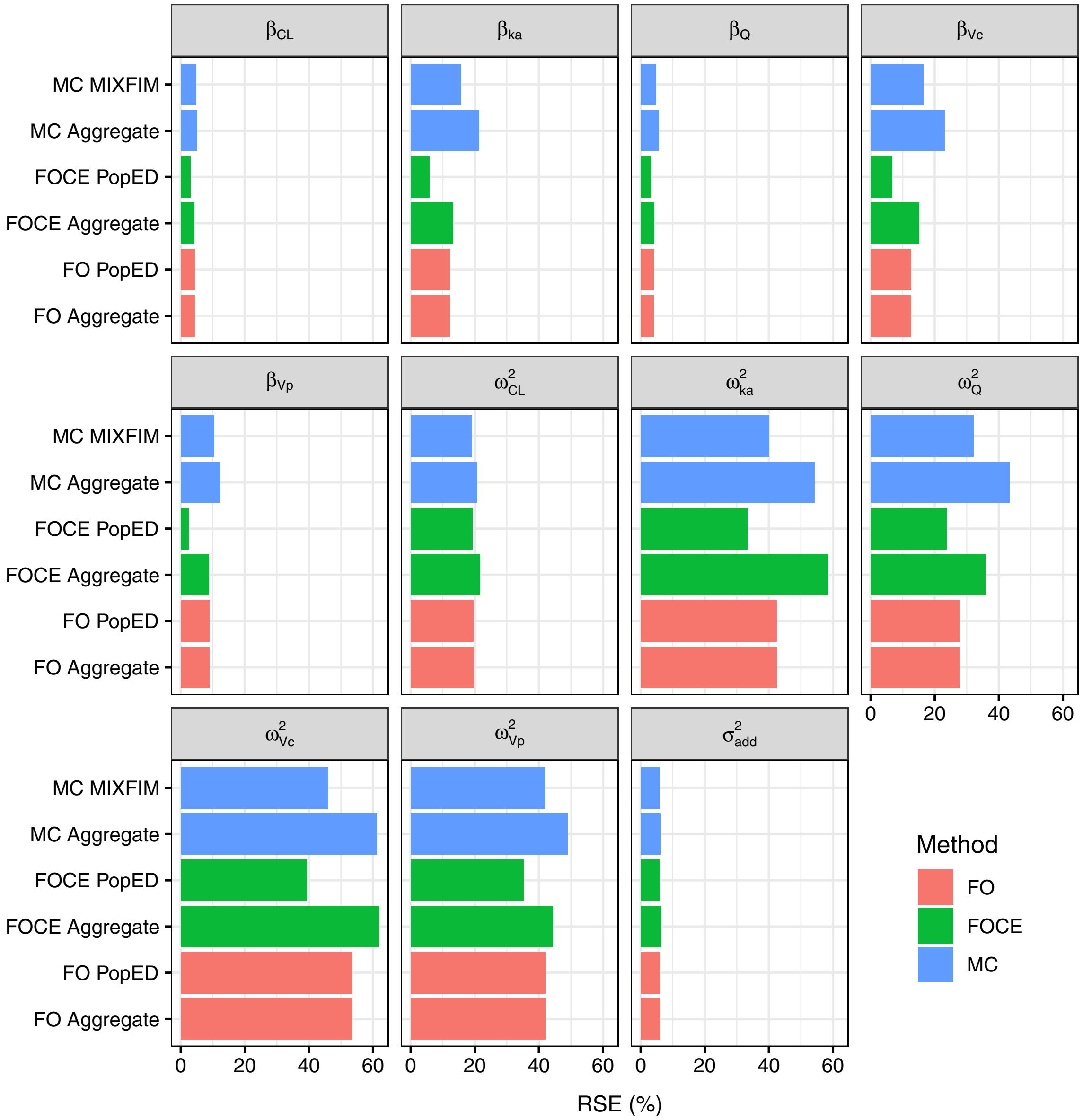


Fig. 2 Predicted relative standard errors for a study of 100 subjects on the basis of different optimal design algorithms. RSE% is relative standard error. FO refers to first-order estimation, FOCE refers to first-order conditional estimation, and MC refers Monte Carlo approximation of the log-likelihood

FOCE aggregate method are higher. Similarly, the standard errors predicted by MIXFIM are lower than the standard errors predicted by the MC aggregate method.

The highest relative standard errors are predicted for the random effects variances of Vc, Vp, Q and KA, followed by the fixed-effects estimates of KA and Vc (Fig. 2). These predictions generally agree with the results presented in Fig. 1.

Case 4: stochastic simulation and estimation

As seen in Fig. 3, there was an almost perfect match between the aggregate-data OD, aggregate-data SSE and individual-data SSE parameter estimates for FO method. Further, for nearly all parameters there is a good agreement between aggregate-data OD and aggregate-data SSE procedure results, for the non-FO estimation methods. These comparisons show that the aggregate-data OD methods reliably predict the sampling distributions of parameters obtained from aggregate-data SSE.

However, the individual-data SSE procedure gave very differing results for non-FO estimation methods. Generally, random effects variances were lower for individual-data SSE procedure for non-FO estimation methods, and residual variance was higher. The fixed-effects estimates also differed, with no clearly identifiable trends. This demonstrates that when the data-generating model is different from the data-analytic model, the individual-data algorithms and the aggregate-data algorithms do not necessarily converge to the same parameter estimates.

Discussion

In this paper, we have presented a method for estimation of nonlinear mixed-effects model parameters based on aggregate data of the actual observations. The aggregate data in this paper refers to the mean vectors and the variance–covariance matrix of observations. We have shown how the expressions can be used for fitting models, performing optimal design, and replacing computation-intensive SSE procedures with a faster and more deterministic alternative: Fitting aggregate data models to asymptotically simulated aggregate data. We have shown both mathematically and via computational examples that the aggregate-data FO and individual-data FO algorithms produce identical results given data.

Using FO method, the individual-data and aggregate-data modelling methods are mathematically identical, as was shown in Appendix 1. This means that the parameter estimation properties and optimal design properties are also identical. In the case of FO method, the $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{V}}$ are the same for each individual with the same study design and covariate values, and the observed mean vector and

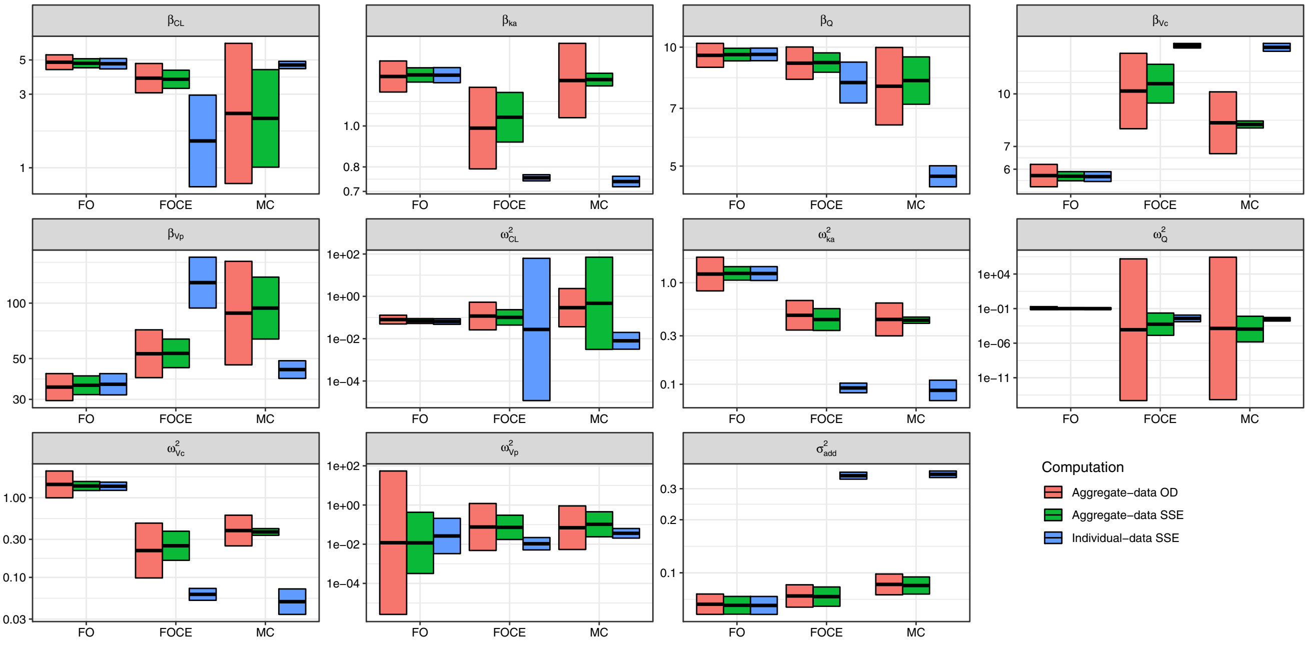


Fig. 3 The parameter estimates and their variances when the data-generating model is different from the data-analytic model. The "Aggregate-data OD" refers to fitting the data-analytic model to the expected data, simulated from the data-generating model. The "Aggregate-data SSE" and "Individual-data SSE" labels refer to repeatedly simulating a dataset of 100 individuals and fitting a model to both the simulated individual data, and to aggregate data calculated from the simulated data. FO refers to first-order estimation, FOCE refers to first-order conditional estimation, and MC refers to SAEM algorithm for individual data, and aggregate-data MC method for aggregate data

variance–covariance matrix are a sufficient statistic for the individual observed data. The implementation presented here is expected to be faster since the log-likelihood for all subjects with an identical design is calculated at once, and not one individual at a time. Therefore, with N subjects, if the individual-data FO log-likelihood evaluation takes x time units, the aggregate-data FO method will take x/N time units plus the time for the computation of the one additional term that is not included in the individual-data log-likelihood (compare Eqs. 10 and 14).

While this manuscript deals with aggregate data, the definition of designs in equations 1 and 2 also allows data from a single individual with a unique design, i.e. unique sampling timepoints. This would subsequently result in individual-level data, because the data could not be grouped with other observations occurring at the same sampling timepoints. If individual data are included as "aggregate data", then the mean vector becomes the vector of observed data, and the variance–covariance matrix becomes a matrix of zeros. Consequently, the aggregate-data log-likelihood outlined equation 14 would collapse to the individual-data log-likelihood outlined in equation 10. In practice, the aggregate-data FO would then function equivalently to individual-data FO estimation, whereas aggregate-data FOCE and aggregate-data MC estimation methods would function as computationally and statistically inefficient individual-data estimation algorithms.

The individual-data FO and the aggregate-data FO methods gave practically identical results in Case 1 and Case 3, however small differences in parameter estimates were observed in Case 2. We suspect that the small differences between the parameter estimates are caused by differences in implementation of computations. Thus, although the mathematical expressions of individual-data and aggregate-data produce identical results (Appendix 1), small differences can be seen in the computational examples due to differences in numerical implementation.

In Case 4: Stochastic simulation and estimation, we directly fitted the data-analytic model to the expected dataset, given the study design and the data-generating model. The results of this analysis were compared to two SSE scenarios, the first of which was performed based on individual-level data, and the second of which was performed based on aggregate-level data. The mean and variability of parameter estimates from the aggregate-level SSE was almost identical to the parameter estimate sampling distributions predicted by aggregate-data OD. However, the individual-data SSE parameter estimates with FOCE and SAEM algorithms differed from the aggregate-data FOCE and aggregate-data MC parameter estimates. This observation reconfirms that there are differences between the individual-level estimation and aggregate-level estimation algorithms, except for the FO approximation.

For the aggregate-data MC method, the required number of simulated subjects, N_{sim}, is dependent on both model and design. A higher number of simulated subjects will likely be required as the number of random effects increases, as the variance of random effects increases, or as the sensitivity of model predictions to random effects increases. Additionally, design aspects such as the number and timing of observations can affect the required number of simulated subjects. An excessively high N_{sim} will result in needlessly slow computation, whereas a too low N_{sim} will result in inaccurate results. In this manuscript, the minimum number of simulated subjects was set to 300, which was empirically found to produce a robust calculation of log-likelihood for the data-analytic model used in Case example 2; for details, see section Results: Case Example 2. As a potential future improvement, it should be possible to dynamically adjust the N_{sim} so that the Monte Carlo approximation error of the log-likelihood function satisfies some constraint.

Potential applications

Model-based meta-analysis traditionally refers to collecting data from multiple studies, possibly using only literature reports of aggregate data such as means and variances. The methods presented here allow the inclusion of aggregate data in the form of variance–covariance matrices. More so, the methods demonstrated here allow the simulation of aggregate data from literature models, and using these simulated aggregate data as part of fitting the meta-analytic model. In other words, data can be extracted from multiple literature-reported models and combined into a single meta-analytic model. This is an improvement, since thus far it has been necessary to choose one model if multiple models have been reported for a phenomenon.

For using literature models as data, it is not necessary to know the variance–covariance matrix of the model parameters; it is sufficient to only know the model structure, the parameter estimates, and the design aspects such as the PK sampling timepoints. If the literature model is not well-informed, then the lack of information will be captured by the sparsity of the data together with the high variances in the simulated variance–covariance matrix. This was practically demonstrated in Case 3, where only the model and the design were required to compute the expected Fisher information of model parameters, from which the variance–covariance matrix of model parameters can be calculated.

For optimal design, the proposed expressions allow optimizing designs of the data-analytic model when the data-generating model is different from the data-analytic model. In practice, this could mean using a PBPK model as the data-generating model and optimizing the sampling schedule for a two-compartment population PK model. In previous work, there have been examples of a multi-step approach, where data are first simulated from a PBPK model, a population PK model is fitted to the simulated data, and then the population PK model is used as the basis for optimal design [19–21]. The attractive alternative presented here is that a one-step approach can be used, i.e. the "fitting" of the population PK model is performed at the same step as the optimal design.

Limitations

The aggregate data estimation methods proposed in the current manuscript are based on the mean vector and the variance–covariance matrix of observations. An important limitation to using literature-reported data is that the covariances of observations are rarely reported, typically only means and variances can be extracted from the literature. If no other data than means and variances are available, then it is still theoretically possible to fit some simple nonlinear mixed-effects models, if the random effects affect the variances in a uniquely identifiable way. This scenario would be similar to fitting individual-data nonlinear mixed effects models to datasets with only one observation per subject, which is also theoretically possible. However, in practice it is difficult to make meaningful inferences from such models. To conclude, if only means and variances are available, fitting nonlinear mixed-effects to these data may not be worthwhile. However, aggregate data of means and variances can readily be used jointly with individual data in nonlinear mixed-effects model estimation.

Apart from FO approximation, population modelling with individual data is expected to always be more powerful than modelling with aggregate data. This was demonstrated in the Case 3, in which the expected standard errors for aggregate-data FOCE were higher than those for individual-data FOCE, and likewise the standard errors for aggregate-data MC estimation were higher than those predicted by MIXFIM. The likely reason for this inefficiency of aggregate data estimation is that information is lost when summarized only by means and variance-covariances, i.e. the first and second statistical moments. This amounts to assuming that the data are normally distributed. The third and fourth statistical moments are skewness and kurtosis. Theoretically, implementing skewness and kurtosis as a form of aggregate data would likely further improve the efficiency of aggregate data MC estimation algorithm, however deriving the log-likelihood expressions for skewness and kurtosis would be a challenge.

Case 4 demonstrated that the parameter estimates obtained by aggregate data estimation can be different from parameter estimates obtained by individual data estimation when the data-analytic and the data-generating models are different. Because the data-analytic model is different from the data-generating model, the "correct" parameter estimates for the data-analytic model are unknown. Thus, although Fig. 3 shows that e.g. the residual variance parameter resulting from the aggregate-data estimation was closer to the residual variance value of 0.04 of the data-generating model, this should not be interpreted as aggregate-data estimation being superior to individual-data estimation. It is more relevant to consider the general properties of aggregate-data versus individual-data estimation when deciding which one is more accurate: Whereas the aggregate data estimation methods must assume that the random effects are distributed perfectly normally, individual data estimation methods can allow some degree of skewness or kurtosis in random effects distribution if it results in a better agreement between predictions and data. As such, the aggregate-data estimation methods are likely less robust towards model misspecification than the individual-data estimation methods. Individual-data FOCE and SAEM estimation methods are thus expected to be superior to the aggregate-data FOCE and MC estimation methods in real-life pharmacometric analyses, where the true data-generating model is unknown. However, it is worth noting that even when conducting individual-data modelling, identifying the distribution of random effects as clearly non-normal would likely lead to model refinement until no obvious misspecification can any longer be detected. Furthermore, even if the individual-data estimation methods are more robust against non-normality of the random effects, this advantage is lost when parametrically simulating future trials from the models.

A further reason for individual-data FOCE being superior to aggregate-data FOCE is that when individual data are available, the log-likelihoods are calculated individually and only summed together at the end. On the other hand, when aggregate data are used, the expected mean vector $\tilde{\boldsymbol{y}}$ and variance–covariance matrix $\tilde{\boldsymbol{V}}$ are calculated based on a set of quasi-random individual parameters (equations 15–16). Then, the mean expected $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{V}}$ are used in the calculation of aggregate data log-likelihood. However, the individual-data log-likelihood (equation 10) involves nonlinear functions such as the inverse of $\tilde{\boldsymbol{V}}$, and taking a mean of inverse is not the same taking an inverse of mean. Therefore, taking a sum of individual log-likelihoods (calculated based on individually predicted $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{V}}$) is more accurate than calculating an aggregate-data log-likelihood on the basis of mean predicted $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{V}}$.

For the above reasons, the usefulness of the aggregate-data FOCE approximation is limited. The approximation is computationally expensive because derivatives need to be calculated for a large set of random effects, while there is no guarantee that estimation results would be identical to those estimated by individual-data FOCE method. The aggregate-data MC approximation is generally faster than aggregate-data FOCE approximation due to not having to calculate derivatives. Further, the aggregate-data MC approximation is guaranteed to asymptotically converge to the correct parameter estimates if the data-analytic model is the same as the data-generating model. Meanwhile, the aggregate-data FO method is guaranteed to give results identical to individual-data FO method, regardless of whether the data-analytic model is the same as the data-generating model. To summarize, aggregate-data FO and MC approximations are considered useful, whereas aggregate-data FOCE is not.

The currently proposed expressions do not include covariates, but it should be easy to extend the framework to include them using e.g. the results of Hooker and coworkers [17, 22]. Further, inter-occasion variability is not included in the currently proposed expressions, but should be easy to include in the same manner as done by Retout and colleagues [12]. Indeed, the currently proposed expressions have reserved the subscript j exactly for the purpose of denoting jth occasion for ith individual.

Conclusions

The presented methods for fitting nonlinear mixed-effects models to aggregate data are considered a valuable addition to the pharmacometric modelling toolbox. Future studies should explore the properties of aggregate-data estimation in model-based meta-analysis, and in conducting optimal design when the data-analytic model is different from the data-generating model.

Appendix 1: Derivation of aggregate-data log-likelihood from individual-data FO log-likelihood

In this Appendix, the FO log-likelihood expression for aggregate data (equation 14) is derived from the FO log-likelihood expression for individual data (equation 10).

We start with the log-likelihood expression for all data, as the sum of log-likelihoods of individual data. We note that in FO method, $\tilde{\boldsymbol{V}}_i$ is same for all individuals with the same design factors and can thus be substituted with $\tilde{\boldsymbol{V}}$.

$$\begin{aligned}\log(L(\boldsymbol{y}|\boldsymbol{\Psi})) &= -\frac{1}{2}\sum_i\left(tr\left(\boldsymbol{R}_i\tilde{\boldsymbol{V}}_i^{-1}\right)+\log\left|\tilde{\boldsymbol{V}}_i\right|\right)\\ &= -\frac{1}{2}\left(tr\left(\left(\sum_i\boldsymbol{R}_i\right)\tilde{\boldsymbol{V}}^{-1}\right)+N\log\left|\tilde{\boldsymbol{V}}\right|\right)\end{aligned}$$

Observing the matrix $\sum_i\boldsymbol{R}_i$

$$\sum_i\boldsymbol{R}_i=\sum_i\boldsymbol{y}_{res,i}^T\times\boldsymbol{y}_{res,i}$$

The mth row and nth column of this matrix correspond to

$$\left(\sum_i\boldsymbol{R}_i\right)_{m,n}=\sum_i(y_{im}-\tilde{y}_m)(y_{in}-\tilde{y}_n)$$

We can twice include the terms for mean observations as follows

$$\left(\sum_i\boldsymbol{R}_i\right)_{m,n}=\sum_i(\overline{y}_m+(y_{im}-\overline{y}_m)-\tilde{y}_m)(\overline{y}_n+(y_{in}-\overline{y}_n)-\tilde{y}_n)$$

Expanding the above gives

$$\left(\sum_i\boldsymbol{R}_i\right)_{m,n}=\sum_i\begin{pmatrix}+\overline{y}_m\overline{y}_n+\overline{y}_m(y_{in}-\overline{y}_n)-\overline{y}_m\tilde{y}_n\\ +(y_{im}-\overline{y}_m)\overline{y}_n+(y_{im}-\overline{y}_m)(y_{in}-\overline{y}_n)-(y_{im}-\overline{y}_m)\tilde{y}_n\\ -\tilde{y}_m\overline{y}_n-\tilde{y}_m(y_{in}-\overline{y}_n)+\tilde{y}_m\tilde{y}_n\end{pmatrix}$$

It can be noted that $\overline{\boldsymbol{y}}$ is not affected by i. Further, $\tilde{\boldsymbol{y}}$ is not affected by i. Even though $(y_{ik}-\overline{y}_k)$ are different at each value of i, the mean of these values is zero. Therefore,

$$\sum_i(y_{im}-\overline{y}_m)(\overline{y}_n-\tilde{y}_n)=(\overline{y}_n-\tilde{y}_n)\sum_i(y_{im}-\overline{y}_m)=0$$

$$\sum_i(y_{in}-\overline{y}_n)(\overline{y}_m-\tilde{y}_m)=(\overline{y}_m-\tilde{y}_m)\sum_i(y_{in}-\overline{y}_n)=0$$

Also, it can be observed that the sum of residual cross-products is actually maximum likelihood variance–covariance matrix:

$$\begin{aligned}\sum_i(y_{im}-\overline{y}_m)(y_{in}-\overline{y}_n) &= N\frac{\sum_i(y_{im}-\overline{y}_m)(y_{in}-\overline{y}_n)}{N}\\ &= NV_{mn}\end{aligned}$$

Substituting this back to $R_{m,n}$ and removing the terms that amount to zero gives

$$\left(\sum_i\boldsymbol{R}_i\right)_{m,n}=N(V_{mn}+(\overline{y}_m-\tilde{y}_m)(\overline{y}_n-\tilde{y}_n))$$

With this result, we can write

$$\sum_i\boldsymbol{R}_i=N\boldsymbol{V}+N(\overline{\boldsymbol{y}}-\tilde{\boldsymbol{y}})^T(\overline{\boldsymbol{y}}-\tilde{\boldsymbol{y}})$$

Further substituting this into the log-likelihood expression for individual data, and omitting the subscript i, gives

$$\begin{aligned}\log(L(\boldsymbol{y}|\boldsymbol{\Psi})) &= -\frac{1}{2}\left(tr\left(N\boldsymbol{V}\tilde{\boldsymbol{V}}^{-1}\right) + N\,tr((\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}})^T(\bar{\boldsymbol{y}}-\ \tilde{\boldsymbol{y}})\tilde{\boldsymbol{V}}^{-1}) + N\log\left|\tilde{\boldsymbol{V}}\right|\right)\\ &= -\frac{1}{2}N\left(tr\left(\boldsymbol{V}\tilde{\boldsymbol{V}}^{-1}\right) + (\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}})^T\ \tilde{\boldsymbol{V}}^{-1}(\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}}) + \log\left|\tilde{\boldsymbol{V}}\right|\right)\end{aligned}$$

Which is termed the aggregate data log-likelihood in the current manuscript.

Appendix 2:Derivation of the general aggregate-data log-likelihood

With a model that only produces aggregate-level data predictions, the log-likelihood for any individual datapoints is

$$\log(L(\boldsymbol{y}|\boldsymbol{\Psi})) = -\frac{1}{2}\sum_i\left(\boldsymbol{y}_{res,i}^T\tilde{\boldsymbol{V}}^{-1}\boldsymbol{y}_{res,i} + \log\left|\tilde{\boldsymbol{V}}\right|\right)$$

where $\tilde{\boldsymbol{V}}$ is the population-level variance–covariance matrix, and $\boldsymbol{y}_{res,i}$ is the vector of individual residuals, calculated on the basis of population-level mean predictions and individual data.

Using the same steps as outlined in Appendix 1, it is then possible to show that the aggregate-data log-likelihood can be expressed as

$$\log(L(\boldsymbol{y}|\boldsymbol{\Psi})) = -\frac{1}{2}N\left(tr\left(\boldsymbol{V}\tilde{\boldsymbol{V}}^{-1}\right) + (\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}})^T\tilde{\boldsymbol{V}}^{-1}(\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}}) + \log\left|\tilde{\boldsymbol{V}}\right|\right)$$

This is an exact expression for the aggregate-data log-likelihood. Therefore, maximizing the log-likelihood of this expression as a function of $\boldsymbol{\Psi}$ will result in the exact maximum likelihood estimates with respect to data, similar to the EM individual-data estimation algorithms such as stochastic approximation EM and importance sampling EM algorithms.

Appendix 3:Derivation of the expected FIM when the data-generating model is the same as the data-analytic model

Retout et al. [23], derived expressions for the expected FIM given a log-likelihood function.

Translating and extending their notation into the format used in the current manuscript and omitting the constant gives

$$\log(L(\boldsymbol{\Psi},\boldsymbol{y})) = (\boldsymbol{y}-\tilde{\boldsymbol{y}})^T\tilde{\boldsymbol{V}}^{-1}(\boldsymbol{y}-\tilde{\boldsymbol{y}}) + \log\left|\tilde{\boldsymbol{V}}\right|$$

The authors [23] used expressions

$$\frac{\delta\log\left|\tilde{\boldsymbol{V}}\right|}{\delta\lambda_k} = tr\left(\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\lambda_k}\right)$$

and

$$\frac{\delta\tilde{\boldsymbol{V}}^{-1}}{\delta\lambda_k} = -\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\lambda_k}\tilde{\boldsymbol{V}}^{-1}$$

To show that

$$E\left(\frac{\delta^2(-\log(L(\boldsymbol{\Psi},\boldsymbol{y})))}{\delta\boldsymbol{\beta}\delta\boldsymbol{\beta}^T}\right) \cong \frac{\delta f^T(\boldsymbol{\beta},\boldsymbol{\xi})}{\delta\boldsymbol{\beta}}\tilde{\boldsymbol{V}}^{-1}\frac{\delta f(\boldsymbol{\beta},\boldsymbol{\xi})}{\delta\boldsymbol{\beta}}$$

$$E\left(\frac{\delta^2(-\log(L(\boldsymbol{\Psi},\boldsymbol{y})))}{\delta\boldsymbol{\beta}\delta\lambda_k}\right) \cong 0$$

$$E\left(\frac{\delta^2(-\log(L(\boldsymbol{\Psi},\boldsymbol{y})))}{\delta\lambda_k\delta\lambda_j}\right) \cong \frac{1}{2}tr\left(\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\lambda_j}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\lambda_k}\right)$$

These expressions were derived with the assumption that the fixed effects do not affect the predicted variance, and the random effects do not affect the predicted mean. Later, the assumptions have been relaxed to allow for fixed effects to affect the predicted variance [12]. The only change is that the expression

$$tr\left(\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\lambda_j}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\lambda_k}\right)$$

becomes more general

$$tr\left(\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\right)$$

and enters also the expressions where fixed-effects are present.

We show that the differentiation of the aggregate data log-likelihood function directly results in the expected FIM expressions [12]. We investigate the most general case where fixed effects can affect both predicted means and variances, and also random effects can affect predicted means and variances.

We go through each of the terms of the aggregate data log-likelihood expression (expression 14). We note that a change in parameters can affect the predicted $\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{y}}$ but not the observed $\boldsymbol{V}$ and $\bar{\boldsymbol{y}}$, and we note that $\boldsymbol{V}\tilde{\boldsymbol{V}}^{-1} = \boldsymbol{I}$. We further note that $\tilde{\boldsymbol{V}}$ is symmetric, $\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{V}}^T$ and that the matrix multiplication order can be rearranged within the trace operator. Starting with the first term,

$$\begin{aligned}\frac{\delta^2 tr\left(\boldsymbol{V}\tilde{\boldsymbol{V}}^{-1}\right)}{\delta\psi_k\delta\psi_j} &= -\frac{\delta}{\delta\psi_k}tr\left(\boldsymbol{V}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1}\right)\\ &= -tr\left(\boldsymbol{V}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1} + \boldsymbol{V}\tilde{\boldsymbol{V}}^{-1}\frac{\delta^2\tilde{\boldsymbol{V}}}{\delta\psi_j\delta\psi_k}\tilde{\boldsymbol{V}}^{-1} - \boldsymbol{V}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\right)\\ &= tr\left(2\boldsymbol{V}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1} - \boldsymbol{V}\tilde{\boldsymbol{V}}^{-1}\frac{\delta^2\tilde{\boldsymbol{V}}}{\delta\psi_j\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\right)\\ &= tr\left(2\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1} - \frac{\delta^2\tilde{\boldsymbol{V}}}{\delta\psi_j\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\right)\end{aligned}$$

Next, we note that

$$\frac{\delta(\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}})}{\delta\psi_j} = -\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_j}$$

Therefore

$$\frac{\delta^2\left((\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}})^T\tilde{\boldsymbol{V}}^{-1}(\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}})\right)}{\delta\psi_j\delta\psi_k}$$

$$=\frac{\delta\left(-\left(\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_j}\right)^T\ \tilde{\boldsymbol{V}}^{-1}(\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}})+(\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}})^T\ \tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1}(\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}})-(\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}})^T\ \tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_j}\right)}{\delta\psi_k}$$

$$=2\left(\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_j}\right)^T\ \tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_k}-2(\bar{y}-\tilde{\boldsymbol{y}})^T\ \tilde{\boldsymbol{V}}^{-1}\frac{\delta^2\tilde{\boldsymbol{y}}}{\delta\psi_k\psi_j}$$

$$=2\left(\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_j}\right)^T\ \tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_k}$$

Because of terms including $(\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}})=0$ canceling out. Finally,

$$\frac{\delta^2\left(\log|\tilde{\boldsymbol{V}}|\right)}{\delta\psi_j\delta\psi_k}=\frac{\delta tr\left(\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\right)}{\delta\psi_j}$$

$$=tr\left(\tilde{\boldsymbol{V}}^{-1}\frac{\delta^2\tilde{\boldsymbol{V}}}{\delta\psi_k\delta\psi_j}-\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\right)$$

Putting all of the parts together,

$$E\left(\frac{\delta^2(-\log(L(\boldsymbol{\Psi},\boldsymbol{y})))}{\delta\psi_j\delta\psi_k}\right)$$

$$\cong-\left(-\frac{1}{2}\right)\left(tr\left(2\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1}-\frac{\delta^2\tilde{\boldsymbol{V}}}{\delta\psi_j\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\right)\right.$$

$$\left.+2\left(\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_j}\right)^T\ \tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_k}+tr\left(\tilde{\boldsymbol{V}}^{-1}\frac{\delta^2\tilde{\boldsymbol{V}}}{\delta\psi_k\delta\psi_j}-\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\right)\right)$$

$$=\frac{1}{2}\left(tr\left(2\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1}\right)-tr\left(\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1}\right)\right.$$

$$\left.-tr\left(\frac{\delta^2\tilde{\boldsymbol{V}}}{\delta\psi_j\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\right)+tr\left(\frac{\delta^2\tilde{\boldsymbol{V}}}{\delta\psi_j\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\right)+2\left(\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_j}\right)^T\ \tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_k}\right)$$

$$=\frac{1}{2}tr\left(\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_k}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\psi_j}\tilde{\boldsymbol{V}}^{-1}\right)+\left(\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_j}\right)^T\ \tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{y}}}{\delta\psi_k}$$

Finally, with these identities we can conclude that under the assumptions of FO approximation, in which the random effects are not able to affect mean predictions,

$$E\left(\frac{\delta^2(-\log(L(\boldsymbol{\Psi},\boldsymbol{y})))}{\delta\beta_j\delta\beta_k}\right)\cong\left(\frac{\delta\tilde{\boldsymbol{y}}}{\delta\beta_j}\right)^T\ \tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{y}}}{\delta\beta_k}$$

$$+\frac{1}{2}tr\left(\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\beta_j}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\beta_k}\right)$$

$$E\left(\frac{\delta^2(-\log(L(\boldsymbol{\Psi},\boldsymbol{y})))}{\delta\beta_j\delta\lambda_k}\right)\cong\frac{1}{2}tr\left(\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\beta_j}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\lambda_k}\right)$$

$$E\left(\frac{\delta^2(-\log(L(\boldsymbol{\Psi},\boldsymbol{y})))}{\delta\lambda_j\delta\lambda_k}\right)\cong\frac{1}{2}tr\left(\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\lambda_j}\tilde{\boldsymbol{V}}^{-1}\frac{\delta\tilde{\boldsymbol{V}}}{\delta\lambda_k}\right)$$

Which is identical to the results derived by Retout et al. [12]. To summarize, in this Appendix we have derived the previously published expressions for expected population FIM using only the expected aggregate data. This approach has pedagogical value because it may be more intuitive for some people than deriving the expected population FIM from individual-data log-likelihood expressions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10928-021-09760-1.

Acknowledgements No financial support was received. The views expressed here are those of the author and do not necessarily reflect the position of Finnish Medicines Agency. The author would like to thank J G Coen van Hasselt and the two anonymous reviewers for their helpful comments on a draft of the manuscript.

Funding Open access funding provided by University of Eastern Finland (UEF) including Kuopio University Hospital.

References

1. Sheiner LB, Rosenberg B, Marathe VV (1977) Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. J Pharmacokinet Biopharm 5:445–479
2. Sheiner LB, Beal SL (1980) Evaluation of methods for estimating population pharmacokinetics parameters. I. Michaelis-Menten model: routine clinical pharmacokinetic data. J Pharmacokinet Biopharm 8:553–571
3. Sheiner BL, Beal SL (1981) Evaluation of methods for estimating population pharmacokinetic parameters. II. Biexponential model and experimental pharmacokinetic data. J Pharmacokinet Biopharm 9:635–651
4. Sheiner LB, Beal SL (1983) Evaluation of methods for estimating population pharmacokinetic parameters. III. Monoexponential model: routine clinical pharmacokinetic data. J Pharmacokinet Biopharm 11:303–319
5. Weber S, Gelman A, Lee D et al (2018) Bayesian aggregation of average data: an application in drug development. Ann Appl Stat 12:1583–1604. https://doi.org/10.1214/17-AOAS1122
6. Mentre F, Mallet A, Baccar D (1997) Optimal design in random-effects regression models. Biometrika 84:429–442
7. Bauer RJ, Guzy S (2004) Monte carlo parametric expectation maximization (MC-PEM) method for analyzing population pharmacokinetic/pharmacodynamic data. In: D'Argenio DZ (ed) Advanced methods of pharmacokinetic and pharmacodynamic systems analysis, vol 3. Springer, US, pp 135–163
8. Delyon B, Lavielle M, Moulines E (1999) Convergence of a stochastic approximation version of the EM algorithm. Ann Stat 27:94–128

9. Kuhn E, Lavielle M (2005) Maximum likelihood estimation in nonlinear mixed effects models. Comput Stat Data Anal 49:1020–1038. https://doi.org/10.1016/j.csda.2004.07.002
10. Lavielle M (2014) Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b17203
11. Jöreskog KG (1970) A general method for estimating a linear structural equation system*. ETS Res Bull Ser 1970:i–41. https://doi.org/10.1002/j.2333-8504.1970.tb00783.x
12. Retout S, Mentré F (2003) Further developments of the Fisher information matrix in nonlinear mixed effects models with evaluation in population pharmacokinetics. J Biopharm Stat 13:209–227. https://doi.org/10.1081/BIP-120019267
13. Wang Y (2007) Derivation of various NONMEM estimation methods. J Pharmacokinet Pharmacodyn 34:575–593. https://doi.org/10.1007/s10928-007-9060-6
14. Riviere M-K, Ueckert S, Mentré F (2016) An MCMC method for the evaluation of the Fisher information matrix for non-linear mixed effect models. Biostat Oxf Engl 17:737–750. https://doi.org/10.1093/biostatistics/kxw020
15. Niederreiter H (1988) Low-discrepancy and low-dispersion sequences. J Number Theory 30:51–70. https://doi.org/10.1016/0022-314X(88)90025-X
16. Fidler M, Wilkins JJ, Hooijmaijers R et al (2019) Nonlinear mixed-effects model development and simulation using nlmixr and related R open-source packages. CPT Pharmacomet Syst Pharmacol 8:621–633. https://doi.org/10.1002/psp4.12445
17. Foracchia M, Hooker A, Vicini P, Ruggeri A (2004) POPED, a software for optimal experiment design in population kinetics. Comput Methods Programs Biomed 74:29–46. https://doi.org/10.1016/S0169-2607(03)00073-7
18. Nyberg J, Bazzoli C, Ogungbenro K et al (2015) Methods and software tools for design evaluation in population pharmacokinetics-pharmacodynamics studies. Br J Clin Pharmacol 79:6–17. https://doi.org/10.1111/bcp.12352
19. Chenel M, Bouzom F, Aarons L, Ogungbenro K (2008) Drug-drug interaction predictions with PBPK models and optimal multiresponse sampling time designs: application to midazolam and a phase I compound. Part 1: comparison of uniresponse and multiresponse designs using PopDes. J Pharmacokinet Pharmacodyn 35:635–659. https://doi.org/10.1007/s10928-008-9104-6
20. Dumont C, Mentré F, Gaynor C et al (2013) Optimal sampling times for a drug and its metabolite using SIMCYP(®) simulations as prior information. Clin Pharmacokinet 52:43–57. https://doi.org/10.1007/s40262-012-0022-9
21. Thai H-T, Mazuir F, Cartot-Cotton S, Veyrat-Follet C (2015) Optimizing pharmacokinetic bridging studies in paediatric oncology using physiologically-based pharmacokinetic modelling: application to docetaxel. Br J Clin Pharmacol 80:534–547. https://doi.org/10.1111/bcp.12702
22. Hooker AC, Foracchia M, Dodds MG, Vicini P (2003) An evaluation of population D-optimal designs via pharmacokinetic simulations. Ann Biomed Eng 31:98–111
23. Retout S, Duffull S, Mentré F (2001) Development and implementation of the population Fisher information matrix for the evaluation of population pharmacokinetic designs. Comput Methods Programs Biomed 65:141–151. https://doi.org/10.1016/s0169-2607(00)00117-6

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.