UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy 223

Improved Methods for Pharmacometric Model-Based Decision-Making in Clinical Drug Development

ANNE-GAËLLE DOSNE

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2016

ISSN 1651-6192
ISBN 978-91-554-9734-7
urn:nbn:se:uu:diva-305697

Dissertation presented at Uppsala University to be publicly examined in B/A1:107a, Biomedicinskt Centrum, Husargatan 3, Uppsala, Friday, 9 December 2016 at 09:15 for the degree of Doctor of Philosophy (Faculty of Pharmacy). The examination will be conducted in English. Faculty examiner: PhD Rik Schoemaker (Occams).

Abstract

Dosne, A.-G. 2016. Improved Methods for Pharmacometric Model-Based Decision-Making in Clinical Drug Development. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy* 223. 91 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-9734-7.

Pharmacometric model-based analysis using nonlinear mixed-effects models (NLMEM) has to date mainly been applied to learning activities in drug development. However, such analyses can also serve as the primary analysis in confirmatory studies, which is expected to bring higher power than traditional analysis methods, among other advantages. Because of the high expertise in designing and interpreting confirmatory studies with other types of analyses and because of a number of unresolved uncertainties regarding the magnitude of potential gains and risks, pharmacometric analyses are traditionally not used as primary analysis in confirmatory trials.

The aim of this thesis was to address current hurdles hampering the use of pharmacometric model-based analysis in confirmatory settings by developing strategies to increase model compliance to distributional assumptions regarding the residual error, to improve the quantification of parameter uncertainty and to enable model prespecification.

A dynamic transform-both-sides approach capable of handling skewed and/or heteroscedastic residuals and a t-distribution approach allowing for symmetric heavy tails were developed and proved relevant tools to increase model compliance to distributional assumptions regarding the residual error. A diagnostic capable of assessing the appropriateness of parameter uncertainty distributions was developed, showing that currently used uncertainty methods such as bootstrap have limitations for NLMEM. A method based on sampling importance resampling (SIR) was thus proposed, which could provide parameter uncertainty in many situations where other methods fail such as with small datasets, highly nonlinear models or meta-analysis. SIR was successfully applied to predict the uncertainty in human plasma concentrations for the antibiotic colistin and its prodrug colistin methanesulfonate based on an interspecies whole-body physiologically based pharmacokinetic model. Lastly, strategies based on model-averaging were proposed to enable full model prespecification and proved to be valid alternatives to standard methodologies for studies assessing the QT prolongation potential of a drug and for phase III trials in rheumatoid arthritis.

In conclusion, improved methods for handling residual error, parameter uncertainty and model uncertainty in NLMEM were successfully developed. As confirmatory trials are among the most demanding in terms of patient-participation, cost and time in drug development, allowing (some of) these trials to be analyzed with pharmacometric model-based methods will help improve the safety and efficiency of drug development.

Keywords: pharmacometrics, nonlinear mixed-effects models, confirmatory trials, residual error modeling, parameter uncertainty, sampling importance resampling, model-averaging

Anne-Gaëlle Dosne, Department of Pharmaceutical Biosciences, Box 591, Uppsala University, SE-75124 Uppsala, Sweden.

ISSN 1651-6192
ISBN 978-91-554-9734-7
urn:nbn:se:uu:diva-305697 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-305697)

Sustaining doubt is harder work than sliding into certainty.

Daniel Kahneman

To my father, the warrior

The cover artwork „*Principiis obsta!*“ (“*Resist the beginnings!”*) by Thorsten Schiffer is inspired by Jackson Pollock’s drip paintings. “It portrays the evolution and result of an idea rooted in disorder as inevitable lawlessness. The tangled lines with no apparent start or end emphasize the bootstrapping aspect of unclear concepts. Observing plots of pharmacometric models from a distance reminded me at times of the beauty of clear hypotheses and design.” (T. Schiffer)

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I. Dosne AG, Bergstrand M, Karlsson MO. (2016) A Strategy For Residual Error Modeling Incorporating Scedasticity Of Variance And Distribution Shape. *J Pharmacokinet Pharmacodyn* 43(2):137-51.

II. Dosne AG*, Niebecker R*, Karlsson MO. (2016) dOFV Distributions: A New Diagnostic For The Adequacy Of Parameter Uncertainty In Nonlinear Mixed-Effects Models Applied To The Bootstrap. *J Pharmacokinet Pharmacodyn* DOI: 10.1007/s10928-016-9487-8.

III. Dosne AG, Bergstrand M, Harling K, Karlsson MO. (2016) Improving The Estimation Of Parameter Uncertainty Distributions In Nonlinear Mixed Effects Models Using Sampling Importance Resampling. *J Pharmacokinet Pharmacodyn* DOI :10.1007/s10928-016-9496-7.

IV. Dosne AG, Bergstrand M, Karlsson MO. An Automated Sampling Importance Resampling Procedure For Estimating Parameter Uncertainty [*Submitted*]

V. Bouchene S, Dosne AG, Marchand S, Friberg LE, Björkman S, Couet W, Karlsson MO. Development Of An Interspecies Whole-Body Physiologically Based Pharmacokinetic Model For Colistin And Colistin Methanesulfonate In Five Animal Species And Evaluation Of Its Predictive Ability In Human [*In manuscript*]

VI. Dosne AG, Bergstrand M, Karlsson MO, Renard D, Heimann G. Model-Averaging For Robust Assessment Of QT Prolongation By Concentration-Response Analysis [*Submitted*]

VII. Dosne AG, Bieth B, Bergstrand M, Karlsson MO, Renard D. Longitudinal Data Analysis Using Model-Averaging: Benefits For Pivotal Clinical Trials, Applied To Rheumatoid Arthritis [*In manuscript*]

Reprints were made with permission from the respective publishers.

*The authors contributed equally to this work

Contents

Introduction 11
Decision-making in clinical drug development 11
Established role of pharmacometrics for learning activities 13
Potential role of pharmacometrics for confirming activities 14
Principle of hypothesis testing 16
Analysis models 18
Cross-sectional models 18
Longitudinal models 18
Hurdles to the use of pharmacometric models for decision-making 21
Distributional assumptions regarding the residual error 21
Distributional assumptions regarding parameter uncertainty 22
Need for model prespecification 22

Aims 23

Methods 24
Residual error modeling 24
Commonly used models 24
Proposed error models: dTBS and the t-distribution 25
Evaluation of the new error models on real data examples 27
Evaluation of the new error models on simulated examples 28
Parameter uncertainty 29
Commonly used methods 29
The dOFV diagnostic: assessing uncertainty adequacy 31
Evaluation of bootstrap adequacy in NLMEM 33
SIR: improving parameter uncertainty estimation 34
Evaluation of the 1-step SIR on simulated data 38
Evaluation of the 1-step SIR on real data 38
Evaluation of the 5-step SIR on real data 39
Application of SIR for decision-making using a WBPBPK model 40
Model prespecification 42
Principle of model-averaging 42
Model-averaged test for QT prolongation assessment 42
Model-averaged test for rheumatoid arthritis trials 45
Software 47

Results 49
Residual error modeling 49
Performance of dTBS on real data examples 49
Performance of the t-distribution on real data examples 52
Simulation results 53
Parameter uncertainty 53
Performance of the dOFV uncertainty diagnostic 53
Evaluation of bootstrap adequacy in NLMEM 53
Performance of the 1-step SIR on simulated data 56
Performance of the 1-step SIR on real data 56
Performance of the 5-step SIR on real data 58
Performance of SIR for decision-making using a WBPBPK model 60
Model prespecification 62
Model-averaged test for QT prolongation assessment 62
Model-averaged test for rheumatoid arthritis trials 64

Discussion 67
Residual error modeling 67
Assessing improvement on the error model level 68
dTBS specificities 69
t-distribution specificities 69
Parameter uncertainty 70
Performance of the dOFV uncertainty diagnostic 70
Evaluation of bootstrap adequacy in NLMEM 70
Performance of the 1-step SIR on simulated data 72
Performance of the 1-step SIR on real data 73
Performance of the 5-step SIR on real data 74
Performance of SIR for decision-making using a WBPBPK model 75
Model prespecification 76
Model-averaged test for QT prolongation assessment 76
Model-averaged test for rheumatoid arthritis trials 77

Conclusion 80

Acknowledgments 81

References 84

Abbreviations

ACR	American College of Rheumatology
AGQ	Adaptive Gaussian Quadrature
AIC	Akaike Information Criterion
AN(C)OVA	Analysis of (Co)Variance
AUC	Area Under the Curve
BIC	Bayesian Information Criterion
CI	Confidence Interval
CL	Clearance
Cmax	Maximum Concentration
CMS	Colistin Methanesulfonate
CWRES	Conditional Weighted Residuals
dOFV	Delta Objective Function Value
dTBS	dynamic Transform-Both-Sides
E0	Baseline
EC50	Concentration leading to half the maximum Effect
ED50	Dose leading to half the maximum Effect
EMA	European Medicines Agency
EMAX	Maximum Effect
FDA	Food and Drug Administration
FIM	Fisher Information Matrix
FO	First-Order (estimation)
FOCE(I)	First-Order Conditional Estimation (with Interaction)
H_0, H_1	Null and Alternative Hypotheses
IIV	Inter-Individual Variability
IOV	Inter-Occasion Variability
IWRES	Individual Weighted Residuals
KA	Absorption rate
K_p	Tissue-to-plasma partition coefficient
LAPLACE	Laplacian (estimation)
(L)L	(Log-)Likelihood
LLP	Log-Likelihood Profiling
MCP-Mod	Multiple Comparison Procedure - Modelling
MID3	Model-Informed Drug Discovery and Development
MISE	Mean Integrated Square Error

NLME(M)	Nonlinear Mixed Effects (Models)
NPDE	Normalized Prediction Distribution Error
OFV	Objective Function Value
OFVi	Individual Objective Function Value
PD	Pharmacodynamic(s)
PDF	Probability Density Function
PK	Pharmacokinetic(s)
PsN	Perl-speaks-NONMEM
QTc	corrected QT interval
(R)SE	(Relative) Standard Error(s)
RUV	Residual Unexplained Variability
(SA)EM	(Stochastic Approximation) Expectation Maximization
SIR	Sampling Importance Resampling
SSE	Stochastic Simulations and Estimations
TLAG	Lag-Time
TQT	Thorough-QT
V	Volume
VPC	Visual Predictive Check
WBPBPK	Whole-Body Physiologically Based Pharmacokinetic

Introduction

Decision-making in clinical drug development

Drug development can be defined as the process of finding a dose regimen at which a candidate drug is safe and effective at treating subjects suffering from a given condition. Drug development is typically divided into preclinical and clinical development. Preclinical development aims at determining the efficacy and safety of the candidate drug based on *in vitro* assays and *in vivo* experiments in various animal species. Clinical development pursues the same goal based on studies performed in human, healthy volunteers or patients. Clinical development is further subdivided into three more or less consecutive phases. Phase I corresponds to the first time the candidate drug is administrated to humans, typically healthy volunteers, and mainly focuses on safety and tolerability. Phase II involves a limited number of patients and aims at establishing the "proof-of-concept", i.e. a first indication of efficacy, as well as at selecting a suitable dosing regimen. Phase III involves a high number of patients and constitutes the pivotal confirmation of a positive benefit/risk ratio, which will lead if successful to an application for market authorization to health authorities. Figure 1 presents a schematic of clinical drug development and the associated decision-making processes at the different milestones.

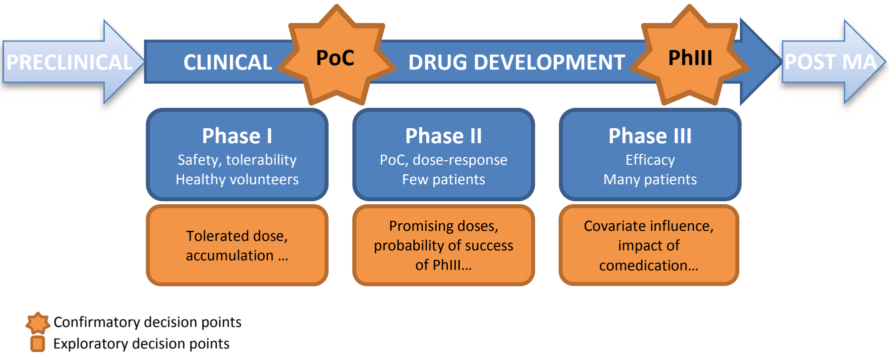
PRECLINICAL
CLINICAL
PoC
DRUG DEVELOPMENT
PhIII
POST MA
Phase I
Safety, tolerability
Healthy volunteers
Phase II
PoC, dose-response
Few patients
Phase III
Efficacy
Many patients
Tolerated dose, accumulation ...
Promising doses, probability of success of PhIII...
Covariate influence, impact of comedication...
Confirmatory decision points
Exploratory decision points

Figure 1. Schematic of clinical drug development and associated decision-making processes. PoC: proof-of-concept, PhIII: Phase III, MA: marketing authorization.

Confirmatory decisions are made at two main points during clinical drug development. These decisions are taken by different stakeholders and bear different consequences. The first confirmatory decision point is the Phase II proof-of-concept trial, which is a major driver of the sponsor's decision of whether or not to pursue the compound's development ("go/no go"). The second confirmatory decision point is the Phase III efficacy trial, which is a key element when regulatory bodies decide whether or not to approve the drug. *Exploratory* decision-making also happens at other points along clinical drug development, for example when selecting which dose regimens to move forward or which drug formulation to use. Clinical trials may differ in a number of ways depending on whether they are confirmatory or exploratory. Confirmatory trials provide firm evidence of efficacy or safety by testing predefined hypotheses with predefined methods[1]. Examples of confirmatory trials are the Phase III efficacy studies. Exploratory trials on the other hand have specific objectives which should increase the knowledge on the compound, but do not necessarily require strict hypothesis testing. Examples of exploratory trials are the Phase I single ascending dose studies. Note that a trial may have both confirmatory and exploratory aspects. This is for example often the case for Phase II dose finding studies, where both the presence of a dose-response is confirmed and a suitable target dose for Phase II is explored.

The concept of confirmatory and exploratory trials is mirrored in the *learning versus confirming* paradigm[2] established by Lewis Sheiner in 1997. This paradigm lays out differences in the aim, design and analysis of studies designed for learning and those of studies designed for confirming. A summary of the main distinctions is provided in Table 1. Confirmatory settings are usually characterized by a single question to be answered by "yes" or "no", typically within a single study referred to as confirmatory or pivotal trial. Examples of such questions are: does the drug show efficacy in selected patients? Does the drug demonstrate an acceptable benefit/risk ratio in a large patient population? Is the new formulation equivalent to the old formulation? Is dose adjustment needed in some populations? The analysis of confirmatory trials is based on hypothesis testing. Answering the given question with high certainty, in particular making sure that the risk of false positives is controlled, is considered of major importance. Learning settings can address multiple questions which are answered by quantitative metrics, possibly resulting from pooled data. Examples of such questions are: what is the tolerated dose in healthy volunteers? What is the dose-response in selected patients and which doses should be carried on to the next phase? What is the probability of success of the next trial? How should the dose be adapted for particular populations? Analysis tools in learning settings are aimed at quantifying one or more metrics of interest which can then be used to generate hypotheses. With the specifics of confirming and learning activities now outlined, the question arises of which analysis tools to use for which activity.

Table 1. Differences between learning and confirming activities

Characteristic	Learning	Confirming
Aim	Hypothesis generation	Hypothesis testing
Type of answer	Quantitative Multiple endpoints	Binary yes/no Single endpoint
Evidenced used	Pooled studies Prior information	Single study (possibly replicated)
Study design	More flexible Can be unbalanced	Highly controlled Balanced
Analysis method	Principles prespecified, but can be data driven	All details prespecified
Aversion to risk	Low/moderate	High

Established role of pharmacometrics for learning activities

Pharmacometrics, which is a key discipline of the Model-Informed Drug Discovery and Development (MID3) framework[3], has been traditionally applied for learning activities. MID3 has been introduced recently as a "quantitative framework for prediction and extrapolation, centered on knowledge and inference generated from integrated models of compound, mechanism and disease level data and aimed at improving the quality, efficiency and cost effectiveness of decision making". It followed up on the previously used term of model-based drug development. Model-based approaches have long been recognized as relevant tools to increase the productivity and sustainability of drug development[4-6]. Pharmacometrics itself has been defined as "the science of developing and applying mathematical and statistical methods to (a) characterize, understand and predict a drug's pharmacokinetic and pharmacodynamic behavior; (b) quantify uncertainty of information about that behavior; and (c) rationalize data-driven decision making in the drug development process and pharmacotherapy"[7].

Pharmacometrics is used for an array of learning activities. For example, modeling clinical trial data has made a major impact on dose selection and optimization through the establishment of dose-response or exposure-response relationships. Pharmacometrics has also greatly impacted the design of clinical trials through the use of clinical trial simulations[8], where models for placebo responses, disease progression, drug responses, and trial execution are developed and integrated in order to predict endpoints such as future trial outcomes[9]. Model-based meta-analysis also proved a valuable tool for benchmarking new compounds in their competitive landscape and assess how likely it is that they outperform readily available treatments[10]. With the capacity of summarizing, integrating and storing knowledge acquired during the course of drug development, pharmacometrics has contributed to improve the quality and the importance of learning phases. While pharmacometric models should continue to be used for learning purposes, their application to confirmatory analyses deserves further attention.

Potential role of pharmacometrics for confirming activities

The use of pharmacometric models based on longitudinal data analysis has so far been limited for hypothesis testing in confirmatory settings despite the recognition of its scientific merit by regulatory authorities[11]. The qualification by the European Medicines Agency (EMA) of Multiple Comparison Procedure - Modelling (MCP-Mod)[12] as an efficient statistical methodology for the design and analysis of Phase II dose finding studies is a first step towards the use of dose-response models in confirmatory settings. Note that this methodology combines both confirmatory and exploratory elements, but is not based on longitudinal data.

More generally, a review of 198 submissions to the Food and Drug Administration (FDA) over the years 2000 to 2008[13] showed that 64% and 67% of pharmacometric reviews contributed to drug approval and labelling decisions, with about half of them providing pivotal or supportive insights into effectiveness and safety. However, model-based primary endpoints were only used in 2.5% of cases, and all of them concerned pediatrics settings. Only 4.5% of the reviews used model-based analyses to confirm effectiveness, and in a presented case study the performed exposure-response modeling was not used for formal hypothesis testing. This setting is actually a particularity of drug approvals by the FDA, which more often than its European counterpart demands evidence from two pivotal Phase III trials[14]. The benefit of pharmacometrics to obviate the need for a second trial, without necessarily being used as a primary analysis, has been recognized previously[15].

The shift to longitudinal model-based analysis for the purpose of confirmation is expected to bring a number of advantages. Pharmacometric model-based analysis is likely to achieve higher power than traditional analysis, thus enabling to reduce the number of patients needed in clinical trials. This is to relate to the fact that modeling exploits the complete longitudinal data instead of a cross-sectional fraction of it, thus increasing the signal to noise ratio. The power gain could be substantial in many therapeutic areas and lead to much smaller studies. For example, a two-arm proof-of-concept Phase II analysis in the stroke indication would require a total of 90 patients for 80% power to detect a drug effect different from 0 using a pharmacometric model-based analysis versus 388 patients using a two-sided t-test, resulting in a 4.3-fold difference in study size[16]. Power gains could be even higher when model-based endpoints associated with low inter-individual variability and precise quantification assays can be used[17]. Furthermore, inferential assessment of endpoints such as disease progression, which are expected to be increasingly used in various therapeutic areas, will require longitudinal approaches[11]. Overall, analyzing smaller studies with pharmacometric model-based approaches would achieve satisfactory power and thus decrease the number of patients exposed to an experimental drug as well as reallocate savings in both costs and time to the development pipeline.

Pharmacometric model-based confirmatory analyses are also particularly useful when traditional analyses are not feasible due to practical limitations in sample size (or more generally in available information) or due to constraints in study design. Limitations in sample size are often observed in small population groups such as pediatrics[18] or personalized medicine settings[19], for which traditional methods would often lead to inconclusive studies due to lack of power. Development of new methods including pharmacometric model-based approaches for the design and analysis of trials in small population groups is currently the focus of a European Consortium[20]. Traditional analysis can also be difficult when the amount of collectable data is limited, rendering the metric of interest not assessable at the individual level or within a single study. This has been observed for example when assessing whether covariates significantly influence the pharmacokinetic (PK) profile of a drug, summarized by the area under the curve (AUC) and the maximum concentration (Cmax), in order to detect a potential need for dose adjustment[21]. Detection of influential covariates is best done in Phase III studies, as they provide a high enough power and represent a large sample of the target population. Sampling of drug concentrations during such studies is typically sparse and does not enable to calculate individual AUC or Cmax using traditional non-compartmental analysis. Alternative pharmacometric methods such as compartmental population PK analysis can then be used to estimate individual AUC or Cmax. Another example where traditional analysis methods have limitations is therapeutic protein-drug interaction studies, which could benefit from pooling relevant data from multiple studies. Dedicated studies may be logistically cumbersome for such compounds, as they typically need to be performed in the patient population due to disease-specific PK profiles and require parallel designs due to long terminal half-lives[22]. Population PK could also address issues faced by therapeutic protein-drug interactions.

Lastly, particularities in study design can also present hurdles for traditional analyses. For example, adaptive designs, for which study design can be modified and improved in a pre-planned manner at interim time points during the study, and seamless designs, which combine in a single study goals that are typically addressed in different trials, have been advocated to enhance clinical development[23]. Dealing with such flexible designs is not always straightforward with traditional analyses. As such designs are likely to be more commonplace in the future, pharmacometric model-based analysis methods would be a natural match for this evolution.

Principle of hypothesis testing

To understand the advantages and drawbacks of pharmacometric models in confirmatory settings, the principles of hypothesis testing need to be specified. A hypothesis test can be defined as a decision rule which uses statistical models and observed data to decide which of usually two mutually exclusive hypotheses is true for a population, based on a sample. In the clinical context the population typically corresponds to all patients to be treated, and the sample to all subjects included in the clinical trial. Hypothesis testing can be summarized in four steps:

1. Define the hypotheses to test H_0 (the null hypothesis) and H_1 (the alternative hypothesis). For example, in the case of testing if a difference exists between the new treatment and the standard of care, H_0 and H_1 would be expressed as:

   $$H_0: \psi(\Theta) \; = 0 \qquad \text{Eq. 1}$$
   $$H_1: \psi(\Theta) \; \neq 0$$

   where $\psi(\Theta)$ is the difference between the two treatments expressed as a function of model parameters Θ. H_0 is generally set as the hypothesis one wants to reject.

2. Identify a test statistic which distribution is known if H_0 is true. For example, in the previous case of testing a difference between two treatments, the Student statistic T_{sample} (Eq. 2) is known to follow a Student's t-distribution with n - 1 degrees of freedom if $\psi(\Theta)$ is normally distributed.

   $$T_{sample} \; = \frac{\hat{\psi}(\hat{\Theta})}{\hat{S}} \sqrt{n-1} \qquad \text{Eq. 2}$$

   where $\hat{\psi}(\hat{\Theta})$ is the estimate of $\psi(\Theta)$ in the sample, $\hat{S}$ its variance and n the sample size. Note that the test statistic can also be the endpoint $\psi(\Theta)$ itself.

3. Calculate the test statistic for the data at hand and derive the probability, also called p-value, that a test statistic greater or equal to the one observed would be obtained if H_0 were true. In our example the p-value would equal:

   $$p_{sample} = \int_{-\infty}^{T_{sample}} PDF(T|H_0)dT \qquad \text{Eq. 3}$$

   where $PDF(T_{sample}|H_0)$ is the probability density function (PDF) of the test statistic if H_0 is true, in this case the PDF of the Student's t-distribution. If the test statistic is the endpoint itself, its confidence interval (CI) at a predefined significance level α is computed.

4. Compare the p-value to a predefined significance level α: if it is lower, H_0 is rejected. In our example, one would test whether:

$$p_{sample} \le \alpha = \int_{-\infty}^{T_{critic}} PDF(T|H_0)dT \qquad \text{Eq. 4}$$

where T_{critic} is the cut-off value defined by the significance level α in the distribution of T. When the test statistic is the endpoint itself, H_0 would be rejected if the CI at the α significance level contains the value defining the hypotheses (e.g. 0 in Eq. 1).

A summary of the hypothesis testing process is provided in Figure 2.

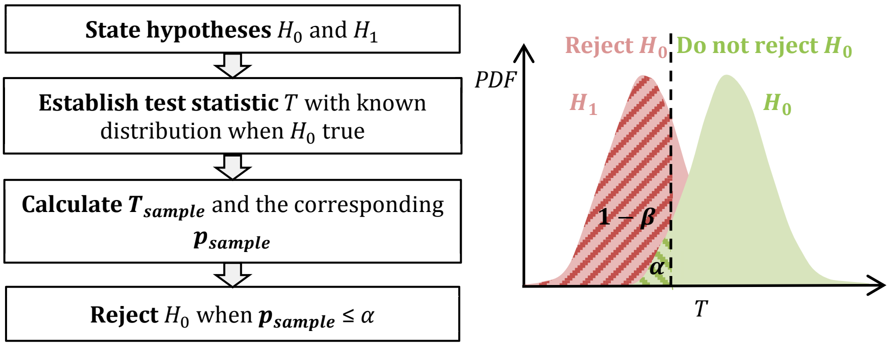
State hypotheses H_0 and H_1
Establish test statistic T with known distribution when H_0 true
Calculate T_{sample} and the corresponding p_{sample}
Reject H_0 when $p_{sample} \le \alpha$
Reject H_0
Do not reject H_0
PDF
H_1
H_0
$1-\beta$
α
T

Figure 2. The principle of hypothesis testing. α and β are the type I and II errors which will be defined below.

Decisions made based on a hypothesis test can be correct or incorrect. The decision will be correct in two cases: if H_0 is true and H_0 is not rejected; and if H_1 is true and H_0 is rejected. Oppositely, the decision will be incorrect if H_0 is true and H_0 is rejected; and if H_1 is true and H_0 is not rejected. The risks associated with making wrong decisions are called type I and type II errors depending on which hypothesis is true (Table 2). These risks need to be managed for a hypothesis test to be accepted as a means of drawing inference, i.e. making a conclusion, with regards to the efficacy or safety of a new drug. The type I error α is the error of major concern for regulatory agencies, as it corresponds to "worst case scenarios" such as falsely declaring a new compound superior to the standard of care, or falsely declaring a generic compound equivalent to the brand name drug. The type I error is equal to the significance level during hypothesis testing, and is mitigated by being set to a small value (e.g. 0.05, meaning that superiority will falsely be concluded at maximum in 5% of cases). Note that the type I risk is only controlled if the assumptions underlying the model used for the test are valid. The type II error β is a concern of the sponsor, and is more commonly used as 1- β, which is referred to as power. To mitigate type II error, the power is set to a high value (e.g. 0.80, meaning that H_0 will be correctly rejected in 80% of cases for a specific H_1). The power conditions the number of patients who need to be recruited for a trial.

Table 2. Correct and incorrect inferences based on hypothesis testing

	Reject H_0	**Do not reject H_0**
H_0 true	Incorrect decision *Type I error α*	Correct decision
H_1 true	Correct decision *Power 1- β*	Incorrect decision *Type II error β*

Analysis models

Cross-sectional models

Cross-sectional models, which typically describe the distribution of an average endpoint at a specific time point, have been mostly used so far in confirmatory settings. They provide good type I error control but moderate power, which can be linked to the fact that they make few assumptions about the endpoint and use only part of the trial data. An example of a cross-sectional model for a given endpoint y at a chosen time would be a normal distribution N with mean μ and standard deviation σ.

$$\begin{aligned} y &= \mu + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2) \end{aligned} \qquad \text{Eq. 5}$$

Such a model can for example be used for the endpoint of the mean decrease in blood pressure in a population at the end of a clinical trial. The main assumption in cross-sectional models typically pertains the distribution of the endpoint at a particular time point.

Model parameters can be directly estimated from the data using the formulas for means and variances, or using least squares estimation methods if regression elements are included in the model.

A test statistic can then be derived from the endpoint model and subsequently be used for hypothesis testing. The distributions of test statistics are often well known under certain assumptions such as independence between measurements and large enough sample sizes. Examples of hypothesis tests based on cross-sectional models include Student’s t-test, which tests whether the means of two sets of data are significantly different from each other, analysis of variance (ANOVA), which is an extension of the t-test to more than two groups, and analysis of covariance (ANCOVA), which enables to add linear regression components such as covariates into ANOVA.

Longitudinal models

Longitudinal mixed effects models, also referred to as pharmacometric models herein, typically describe the distribution of individual endpoints over the full time course of the study, and could improve the power of hypothesis tests by utilizing additional study information. However, the increased number of assumptions raises regulatory concerns towards a lack of type I control. An example of a longitudinal model for a given endpoint would be a function of one or more explanatory variables such as time. A common family of longitudinal models is nonlinear mixed effect models (NLMEM). For continuous data, such a model can be expressed as:

$$
\begin{aligned}
&y_{ij} = f\left(t_{ij}, \theta_i, z_{ij}\right) + g\left(f\left(t_{ij}, \theta_i, z_{ij}\right), \varepsilon_{ij}\right) \\
&\theta_i = h(\theta, \eta_i) \\
&\eta_i \sim N(0, \Omega) \\
&\varepsilon_{ij} \sim N(0, \Sigma)
\end{aligned}
\qquad \text{Eq. 6}
$$

where y_{ij} is the observation of individual i at time j and $f(\cdot)$ is the structural model depending on time t_{ij}, on the vector of individual parameters θ_i and on the vector of individual covariates z_{ij}. $g(\cdot)$ is the residual error function with residuals ε_{ij}. Individual parameters are obtained from the population fixed effects parameters θ and individual random effects η_i via a function $h(\cdot)$ which is typically additive or exponential. The η_i and ε_{ij} are assumed normally distributed with mean 0 and variance-covariance matrices Ω and Σ, respectively. They are independent between individuals as well as between each other. For discrete data, the model can be expressed as:

$$p(y_{ij} = \mathrm{x}) = l\left(t_{ij}, \theta_i, z_{ij}\right) \qquad \text{Eq. 7}$$

where p is the probability of observing x given the probability density function $l(\cdot)$. Assumptions in longitudinal models pertain the structural model and the distributions of the random effects and the residual variability. The validity of these assumptions can be checked by inspecting distribution plots of random effects and residuals, simulation-based diagnostics such as Visual Predictive Checks (VPC), and other metrics.

Model parameters $\Theta = (\theta, \Omega, \Sigma)$ are estimated using maximum likelihood, which aims at finding the parameter values which maximize the likelihood of observing the sample data given the parameters. In NLMEM, the likelihood of the full set of data is equal to the product of the likelihoods of the data of each individual. For ease of computation, the individual likelihood can be obtained on the log scale by integrating over the random effects:

$$\mathrm{LL}_i(\mathrm{y}_i|\Theta) = \log \int_{-\infty}^{\infty} p(\mathrm{y}_i, \eta_i; \Theta) d\,\eta_i = \log \int_{-\infty}^{\infty} u(\mathrm{y}_i|\eta_i; \Theta) v(\eta_i; \Theta) d\,\eta_i \qquad \text{Eq. 8}$$

where $\mathrm{LL}_i(\mathrm{y}_i|\Theta)$ is the log-likelihood of the observed data from one individual, $p(\mathrm{y}_i, \eta_i; \Theta)$ is the likelihood of the complete data (y_i, η_i) of subject i, $u(\mathrm{y}_i|\eta_i; \Theta)$ is the conditional density of the observations given the individual random effects, and $v(\eta_i; \Theta)$ is the density of the individual random effects. $u(\cdot)$ is generally equal to the probability density function of the normal distribution for continuous data, and to the probability density function defined by $l(\cdot)$ for discrete data. Parameter estimation is performed by minimizing the objective function value (OFV), which corresponds to minus two times the log-likelihood of all data up to a constant. The calculation of individual likelihoods is not trivial and often no analytical solution exists for Eq. 8.

Numerical approximations are therefore used, which can be divided into gradient-based algorithms and expectation-maximization (EM) algorithms. Gradient-based algorithms use the derivative of the approximation of the log-likelihood (LL) to guide parameter search. They include the first-order (FO), first-order conditional estimation (with interaction) FOCE(I), Laplacian (LAPLACE) and Adaptive Gaussian Quadrature (AGQ) algorithms, which differ in the number of quadrature points used to approximate the integral (freely chosen for AGQ, 1 for the others), the order of the approximation (second for AGQ and LAPLACE, first for FOCE(I) and FO) and the location of the approximation ($\eta_i = 0$ for FO, $\eta_i = \hat{\eta}_i$ for the others). EM algorithms are based on the alternation of a step estimating the conditional mean parameters for each individual and a step maximizing the likelihood of the full individual data with respect to the population parameters. EM algorithms include the iterative two-stage, importance sampling and stochastic approximation expectation-maximization (SAEM).

Hypothesis tests are typically carried out on model parameters based on the log-likelihood ratio or the Wald statistic. For example, one can test whether the drug effect parameter is different from 0 by calculating the statistic:

$$T_{sample} = \mathrm{LL}_i\left(\mathrm{y}_i \middle| \hat{\Theta}\right) - \mathrm{LL}_i\left(\mathrm{y}_i \middle| \hat{\Theta}_0\right) \qquad \text{Eq. 9}$$

where $\hat{\Theta}_0$ is the estimated vector of model parameters when the drug effect is fixed to 0. This statistic, also referred to as the delta Objective Function Value (dOFV), follows a chi-square distribution with a degree of freedom equal to the difference in the number of estimated parameters between $\hat{\Theta}$ and $\hat{\Theta}_0$. The Wald statistic is based on the estimate of the parameter uncertainty obtained from the Fisher Information Matrix (FIM) and also follows a chi-square distribution under H_0. These tests are asymptotically equivalent. Note that hypothesis testing can also be carried out on functions of model parameters (e.g. endpoint at the end of study), and thus be used for the same type of tests than cross-sectional models. A comparative summary of cross-sectional and longitudinal models is provided in Table 3.

Table 3. Differences between cross-sectional and pharmacometric models and tests

Characteristic	**Cross-sectional model**	**Longitudinal (pharmacometric) model**
Endpoint	Clinical endpoint (observed) at specific time point	Model parameter(s) (often unobserved), sometimes function of model parameters
Assumptions	Endpoint and test statistic distributions	Structural longitudinal model, random effects and test statistic distributions
Parameter estimation	Least squares	Maximum likelihood
Example test statistics	ANCOVA, endpoint	likelihood ratio, Wald, endpoint

Hurdles to the use of pharmacometric models for decision-making

There are a number of reasons why pharmacometric models are not used as primary analyses for decision-making despite the potential gains they may entail. A first reason might be purely organizational: today both drug developers and regulators are experienced in designing and interpreting confirmatory trials based on traditional analyses and there is an abundance of trained personnel qualified to do this[24]. Another reason is the lack of identification of the potential benefits and risks of such a shift. Identification of which situations are likely (or not) to gain from pharmacometric model-based analysis lacks systematic investigation. Quantification of the benefits through direct comparison of traditional versus pharmacometric model-based analysis has been even rarer, which is partly due to the current use of modeling mainly in cases where traditional analysis is not possible[22,25]. Quantification of the risks expressed in the type I and type II errors often stays unanswered, which leads to a certain discomfort regarding the use of such methods. In addition, the calculation of appropriate sample sizes may have been seen as a drawback of pharmacometric methods, as it requires computer-intensive Monte Carlo simulations except in simple cases[26]. However, recently developed strategies[16,27] have made model-based sample size determination very accessible, which should foster the implementation of model-based analysis in confirmatory settings. The real challenge for pharmacometric model-based analysis in confirmatory settings lies in the mitigation of its inherent risks. These risks can be summarized into three categories: distributional assumptions of the model, distributional assumptions of the uncertainty of its parameters, and data-driven model-building.

Distributional assumptions regarding the residual error

In order to take correct decisions based a pharmacometric model, the assumptions made during the modeling process regarding the distributions of the random effects and the residual error need to be met. Semiparametric distribution with estimated shape parameters have already been proposed for random effects, allowing a more flexible description of their distribution and thus increasing compliance to modeling assumptions[28]. Such possibilities have been advocated for the residual error[29-31], but no framework has yet been proposed and thoroughly investigated for the residual error model. The development and evaluation of strategies able to increase compliance to residual unexplained variability (RUV) assumptions will thus be the focus of the first part of this thesis work.

Distributional assumptions regarding parameter uncertainty

Further assumptions are needed for quantifying the uncertainty around a given decision using pharmacometric analysis, for example when computing the CI around the endpoint to test. These assumptions relate to the distribution of the uncertainty of model parameters. A number of methods to estimate parameter uncertainty exist, but their performance for NLMEM is not well defined, and they might display considerable limitations. In addition, no diagnostic exists to judge their appropriateness for a given context. The focus of the second part of this thesis work will thus be the development of a diagnostic to judge the appropriateness of parameter uncertainty estimates, and the development and application of a method improving parameter uncertainty estimation for NLMEM.

Need for model prespecification

Lastly, in confirmatory settings the analysis model typically needs to be fully prespecified, i.e. all details need to be laid out in advance in the study's analysis plan. This is difficult for NLMEM, which typically undergo data-driven model-building. The multiplicity of the model building process, which can be seen as a series of testing steps leading to a final model, is not and cannot easily be accounted for when this model is used for hypothesis testing. This may lead to an unacceptable increase in type I error, which is particularly worrisome since the increase itself cannot be precisely quantified. In addition, the data-driven building process also renders the procedure subjective and non-reproducible: models can differ depending on the modeler. In order to ensure reproducibility and type I error control, fully prespecified model-based analyses have been proposed, where model building is completely avoided[25] or limited to a very limited number of steps with precisely defined selection criteria[32]. However, type I error may still not be controlled if the prespecified model happens to be misspecified. Model-averaging, which consists of analyzing the data with a set of prespecified models and weighting each model based on its fit to the data, has been proposed as a way to control type I error while guarding against model misspecification[33], but experience with this type of approach is lacking. The last part of this thesis work will propose and evaluate the performance of two model-averaged tests to be used in the context of safety and efficacy confirmatory trials.

Aims

The overall aim of this PhD thesis was to address current hurdles hampering the use of pharmacometric model-based analysis for decision-making in clinical drug development. The goal was to extend the application of pharmacometrics to key decision points such as confirmatory trials, thus enabling more efficient drug development.

The specific aims were:

- to develop strategies to increase model compliance to distributional assumptions regarding the residual error.
- to develop methods to judge the appropriateness and improve the quantification of parameter uncertainty in pharmacometric model-based analysis.
- to develop and evaluate modeling strategies suitable for model-based analysis of confirmatory trials, with particular considerations regarding prespecification of the analysis and type I error control.

Methods

Methods will be presented in sequence for each of the three components of this thesis work: residual error modeling (Paper I), parameter uncertainty (Paper II-V) and model prespecification (Paper VI-VII).

Residual error modeling

Commonly used models

Commonly used error models in NLMEM can be expressed using Eq. 10.

$$g\left(f\left(t_{ij},\theta_i,z_{ij}\right),\varepsilon_{ij}\right) = f\left(t_{ij},\theta_i,z_{ij}\right)^{\zeta} \times \varepsilon_{ij,slope} + \varepsilon_{ij,intercept} \qquad \text{Eq. 10}$$

where $g(\cdot)$ is the residual error function, $f(\cdot)$ is the structural model, t_{ij} is the observation time of individual i at time j, θ_i is the vector of individual parameters, z_{ij} is the vector of individual covariates, ε_{ij} are residual terms, ζ is a power exponent, and $\varepsilon_{ij,slope}$ and $\varepsilon_{ij,intercept}$ are residuals with mean 0 and variances σ^2_{slope} and $\sigma^2_{intercept}$ respectively. $\varepsilon_{ij,slope}$ and $\varepsilon_{ij,intercept}$ are independent within and between individuals, as well as between each other. Eq. 11 displays the variance of the observations y_{ij} based on Eq. 10.

$$var(y_{ij}) = f\left(t_{ij},\theta_i,z_{ij}\right)^{2\zeta} \times \sigma^2_{slope} + \sigma^2_{intercept} \qquad \text{Eq. 11}$$

Three commonly used error models are the additive, proportional and combined (additive plus proportional) models. The combined error model corresponds to the linear slope-intercept model ($\zeta = 1$). The additive model is obtained by setting $\sigma^2_{slope} = 0$. In this case the error is homoscedastic, i.e. it does not depend on the model predictions. The error using the proportional model, which is obtained by setting $\sigma^2_{intercept} = 0$, is heteroscedastic as it depends on model predictions.

In NLME modeling based on maximum likelihood, estimated model parameters correspond to maximum likelihood estimates only if the scedasticity and the distribution shape of the residual error are correctly specified, i.e. if the variance is correctly specified and the residuals are normally distributed (Figure 3). However, the relationship between residuals and model predictions may not be additive and/or proportional. The distribution of the residuals may be skewed or contain more extreme values than the normal distribution, which could alter maximum likelihood properties.

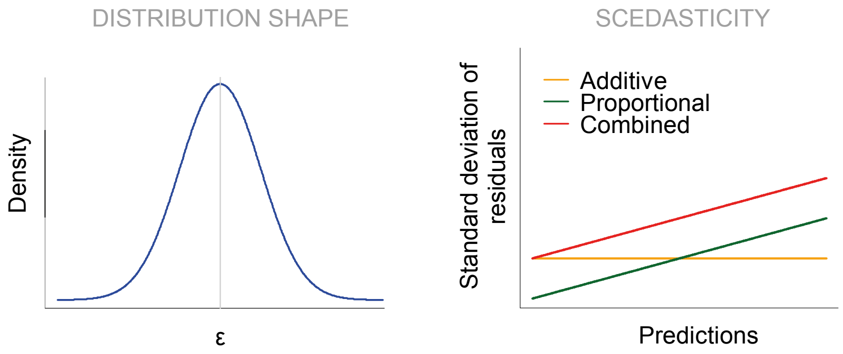

Figure 3. Maximum likelihood assumptions regarding the residual error: residuals ε are normally distributed (distribution shape) and their relationship to model predictions (scedasticity) is correctly specified.

Proposed error models: dTBS and the t-distribution

Two strategies can be envisaged to render residual error models more flexible, so that maximum likelihood assumptions regarding the residual error are more easily met in the presence of skewed or outlying residuals.

dTBS

The first strategy is called dynamic Transform-Both-Sides (dTBS) and allows skewed and/or heteroscedastic residuals by performing parameter estimation on a transformed scale, on which the residuals are normally distributed. With dTBS, observations and model predictions are transformed using a Box-Cox distribution with shape parameter λ (Eq. 12). The residual function is a power ζ of the untransformed model predictions (Eq. 13).

$$\begin{cases} h(x,\lambda) = \dfrac{x^{\lambda}-1}{\lambda} \text{ if } \lambda \neq 0 \\ h(x,\lambda) = \log(x) \text{ otherwise} \end{cases} \qquad \text{Eq. 12}$$

$$h(y_{ij},\lambda) = h(f(t_{ij},\theta_i,z_{ij}),\lambda) + f(t_{ij},\theta_i,z_{ij})^{\zeta} \times \varepsilon_{ij} \qquad \text{Eq. 13}$$

where x is a variable, $h(x,\lambda)$ is its Box-Cox transform with shape parameter λ, y_{ij} are observations, $f(t_{ij},\theta_i,z_{ij})$ are model predictions, ζ is a power parameter and ε_{ij} are residual terms. All model parameters including the transformation (shape) parameter λ and the power parameter ζ can be estimated using maximum likelihood assuming that the residuals on the transformed scale are normally distributed with correctly specified scedasticity. The OFV is set to minus two times the log-likelihood of the data on the *untransformed* scale given all parameters in order to allow for the estimation of λ. This OFV can be obtained from the likelihood of the data on the *transformed* scale using the change of variable formula (Eq. 14).

$$\begin{cases} L_{\mathrm{Y}} = L_{h(\mathrm{Y},\lambda)} \times \dfrac{d\big(h(\mathrm{Y},\lambda)\big)}{dY} = L_{h(\mathrm{Y},\lambda)} \times \mathrm{Y}^{\lambda-1} \\ OFV = -2\,LL_{\mathrm{Y}} = -2LL_{h(\mathrm{Y},\lambda)} - 2(\lambda-1)\log(\mathrm{Y}) \end{cases} \qquad \text{Eq. 14}$$

where L_Y is the likelihood of the *untransformed* data and $L_{h(Y,\lambda)}$ the likelihood of the *transformed* data. λ > 1 indicates that the distribution of the residuals is left skewed on the *untransformed* scale and λ < 1 indicates that the distribution is right skewed. If λ = 1, the residuals are normally distributed, and if λ = 0 they are log-normally distributed. With dTBS the error is proportional to the ζ power of the model predictions on the *transformed* scale, which corresponds to an error approximately proportional to the (1 – λ + ζ) power of the model predictions on the *untransformed* scale (Eq. 15). The dTBS model includes the additive (λ = 1 and ζ = 0) and proportional (λ = 1 and ζ = 1) error models, as well as the additive-on-log error model (λ = 0 and ζ = 0).

$$Var(y_{ij}) = Var\left(h(y_{ij},\lambda)\right) \times \left[\frac{dh\big(f(t_{ij},\theta_i,z_{ij}),\lambda\big)}{df(t_{ij},\theta_i,z_{ij})}\right]^{-2} \qquad \text{Eq. 15}$$

$$\approx f(t_{ij},\theta_i,z_{ij})^{2\zeta} \times \sigma^2 \times f(t_{ij},\theta_i,z_{ij})^{2(1-\lambda)} = \sigma^2 \times f(t_{ij},\theta_i,z_{ij})^{2(1-\lambda+\zeta)}$$

where y_{ij} are observations, h($y_{ij,}$,λ) are the Box-Cox transforms with shape parameter λ, $f(t_{ij},\theta_i,z_{ij})$ are model predictions, ζ is the power parameter and σ^2 is the residual variance.

Student's t-distribution

Instead of transforming both sides to obtain normally distributed residuals, an alternative strategy for increasing the flexibility of the residual error model is to change the distributional assumption. In this work we used the Student's t-distribution, which is a symmetric distribution characterized by its degree of freedom ν. The degree of freedom governs the heaviness of the tails of the distribution: the t-distribution approaches the normal distribution when $\nu \rightarrow +\infty$, and displays heavier tails when $\nu \rightarrow 0$ (Figure 4).

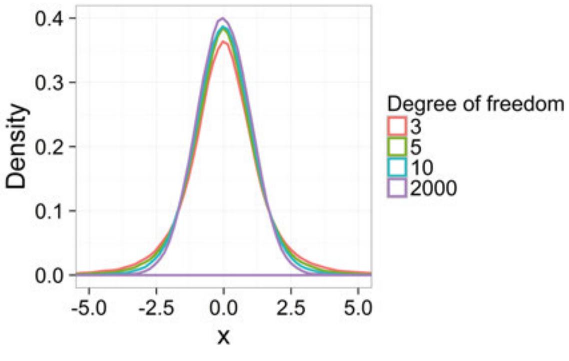

Figure 4. Student's t-distributions with varying degrees of freedom ν.

Model parameters and the degree of freedom of the t-distribution can be estimated by minimizing the OFV expressed as PDF of the Student's t-distribution:

$$OFV = -2\log\left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu Var(Y)}} \times \left(1+\frac{1}{\nu}\frac{\left(Y-f(t,\theta,z)\right)^2}{Var(Y)}\right)^{-\frac{\nu+1}{2}}\right) \qquad \text{Eq. 16}$$

where Γ is the gamma function, ν is the degree of freedom, Y are the observations and $f(t,\theta,z)$ are model predictions. For estimation, the lower bound of ν was set to 3 and its upper bound to 200, as the variance of the t-distribution is undefined for $\nu < 3$ and high ν approximates a normal distribution.

Evaluation of the new error models on real data examples

The dTBS and t-distribution approaches were tested on 10 previously published real data examples[34-42] detailed in Table 4.

Table 4. Description of the 10 real data examples used to investigate the dTBS and t-distribution approaches

Model	Data type	Model type	Error model	Trans-formation	Number of obs.	Number of ID
ACTH/ cortisol[34]	PD	turnover	combined[a]	-	364	7
Cladribine[36]	PK	i.v. 3CMT	combined	-	488	65
Cyclophos-phamide/ Metabolite[37]	PK	oral 4CMT, CL induction	additive (parent), combined (metabolite)	-	383	14
Ethambutol[38]	PK	oral 2CMT, transit	combined	log	1869	189
Moxonidine PK[39]	PK	oral 1CMT	additive	log	1021	74
Moxonidine PD[39]	PD	Emax	additive	log	1364	97
Paclitaxel[40]	PD	neutrophil	additive	fixed Box-Cox (λ=0.2)	523	45
Pefloxacin[41]	PK	i.v. 1CMT	proportional	-	337	74
Phenobarbital[42]	PK	i.v. 1CMT	proportional	-	155	59
Prazosin[35]	PK	oral 1CMT	proportional	-	887	64

[a] additive component fixed; obs.: observations; ID: individuals; PK: pharmacokinetic; PD: pharmacodynamic; CMT: compartment; i.v.:intravenous.

In the dTBS approach, the Box-Cox and power parameters were estimated either simultaneously or alone (the latter by fixing the other parameter to its value in the original model) using the FOCEI and SAEM algorithms. For the t-distribution approach, the scedasticity model was kept identical to the original model and the degree of freedom was estimated simultaneously to all other model parameters using the LAPLACE method with user-defined likelihood. In terms of implementation, the dTBS model could be used without modification of the model file on the *untransformed* scale in the PsN software[43] using the *–dtbs* option. The t-distribution was coded manually.

The impact of the new error models was assessed based on the likelihood ratio test. All model parameter estimates, and their standard errors (SE) if available, were compared. Plots of observations versus individual predictions, individual plots and VPC were inspected, as well as the distribution and scedasticity of conditional weighted residuals (CWRES), normalized prediction distribution errors (NPDE) and individual weighted residuals (IWRES). Changes in individual OFV (OFV_i) were investigated to identify whether subsets of individuals benefited more than others from a given residual error model. Cross-validation techniques were also used to assess the predictive performance of the dTBS approach.

Evaluation of the new error models on simulated examples

The bias, precision and type I error rates of the new error parameters were investigated using stochastic simulations and estimations. Drug plasma concentration data was simulated according to a one-compartment disposition, first order absorption and elimination PK model displaying additive, proportional or additive-on-log error models. Population values used for simulation were a clearance (CL) of 10 liters/hour, a volume of distribution (V) of 100 liters and an absorption constant (KA) of 1/hour, with all inter-individual variabilities (IIV) set to 30% on the standard deviation scale. The RUV was 0.2 for the additive and additive-on-log models and 20% for the proportional model. For each scenario, 500 datasets comprising 400 observations from 50 patients with PK samples at 0.25, 0.5, 1, 2, 5, 8, 12 and 24 hours (h) after administration of a single oral dose of 1000 milligrams (mg) were simulated. The FOCEI and SAEM estimation methods were used for the dTBS scenarios, and the LAPLACE estimation method was used for the t-distribution scenarios.

Parameter uncertainty

Apart from the assumptions about the residual error, which are necessary for appropriate model estimation, further distributional assumptions about the uncertainty around model parameters often need to be made when taking model-based decisions. A number of methods are available to compute the uncertainty around model parameter estimates: the covariance matrix, the bootstrap, likelihood profiling and stochastic simulation and estimations. However, how well these different methods perform in NLMEM remains insufficiently understood.

Commonly used methods

The most commonly used method to assess parameter uncertainty is through the **covariance matrix**. Based on maximum likelihood theory, assuming the number of individuals is high and the random effects and residual variability are normally distributed, the distribution of the maximum likelihood estimates can be approximated by a multivariate normal distribution with mean $\hat{\Theta}$ and covariance matrix $\hat{V}$. The covariance matrix $\hat{V}$ is estimated at the maximum likelihood estimates $\hat{\Theta}$, and the square roots of its diagonal elements correspond to the standard errors of the model parameters[44]. Three estimators of the covariance matrix are commonly used[45].

The first estimator is the inverse of the FIM, which corresponds to the negative of the Hessian matrix, i.e. the square matrix of the second-order partial derivatives of the likelihood function (Eq. 17). It is referred to as the R matrix in NONMEM[46]. R is a consistent estimator of the covariance matrix.

$$R = \left(-\frac{1}{2}\frac{\partial^2(-2LL)}{\partial\Theta\partial\Theta^T}\right)^{-1} \qquad \text{Eq. 17}$$

where -$2LL$ is minus two times to log-likelihood of the data and Θ is the vector of model parameters.

The second estimator is the S matrix, which corresponds to the inverse of the gradient product matrix based on the first-order derivatives of the likelihood function (Eq. 18). It is referred to as the S matrix in NONMEM. S is also a consistent estimator of the covariance matrix.

$$S = \left(\frac{1}{4}\frac{\partial(-2LL)}{\partial\Theta} \times \frac{\partial(-2LL)}{\partial\Theta^T}\right)^{-1} \qquad \text{Eq. 18}$$

where -$2LL$ is minus two times the log-likelihood of the data and Θ is the vector of model parameters.

The last and most commonly used estimator in NLMEM is the “sandwich” matrix SW, which is a combination of the R and S matrices (Eq. 19):

$$SW = RS^{-1}R \qquad \text{Eq. 19}$$

Note that with rich data and normally distributed random effects, all three estimators are expected to converge to the same value. The sandwich estimator is usually preferred for continuous data because it is expected to be more robust to misspecifications of the random effect distributions. For discrete data, for which the normality assumption on the random effects is more likely to hold, R is typically preferred. The standard errors of model parameters are derived directly from the estimator as the square roots of the diagonal elements of the matrix. Approximate asymptotic CI for a given parameter $\hat{\theta}_i$ at the α confidence level can then be computed as $\left[\hat{\theta}_i \pm z_{1-\alpha/2} SE_{\hat{\theta}_i}\right]$, where $z_{1-\alpha/2}$ is the 1-α/2 quantile of the normal distribution. Note that covariance-matrix based CI are symmetric.

The method which is often considered the gold standard to assess parameter uncertainty is the **bootstrap**[47]. With bootstrap, parameter uncertainty distributions are typically represented by a set of model parameter vectors, which are obtained by estimating model parameters on a number of bootstrapped datasets (e.g. 1000). In NLMEM, bootstrapped datasets are typically obtained from the original data using case bootstrap, where the full data of one individual is resampled with replacement. Bootstrapped datasets thus contain the same number of individuals as the original dataset, with some individuals appearing more than once and some individuals not present at all. Stratification, i.e. performing the resampling within subgroups of the data typically defined by design variables such as treatment arm or sex, is often used to obtain bootstrapped datasets similar in structure to the original dataset. Parameters are estimated based on the bootstrapped datasets and the final model, with initial parameter estimates set to $\hat{\theta}$. Nonparametric percentiles-based CI can be derived for each parameter from the bootstrap parameter vectors as $\left[\hat{\theta}_{i,p_{boot,\alpha/2}}; \hat{\theta}_{i,p_{boot,1-\alpha/2}}\right]$, where $p_{\alpha/2}$ and $p_{1-\alpha/2}$ are the α/2th and (1-α/2)th percentiles of the ordered parameter values. Bootstrap CI are not necessarily symmetric. Note that many other ways of performing the resampling and of computing bootstrap CI exist[48-50], but they will not be discussed here.

Log-likelihood profiling[51] (LLP), often referred to as profile likelihood or likelihood profiling, can also be used to assess parameter uncertainty. With LLP, the CI around a parameter is computed by estimating the OFV for an array of fixed values of this parameter, while estimating the remaining parameters. Values of the parameter which lead to OFV increases of 3.84 compared to the OFV of the final model are taken as the bounds of the 95% CI of the parameter. The critical value (3.84) corresponds to the value of the chi-square distribution for one degree of freedom and at the α confidence level ($\chi_{\alpha}^{df=1}$). The LLP CI can be expressed as $\left[\hat{\theta}_{i,\mathrm{lb},\chi_{\alpha}^{df=1}};\hat{\theta}_{i,\mathrm{ub},\chi_{\alpha}^{df=1}}\right]$, where *lb* and *ub* are the lower and upper bound, respectively. Despite some work on multivariate implementation[52], LLP currently does not provide full uncertainty distributions, but only univariate bounds. LLP CI are not necessarily symmetric.

Lastly**, stochastic simulations and estimations** (SSE), also known as parametric bootstrap, can be used to estimate parameter uncertainty. With SSE, a given number of datasets (e.g. 1000) identical in design to the original dataset are simulated using the final model and the final parameter estimates. The simulated datasets are then used in the same manner as the bootstrapped datasets to re-estimate the model parameters. Percentiles-based CI can then be derived for each parameter from the re-estimated parameter vectors as $\left[\hat{\theta}_{i,p_{SSE,\alpha/2}};\hat{\theta}_{i,p_{SSE,1-\alpha/2}}\right]$, where $p_{\alpha/2}$ and $p_{1-\alpha/2}$ are the α/2th and (1-α/2)th percentiles of the ordered parameter values. This method is slightly different from the other presented methods as it evaluates the uncertainty using *simulated* data. SSE uncertainty is obtained using the same model for data simulation and data fitting, which is not the case when using real data. The SSE thus corresponds to the uncertainty distribution of a given model and design in the absence of model misspecification. A comparison of the different methods is provided in Table 5.

Table 5. Comparison of methods to estimate parameter uncertainty

Characteristic	Cov. matrix	Bootstrap	LLP	SSE	SIR*
Computation time	rapid	long	middle	long	middle
Obtention	potential numerical difficulties	high number of estimations	moderate number of estimations	high number of estimations	no estimation, high number of evaluations
Data structure	rich data	big enough groups, balanced designs	no restriction	no restriction	no restriction
Distribution assumptions	multivariate normal, full, symmetric	nonparametric, full, asymmetric	bounds only, asymmetric	nonparametric, full, asymmetric	nonparametric, full, asymmetric

Cov.: covariance; *Sampling Importance Resampling, described in a later subsection (p.34).

The dOFV diagnostic: assessing uncertainty adequacy

The covariance matrix, bootstrap, LLP and SSE may lead to different uncertainty estimates. However, it is difficult to know which method to rely on in a given case, as no diagnostic assessing the adequacy of a given parameter uncertainty distribution is routinely used in NLMEM. A new diagnostic was thus proposed to assess and compare the adequacy of parameter uncertainty distributions. The diagnostic was based on the comparison of the dOFV distribution of a given uncertainty estimate with a theoretical distribution. The diagnostic was developed for the bootstrap, but can be applied to any uncertainty estimate for which parameter vectors can be obtained (i.e. all presented methods except the LLP).

The dOFV distribution of the uncertainty estimate to assess is obtained by *evaluating* the OFV (i.e. MAXEVAL = 0 in NONMEM[46]) of the original data D for N parameter vectors $\widehat{\Theta}_n$ ($n = 1, \ldots, N$) sampled from the uncertainty estimate. The OFV of the original data with the final parameter estimates $\widehat{\Theta}$ is then subtracted from each of these OFV to obtain N dOFV (Eq. 20).

$$dOFV_n = OFV_{\widehat{\Theta}_n,D} - OFV_{\Theta,D} \qquad \text{Eq. 20}$$

where $dOFV_n$ is the n^{th} bootstrap dOFV. The first index of the OFV corresponds to the parameter vector used, and the second to the dataset the parameter vector is estimated/evaluated on. $\widehat{\Theta}_n$ is the parameter vector estimated on the n^{th} bootstrap dataset, and $\widehat{\Theta}$ is the parameter vector estimated on the original dataset D.

The theoretical dOFV distribution corresponds to a chi-square distribution with degrees of freedom equal to the number of estimated model parameters.

The proposed diagnostic displays the quantile function, also known as the inverse cumulative distribution function, of the two just described dOFV distributions as illustrated in Figure 5.

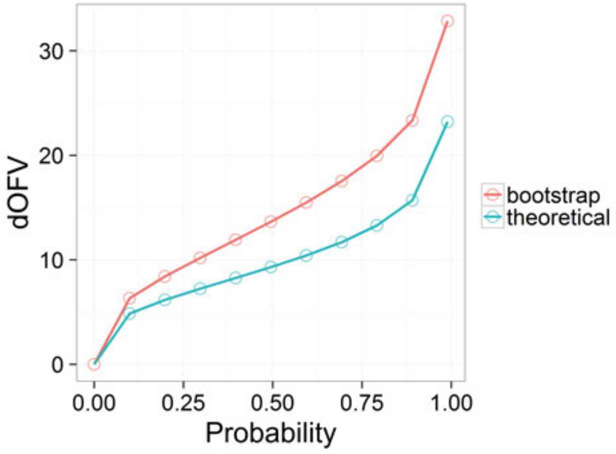

Figure 5. Example of the dOFV distribution plot diagnostic where the parameter uncertainty estimate to evaluate was obtained by bootstrap.

The principle behind the dOFV diagnostic is that if the parameter vectors sampled from the uncertainty estimate were representative of the true uncertainty, their dOFV distribution should follow a chi-square distribution[53]. The degrees of freedom of this chi-square distribution should be asymptotically equal to the number of estimated parameters for unconstrained fixed effects models. This means that the dOFV distribution of the uncertainty estimate should overlay the theoretical dOFV distribution. However, for NLMEM the exact degree of freedom is unknown. This is due to a number of factors, including the estimation of bounded parameters such as variances, which may not account for full degrees of freedom, or properties of the estimation method[54]. The dOFV distribution is thus not necessarily expected to collapse to the theoretical dOFV distribution when the uncertainty is appropriate. However, as the degree of freedom cannot exceed the number of estimated parameters, the diagnostic considers the uncertainty estimate appropriate if its dOFV distribution is *at or below* the theoretical distribution.

While the dOFV distribution is not necessarily expected to collapse to the theoretical dOFV distribution, it is expected to collapse to the SSE dOFV distribution, which corresponds to the expected dOFV distribution of a given NLMEM in the absence of model misspecification. The SSE could not be routinely used as a reference distribution in the dOFV diagnostic due to its high computational burden. It was computed for the investigated examples in order to judge whether using the theoretical distribution as a surrogate for the SSE distribution was appropriate.

Evaluation of bootstrap adequacy in NLMEM

The dOFV diagnostic was applied to the bootstraps of two real data and two simulation examples. The two real data examples were the phenobarbital[42] and pefloxacin[41] examples previously described in Table 4. The first simulation example consisted of an intravenous (i.v.) 1-compartment PK model with linear elimination. CL and V were set to 1, exponential IIV was 30% on both parameters, and the RUV was additive on the log scale with a standard deviation of 0.2. Three different dataset sizes were investigated: 20, 50 and 200 individuals, with four observations each at 0.25, 0.5, 1 and 2 units after single dose administration. The second simulation example consisted of a pharmacodynamic (PD) dose-response sigmoidal Emax model, with a baseline E0 of 10, an additive maximum effect EMAX of 100, a dose leading to half the maximum effect ED50 of 5, a Hill factor of 0.7, 30% exponential IIV on E0 and ED50, and a 10% proportional RUV. Three different dataset sizes were investigated: 20, 50 and 200 individuals with four observations each at doses of 0, 2.5, 5 and 15.

For the real data examples, bootstrap and SSE dOFV distributions were assessed for the original dataset, 10 simulated datasets using the original design, and 10 datasets simulated using an 8-fold increase in the number of individuals. The 8-fold increased datasets were used to test the influence of sample size on bootstrap performance. In the simulation examples, bootstrap and SSE dOFV distributions were assessed for 100 datasets for each dataset size. The degree of freedom of all dOFV distributions was calculated as the mean of each dOFV distribution. The adequacy of parameter uncertainty was evaluated based on parameter CI, using CI obtained from the SSE as a reference. Coverage at the 90% level was investigated for each parameter by calculating the proportion of datasets for which the 90% CI included the true simulation value for that parameter. From statistical theory, the expected coverage at the 90% level is 0.90, i.e. 90% of the simulated datasets should include the true simulation values in their 90% CI.

SIR: improving parameter uncertainty estimation

Given the observed limitations of currently available methods highlighted in Table 5, a method based on Sampling Importance Resampling[55] (SIR) was proposed to improve the estimation of parameter uncertainty distributions in NLMEM. SIR provides a set of m parameter vectors representative of the true and unknown parameter uncertainty distribution. SIR is performed in three steps:

1. Step 1 (**sampling**): M ($M > m$) parameter vectors are randomly sampled from a proposal multivariate distribution.

2. Step 2 (**importance weighting**): an importance ratio is computed for each of the M sampled parameter vectors. It corresponds to the probability of being sampled in the true parameter uncertainty distribution and is computed as the likelihood of the data given the parameter vector, weighted by the likelihood of the parameter vector in the proposal distribution (Eq. 21).

   $$IR = \frac{exp(-0.5 \times dOFV)}{relPDF} \qquad \text{Eq. 21}$$

   where *IR* is the importance ratio, *dOFV* is the difference between the OFV of the sampled parameter vector and the OFV of the final parameter estimates, and *relPDF* is the probability density function of the parameter vector relative to the probability density of the final parameter estimates given the proposal distribution.

3. Step 3 (**resampling**): in the last step, m parameter vectors are resampled from the pool of M sampled vectors based on their importance ratio.

When performing SIR, three settings need to be chosen: the proposal distribution, the number of samples M and the number of resamples m. SIR was first developed using a non-iterative 1-step procedure starting from the sandwich covariance matrix as proposal distribution, with M = 5000 and

m = 1000 (Paper III). Diagnostics were developed to assess whether the chosen SIR settings were appropriate, in which case SIR results were considered final. After investigating the influence of different SIR settings, an improved iterative 5-step SIR procedure was developed (Paper IV), starting from the sandwich covariance matrix, a limited bootstrap (e.g. 200 samples or less) or a generic covariance matrix (e.g. with 30% relative standard error (RSE) on fixed effects, 50% RSE on random effects and 10% RSE on residual variability). The resamples of the first step were used as proposal distribution of the second step, and the procedure was repeated for a series of M = 1000, 1000, 1000, 2000, 2000 samples and m = 200, 400, 500, 1000, 1000 resamples. A summary of the 5-step SIR procedure is provided in Figure 6.

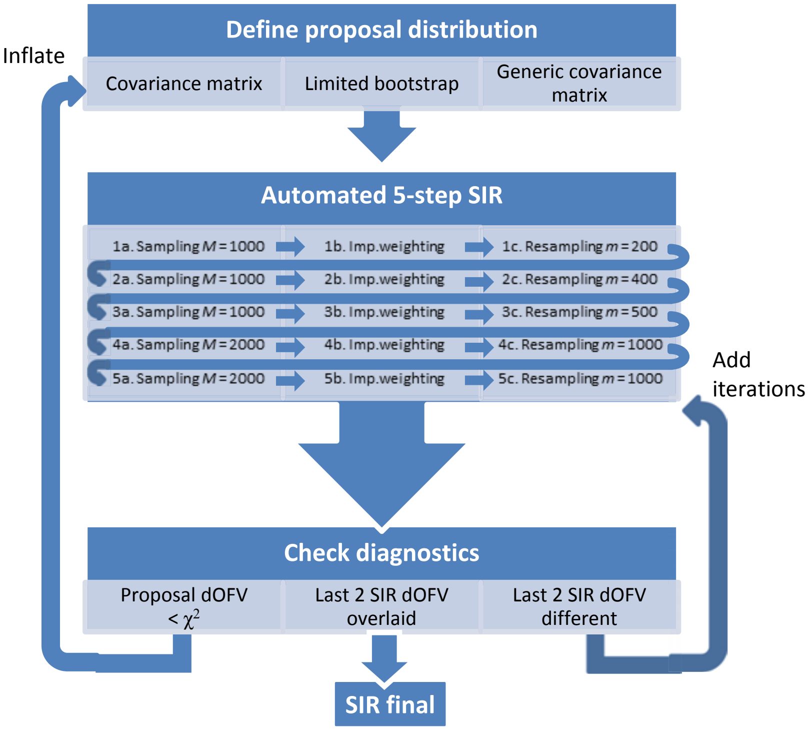

Figure 6. SIR workflow. To obtain SIR parameter uncertainty for a given model, a proposal distribution first needs to be chosen by the user. Choices for this distribution in decreasing order of efficiency are the covariance matrix, a limited bootstrap or a generic covariance matrix. Once the proposal is chosen, the PsN *sir* function is used to automatically perform 5 SIR iterations. SIR results are considered final if the dOFV distributions of the last 2 iterations are overlaid.

Three graphical diagnostics were used to judge whether SIR results could be considered final at the end of the 5-step procedure: the dOFV distribution plot (developed in the previous subsection), the spatial trends plot and the temporal trends plot. Each plot will now be described.

1. The **dOFV distribution plot** (Figure 7) is the main diagnostic plot and diagnoses SIR convergence. It displays the dOFV distributions of the samples *M* and resamples *m* at each SIR iteration. SIR results are considered final when the resamples dOFV distributions of the last two SIR iterations are overlaid up to sampling noise, provided the initial proposal is above the theoretical distribution. Overlaid distributions correspond to a case where the uncertainty estimate cannot be further improved. An initial proposal below the theoretical indicates that the initial proposal is too narrow, in which case SIR should be restarted using an inflated proposal distribution, i.e. by multiplying the covariance matrix by a single factor until the proposal is above the reference chi-square. Details on the need for inflation will be discussed later. In case the last two SIR dOFV distributions are not overlaid, further iterations (e.g. with *M* = 2000 and *m* = 1000) should be added until convergence.

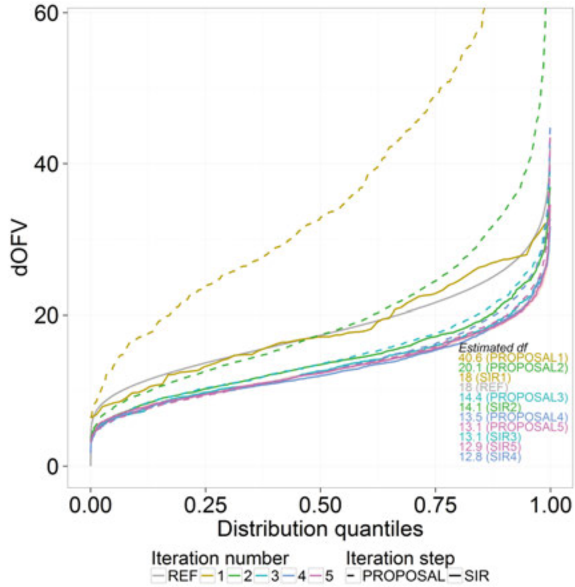

Figure 7. Example dOFV diagnostic plot showing convergence of the 5-step SIR procedure for a model with 18 estimated parameters.

2. The **spatial trends plot** (Figure 8) assesses the adequacy of the proposal distribution for each parameter. It displays the number of resampled parameters divided by the number of available parameters, i.e. the resampling proportion, in 10 different bins of the parameter space. The bins were obtained by binning parameters ordered by increasing value ("spatial" bins). Four types of trends can be observed in this plot: *horizontal trends* (i.e. no trend), which mean that the proposal distribution is close to the true uncertainty; *bell-shaped trends*, which mean that the proposal distribution is wider than the true distribution; *u-shaped trend*s,

which mean that the proposal distribution is narrower than the true distribution; and *diagonal trends*, which mean that the proposal distribution has a different (a)symmetry than the true distribution.

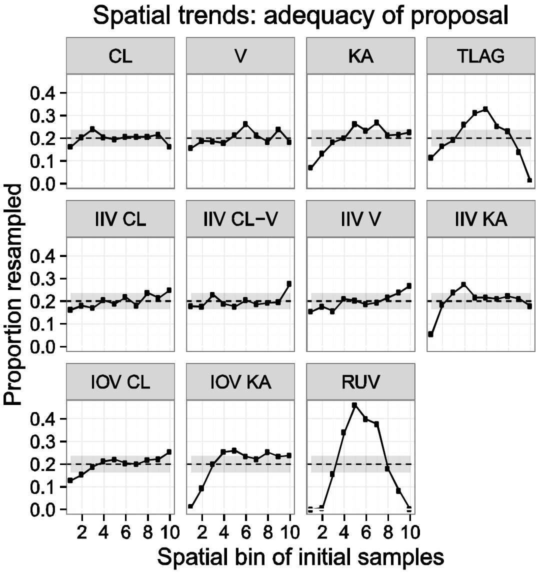

Figure 8. Example spatial trend plot showing the (in)adequacy of the proposal distribution. In this example the proposal appears adequate for CL, V and their IIV (horizontal trends), too wide for TLAG and RUV (bell-shaped trends) and lacking asymmetry for KA and its IIV and IOV (diagonal trends). The dotted line represents the expected proportion and the grey shaded area the stochastic noise around the expected proportion.

3. The **temporal trends plot** (Figure 9) indicates, for each parameter, whether *M,* or more specifically the *M/m* ratio, was high enough to compensate for the inadequacy of the proposal distribution potentially observed in the spatial trends plot. The temporal trends plot focuses on the top spatial bin, defined from the spatial trends plot as the bin with the highest resampling proportion. Instead of binning sampled parameters based on their value as for the spatial trends plot, resampled parameters are now binned based on the order in which they were resampled ("time" bins). Two trends can be observed for this diagnostic: *horizontal trends* (i.e. no trend), which mean that *M/m* was sufficient to compensate for a potential inadequacy of the proposal distribution; and *downward diagonal trend*s, which mean that there were not enough samples in the SIR procedure to fully correct the proposal uncertainty.

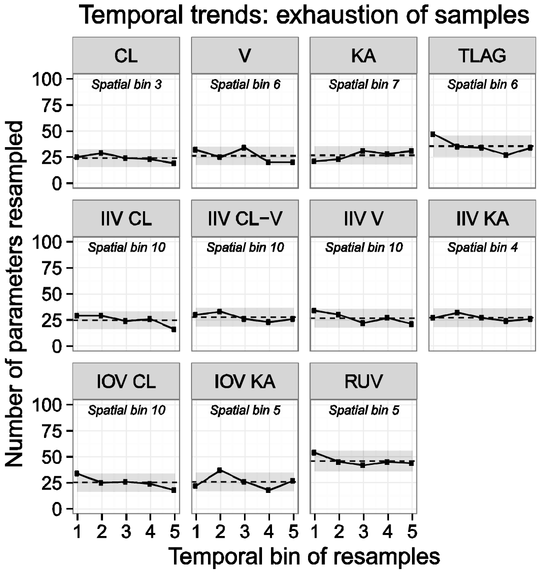

Figure 9. Example temporal trend plot showing the exhaustion of samples. In this example, there appears to be no exhaustion of samples (horizontal trends). The dotted line represents the expected number of resamples and the grey shaded area the stochastic noise around the expected number of resamples.

Evaluation of the 1-step SIR on simulated data

The properties of the initially developed 1-step SIR procedure (sandwich covariance matrix as proposal distribution, M = 5000 samples and m = 1000 resamples) were investigated on two simulation examples, an i.v. 1-compartment PK model with first-order elimination and a PD dose-response Emax model. The simulation examples were identical to those used for the investigation of the dOFV diagnostic (see page 28) except that the Hill factor was fixed to 1. Parameter uncertainty was computed using SIR and using the covariance matrix for each of the parameters and each of the 500 simulated datasets. The coverage at the 95% level, i.e. the proportion of datasets for which the computed 95% CI included the true simulation value, was calculated. The coverage obtained with SIR was compared to the coverage obtained with the covariance matrix.

Evaluation of the 1-step SIR on real data

The 1-step SIR procedure was also applied to three real data PK examples: the moxonidine[39], pefloxacin[41] and phenobarbital[42] examples, which were i.v. and oral 1-compartment PK models (Table 4). The developed diagnostics were used to judge whether the default SIR settings were appropriate for these examples. The parameters' 95% CI were compared between SIR, the covariance matrix, bootstrap (1000 samples, no stratification) and LLP.

In addition, the real data examples were used to investigate the influence of SIR settings, i.e. the number of samples *M* and the proposal distribution, on SIR results. Different numbers of samples were investigated (*M* = 2000, 4000, 6000, 8000 and 10000). The number of resamples *m* was not modified (*m* = 1000), as this number was chosen in order to have sufficient precision on the outer bounds of the CI of interest. Corresponding *M/m* ratios were thus 2, 4, 6, 8 and 10. Different proposal distributions were also investigated, which corresponded to inflations and deflations of the covariance matrix. Variances and covariances of the covariance matrix were either increased or decreased by factors of 0.5, 0.75 1.5 and 2.

Evaluation of the 5-step SIR on real data

Based on the results obtained with the 1-step procedure, an improved 5-step SIR workflow was developed and evaluated on 25 NLMEM[41,56-79]. A summary of the characteristics of the models is provided in Table 6. The evaluation of the 5-step procedure was based on the number of iterations needed until stabilization, and on the degree of freedom of the dOFV distribution (calculated as the mean of the dOFV distribution) obtained at stabilization. Typical and atypical behaviors were reported and analyzed.

The influence of the initial proposal distribution was investigated by performing the 5-step procedure using a generic covariance matrix as initial proposal distribution ("generic SIR") instead of the covariance matrix or the limited bootstrap ("informed SIR"). The generic covariance matrix was set to a multivariate normal distribution with 30% RSE on fixed effects, 50% RSE on inter-individual and inter-occasion variabilities, 10% on residual variabilities, and no correlations between any of the parameter uncertainties. Final parameter uncertainty distributions, the number of iterations until stabilization, the runtime and the final degrees of freedom were compared between the informed SIR and the generic SIR.

The uncertainty obtained with the informed SIR was also compared to the uncertainty obtained with three other methods: the covariance matrix, bootstrap and SSE, based on 1000 samples for all methods. Following metrics were compared: i) RSE, ii) relative widths of the parameters' 95% CI, calculated as the distance between the CI's upper and lower bounds divided by the final parameter estimate, and iii) the asymmetry of the CI, quantified by the ratio of the distance between the CI's upper bound and the median divided by the distance between the CI's lower bound and the median. Runtime comparisons were performed between SIR and bootstrap using the ratio between the time 7000 likelihood *evaluations* were expected to take for SIR versus the time 1000 likelihood *estimations* were expected to take for the bootstrap.

Table 6. Summary of the 25 NLMEM models used to investigate the 5-step SIR

Model characteristic	**Categories / Mean value [Range]**
Type of model	10 PK, 15 PD (total 25)
Type of data	21 continuous, 4 categorical
Number of estimated parameters	15 [1-39]
Proportion of random effects (%)	27 [0-77]
Number of individuals	115 [6-551]
Number of observations	4076 [58-47784]
Number of observations/individual	28 [1-102]
Estimation method	3 FO, 5 FOCE, 10 FOCEI, 7 LAPLACE

PK: pharmacokinetic; PD: pharmacodynamic; FO: first-order; FOCE: first-order conditional estimation; FOCEI: first-order conditional estimation with interaction; LAPLACE: Laplacian.

Application of SIR for decision-making using a WBPBPK model

The developed 5-step SIR procedure was used to estimate parameter uncertainty in an interspecies whole-body physiologically based pharmacokinetic (WBPBPK) model describing colistin and colistin methanesulfonate (CMS) PK. Parameter uncertainty obtained with SIR was used to drive model-building decisions and to predict human plasma concentrations.

The interspecies WBPBPK model was adapted from a model previously developed in rats[62] using plasma concentration data from five animal species: mice, rats, rabbits, baboons and pigs. A schematic of the model is presented in Figure 10. Species-independent priors derived from rat tissue homogenate experiments were implemented on all tissue-to-plasma partition coefficients (K_p). Physiological parameters such as tissue volumes, blood and urinary flow rates, hematocrit, plasma unbound fraction and glomerular filtration rates were fixed to values obtained from the literature. CMS hydrolysis to colistin ($CL_{hyd\text{-}CMS}$) was scaled across species using allometric scaling on tissue volume with estimated exponent. CMS renal clearance ($CL_{r\text{-}CMS}$) was scaled across species using allometric scaling on the glomerular filtration rate with exponent fixed to 1. Because colistin elimination remains poorly understood, three scaling models were tested for colistin nonrenal clearance ($CL_{nr\text{-}coli}$): allometric scaling based on volume with estimated exponent (Model A), allometric scaling based on volume with estimated exponent and corrected by maximum lifespan potential (Model B), and no scaling, i.e. species-dependent $CL_{nr\text{-}coli}$ (Model C). Parameter uncertainty was obtained via the 5-step SIR procedure starting from the sandwich covariance matrix and was used to evaluate the three scaling models.

The interspecies WBPBPK model was then used to predict plasma concentrations of colistin and CMS in human taking parameter uncertainty into account. Model predictions were compared to human data obtained from a published PK study which included 12 healthy volunteers who received a single i.v. infusion of 80 mg of CMS sodium. Colistin and CMS plasma concentrations were measured at 13 time points up until 18h post dose. Individual body weights, glomerular filtration rates and urinary flow rates were available from the published data; other physiological parameters were fixed to literature values. Plasma concentration predictions were performed using the same design as the published study and with all scaling models for which parameter uncertainty had been found reasonable. If Model C (no scaling) was retained, $CL_{nr\text{-}coli}$ would be set to the $CL_{nr\text{-}coli}$ estimated in baboons, scaled by the ratio of the mean body weights of both species to the -0.25 power. The adequacy of median profile predictions was assessed based on 200 datasets simulated including IIV, RUV and parameter uncertainty.

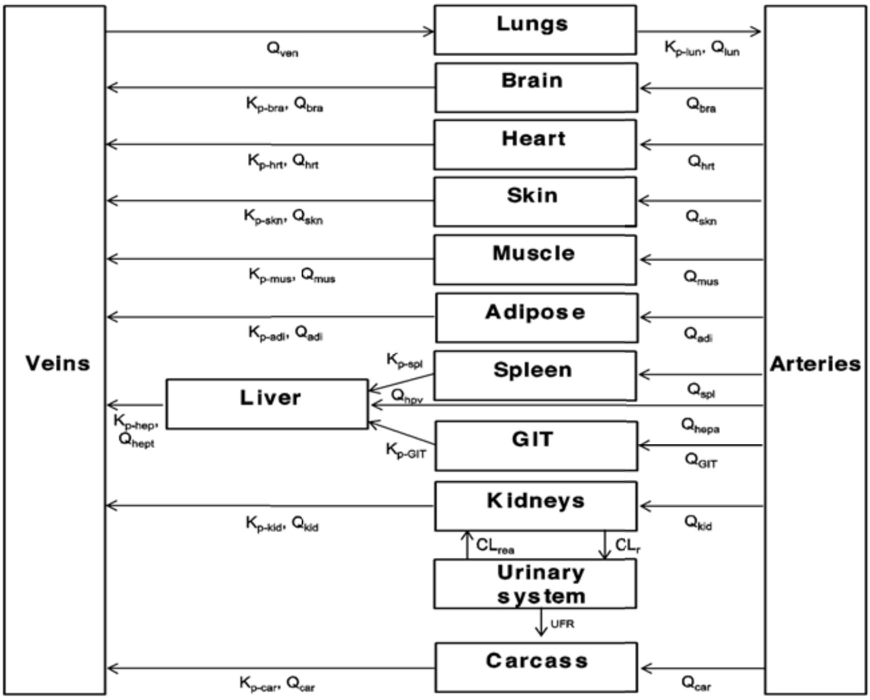
Veins
Lungs
Arteries
Q_{ven}
$K_{p\text{-}lun}$, Q_{lun}
Brain
$K_{p\text{-}bra}$, Q_{bra}
Q_{bra}
Heart
$K_{p\text{-}hrt}$, Q_{hrt}
Q_{hrt}
Skin
$K_{p\text{-}skn}$, Q_{skn}
Q_{skn}
Muscle
$K_{p\text{-}mus}$, Q_{mus}
Q_{mus}
Adipose
$K_{p\text{-}adi}$, Q_{adi}
Q_{adi}
$K_{p\text{-}spl}$
Spleen
Q_{spl}
Liver
Q_{hpv}
$K_{p\text{-}hep}$, Q_{hept}
Q_{hepa}
GIT
$K_{p\text{-}GIT}$
Q_{GIT}
Kidneys
$K_{p\text{-}kid}$, Q_{kid}
Q_{kid}
CL_{rea}
CL_r
Urinary system
UFR
Carcass
$K_{p\text{-}car}$, Q_{car}
Q_{car}

Figure 10. Schematic representation of the WBPBPK model. Q_{tissue}: physiological regional blood flow; V_{tissue}: physiological volume; $K_{p\text{-}tissue}$: tissue-to-plasma partition coefficient; CL_r: renal clearance; CL_{rea}: reabsorption clearance; UFR: urinary flow rate; GIT: gastro-intestinal tract.

Model prespecification

The previous subsections focused on parameter uncertainty as a key component of decision-making. A further layer of uncertainty which needs to be taken into account for decision-making, but is often neglected, is model uncertainty. To be used as primary analysis in confirmatory trials, pharmacometric models need to be prespecified, i.e. the model needs to be decided prior to data analysis and no data-driven model building is allowed. Note that this work focuses only on the uncertainty of the structural part of the model. Model uncertainty also exists on the random effects and residual error models, but flexible models covering a wide range of possible shapes (e.g. Box-Cox distributions, dTBS) are more common for these aspects than for the structural model. In addition, in the considered cases the impact of the IIV and RUV models was relatively limited, which is why only structural model uncertainty was considered.

Principle of model-averaging

Model-averaging[80] can be used as one way of addressing model uncertainty for a prespecified analysis. The principle of model-averaging is to conduct the analysis using a number M_{av} of prespecified models, instead of using a single model. First, each model is fitted separately to the data. Then, model predictions are computed as the weighted average of the predictions of each model using weights based on the respective fit of each model to the data. Hypothesis testing can be carried out on the endpoint to test (for example blood pressure reduction at the end of the trial), which is expressed as the weighted sum of the estimates from the different models:

$$\hat{\delta}_{av} = \sum_{m=1}^{m=M_{av}} \hat{w}_m \, \hat{\delta}_m \qquad \text{Eq. 22}$$

where $\hat{\delta}_{av}$ is the model-averaged endpoint, $\hat{w}_m$ is the estimated weight for model m and $\hat{\delta}_m$ is the endpoint estimated with model m. The uncertainty around the model-averaged estimate can be obtained similarly to single model estimates, based on the covariance matrix, bootstrap or SIR. The workflow of a model-averaged analysis is provided in Figure 11.

Model-averaged test for QT prolongation assessment

Model-averaging was first applied in the context of thorough-QT (TQT) studies. TQT studies are cross-over or parallel studies aiming at detecting a drug's potential for prolonging the QT interval. TQT studies typically involve four arms: placebo, therapeutic and supra-therapeutic dose of the test drug, and a positive control. Drug concentrations and QT intervals are measured simultaneously at *N* time points post-dose, and the QT interval is also measured at *M* time points pre-dose. The QT intervals are corrected for heart rate (QTc) with one of various possible correction methods (e.g. Fridericia's[81], Bazett's).

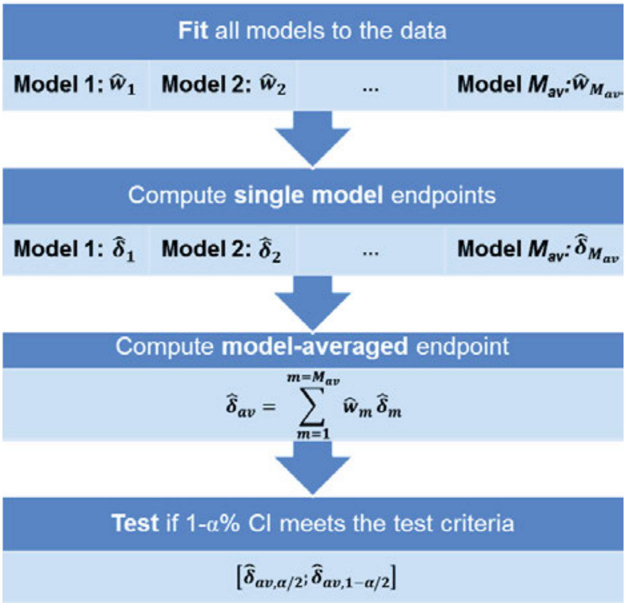
Fit all models to the data
Model 1: $\hat{w}_1$
Model 2: $\hat{w}_2$
...
Model M_{av}: $\hat{w}_{M_{av}}$
Compute single model endpoints
Model 1: $\hat{\delta}_1$
Model 2: $\hat{\delta}_2$
...
Model M_{av}: $\hat{\delta}_{M_{av}}$
Compute model-averaged endpoint
$\hat{\delta}_{av} = \sum_{m=1}^{m=M_{av}} \hat{w}_m \hat{\delta}_m$
Test if 1-α% CI meets the test criteria
$[\hat{\delta}_{av,\alpha/2}; \hat{\delta}_{av,1-\alpha/2}]$

Figure 11. Workflow of a model-averaged analysis.

Concentration-response analysis of TQT studies uses data from all study arms except the positive control, which is only used to validate the sensitivity of the QT measurement method. The corrected QT interval (QTc) is typically assumed to depend on a function of clock time (circadian rhythm) and drug concentrations as expressed in Eq. 23.

$$QTc_{klt} = p_t + \vartheta C_{klt} + \eta_{kl} + \varepsilon_{klt} \qquad \text{Eq. 23}$$

where the subscript *k* indicates the subject, *l* the treatment arm (*l* = *high* or *low*) and *t* the time point at which the observation was made. The *N* fixed effect parameters p_t describe the circadian effects separately for each time point. The slope ϑ is also regarded as a fixed effect. C_{klt} corresponds to the individual observed drug concentrations. The random subject effects η_{kl} describe the between subject variability. They are assumed to be independent across subjects and normally distributed with mean 0 and variance ω^2. The ε_{klt} describe the residual noise and are assumed to be normally distributed random variables with mean 0 and variance σ^2, which are independent of the η_{kl}, and independent within and between subjects.

Here we use change from mean individual baseline in QTc (ΔQTc) as primary endpoint (Eq. 24).

$$\Delta QTc_{klt} = p'_t + \vartheta C_{klt} + \eta'_{kl} + \varepsilon_{klt} \qquad \text{Eq. 24}$$

where p'_t are derived from Eq. 23. The random subject effects η'_{kl} are assumed normally distributed with mean 0 and with variance $1/M\ \sigma^2$, which is a fraction of the residual variability of the ε_{klt}. Under this model the ΔQTc observations are independent across subjects, and within subjects there is a compound symmetric correlation structure with all variance terms equal to $(1+1/M)\ \sigma^2$ and all covariance terms equal to $1/M\ \sigma^2$. The linear drug effect concentration-response model can be embedded in a more general class of models (Eq. 25) where f is a monotonically increasing function with $f(0) = 0$.

$$\Delta QTc_{klt} = p'_t + f(C_{klt}) + \eta'_{kl} + \varepsilon_{klt} \qquad \text{Eq. 25}$$

In the context of concentration-QT analysis, a mean increase of ΔQTc of at least 10 milliseconds (ms) over placebo at the mean over the individual maximum concentrations is regarded to be a potential safety risk[82]. This can be formalized for dose group l via the hypotheses stated in Eq. 26.

$$H_0: f(\gamma_{max,l}) \geq 10\, ms \quad versus \quad H_1: f(\gamma_{max,l}) < 10\, ms \qquad \text{Eq. 26}$$

where $\gamma_{max,l}$ is the geometric mean over the individual maximum concentrations. Three estimators of $f(\gamma_{max,l})$ were investigated: a parametric linear estimator $\hat{\vartheta}\hat{\gamma}_{max,l}$, a nonparametric estimator $\hat{f}(\hat{\gamma}_{max,l})$ based on I-splines[83], and a model-averaged estimator defined in Eq. 27.

$$\hat{\pi}\, \hat{\vartheta}\, \hat{\gamma}_{max,l} + (1-\hat{\pi})\hat{f}\, (\hat{\gamma}_{max,l}) \qquad \text{Eq. 27}$$

The data-driven weights $\hat{\pi}$ were adapted from the global Mean Integrated Square Error (MISE) weights from Yuan et al.[84] to the concentration-response context. Local MISE weights[84] and weights based on the Bayesian Information Criterion (BIC[85]) were also investigated. The two-sided 90% CI used for hypothesis testing were obtained via bootstrap (N = 999 samples).

Simulation study

Parallel group studies including a placebo arm, a therapeutic dose test drug arm and a supra-therapeutic dose test drug arm were simulated in order to evaluate the performance of the model-averaged test. The study design was identical to the real data example presented in the next paragraph. QTc were simulated according to Eq. 23 using values estimated from the real data, with individual concentrations simulated based on a two-compartment PK model with first-order absorption and first-order elimination. Four simulation settings were varied: noise level, sample size, drug effect model and drug effect size (Table 7). 1000 datasets were simulated for each combination of the different settings. Data analysis was performed with the three estimators presented above. Nonparametric estimators require particular settings to be chosen. In the case of I-splines, a knot sequence, corresponding to a vector of concentrations, needs to be defined. The knot sequence governs the number of splines used and the intervals on which each spline is defined. Based on a small pilot study, knot sequences were chosen to be 5 knots (for 50 subjects/arm) and 10 knots (for 100 subjects/arm) placed at percentiles of the observed concentration data. Type I error, power, bias in the predicted endpoint and weights were investigated for each estimator.

Table 7. Summary of simulations settings

Setting	**Values**
Noise level (standard deviation)	3.6, 7.5 and 15 ms (low/middle/high)
Sample size	50 and 100 subjects/arm
Drug effect model	linear, Emax, sigmoid Emax and quadratic
Drug effect size at $\gamma_{max,high}$	7, 8, 9 and 10 ms

Real data example

The real data consisted of a TQT study carried out in 239 male healthy volunteers according to the Declaration of Helsinki. The subjects received either a single oral dose of placebo, a therapeutic or a supra-therapeutic, 3-fold higher dose of the test drug. QT intervals were measured at three time points pre-dose (-25, -24 and -23h) and eight time points post-dose (1, 2, 3, 4, 5, 8, 12 and 24h). A total of 1912 post-dose observations were available. Heart rate correction had been done according to Fridericia's formula. Estimated QT prolongations at $\gamma_{max,high}$ were compared between the three estimators using 10 percentiles-based knots for the nonparametric estimator.

Model-averaged test for rheumatoid arthritis trials

Model-averaging was also applied in the context of efficacy confirmatory trials in rheumatoid arthritis. Considered trials were two-arm parallel studies, where American College of Rheumatology 20 (ACR20)[86] assessments were taken at 2, 4, 6, 8, 12, and 24 weeks. All patients were non-responders at the start of the study. They were treated either with a new or with a reference product (standard of care) during the entire study length. The endpoint to test was set to the proportion of ACR20 responders at week 24. Whether the new treatment was different from the reference, i.e. whether the responder rate difference between the two products was different from 0, was formalized via the hypotheses stated in Eq. 28.

$$H_0: p_1 - p_0 = 0 \quad versus \quad H_1: p_1 - p_0 \neq 0 \qquad \text{Eq. 28}$$

where p_1 is the responder rate in the treatment group and p_0 is the responder rate in the reference group. Estimates of p_1 and p_0 were obtained via classical analysis, single model analysis and model-averaged analysis as detailed in the next paragraphs. The null hypothesis was rejected (i.e. a difference concluded) when the 95% CI of the responder rate difference excluded 0.

Classical analysis

For the classical analysis, only the ACR20 data at week 24 was utilized. The ACR20 status was assumed to be a binomial variable. $\hat{p}_1$ and $\hat{p}_0$were estimated as $\hat{p}_l = Y_l/N_l$, where Y_l is the number of patients meeting the ACR20 criterion in group l and N_l is the number of patients in group l (l = 1 treatment, l = 0 reference product). The asymptotic 95% CI around the responder rate difference was computed using Eq. 29, where $z_{97.5}$ is the 97.5th quantile of the normal distribution.

$$\left[(\hat{p}_1 - \hat{p}_0) \pm z_{97.5}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{N_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{N_0}}\right] \qquad \text{Eq. 29}$$

Single model longitudinal analysis

The framework used for the longitudinal analysis of the ACR20 data was first-order Markov mixed-effects modeling[87]. The ACR20 status was described by two states, responder and non-responder (Figure 12). At each visit, individuals could move from one state to the other according to transition probabilities which depended on their current state.

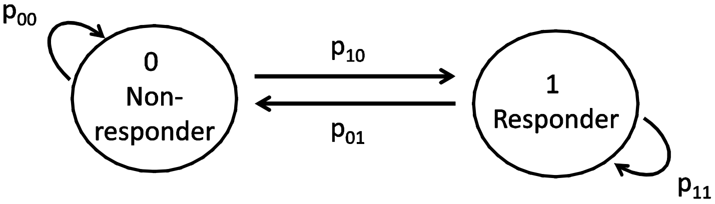

Figure 12. Markov model for the ACR20 response, adapted from Lacroix et al[87]. p_{00} corresponds to the probability of staying a non-responder, p_{10} of becoming a responder, p_{01} of becoming a non-responder, and p_{11} of staying a responder compared to the previous visit.

A generic model for the transition probabilities is displayed in Eq. 30.

$$\begin{cases} logit(p_{10,i,k}) = \beta_0 + \beta_1 TRT_i + (\beta_2 + \beta_3 TRT_i)\log(t_k - t_0) + \eta_i \\ logit(p_{11,i,k}) = \beta'_0 + \beta'_1 TRT_i + (\beta'_2 + \beta'_3 TRT_i)\log(t_k - t_0) + \eta'_i \end{cases} \qquad \text{Eq. 30}$$

where TRT is the treatment indicator (TRT = 1 treatment, TRT = 0 reference product), t_k are the visit times, β are fixed effects parameters, η_i and η'_i are subject-specific random effects defined to capture inter-individual variability in transition probabilities. η_i and η'_i are assumed to follow normal distributions with mean 0 and variances ω and ω', respectively. Transitions were only allowed to happen at t_k. p_{00} and p_{01} can be derived from p_{10} and p_{11}, as by definition $p_{00} = 1 - p_{10}$ and $p_{01} = 1 - p_{11}$.

Model parameters and their uncertainty were obtained through nonlinear mixed-effects modeling. The overall responder rates of the treatment and reference arms at week 24 were derived by integrating over the random effects using simulations. The 95% CI around the responder rate difference was obtained from the asymptotic variance-covariance matrix of the model parameters.

Model-averaged longitudinal analysis

The proposed model-averaged longitudinal analysis utilized a set of 10 models (Table 8). The models were variants of the generic model displayed in Eq. 30 and differed in the number of parameters estimated (from 4 to 16) as well as in the function describing the time course of the transition probabilities (log-linear, linear, quadratic or discrete time). The inter-individual variability model was identical for all models. BIC-based weights were used to compute the model-averaged estimate.

Simulation study

A high number of studies were simulated according to 12 scenarios, with each scenario corresponding to a different data-generating model. Data-generating models corresponded to the models included in the model-averaging pool, with the discrete time model used for three scenarios (Scenarios 10, 11 and 12). For each scenario, 1000 datasets were simulated using three sample sizes: $N_l = 100$, 300 and 1000 patients/arm. Model parameters were chosen to lead to realistic responder rates in the range of 50 to 75% at 24 weeks for the reference treatment[10]. Responder rates of the test treatment were chosen to be 0%, 5% or 10% higher than the reference treatment. Dropout was not included in the simulations. Type I error, power, bias in the predicted endpoint and weights were investigated for each analysis.

Software

NONMEM[46] 7.2 and 7.3 aided by the PsN software[43] version 3.5.2 and higher were used for data simulation and analysis in the work related to residual error modeling and parameter uncertainty. R 3.1.2[88] was used for data simulation and analysis in the work related to model-averaging. In the QT work, the *lm* function was used for the parametric estimator, and the *isb* function of the *SVMMaj* package as well as the *constrOptim* function were used for the nonparametric estimator. In the rheumatoid arthritis work, the *glmer* function of the *lme4* package with the Gaussian quadrature algorithm using 5 support points was used for model fitting. RStudio 0.98 using R 3.1.2[88] and Xpose[89] 4.3.3 and higher were used for graphical outputs.

Table 8. Description of the 10 models included in the model-averaged analysis and used for the simulation scenarios

Mod.	**Mathematical description**	**Name**	**N**
1	$logit(p_{10,i,k}) = \beta_0 + \beta_1 \times TRT_i + (\beta_2 + \beta_3 \times TRT_i)\log(t_k - 13) + \eta_i$ $logit(p_{11,i,k}) = \beta'_0 + \beta'_1 \times TRT_i + (\beta'_2 + \beta'_3 \times TRT_i)\log(t_k - 13) + \eta_i$	Full Markov	9
2	$logit(p_{10,i,k}) = \beta_0 + \beta_1 \times TRT_i + (\beta_2 + \beta_3 \times TRT_i)\log(t_k - 13) + \eta_i$ $logit(p_{11,i,k}) = (\beta'_2 + \beta'_3 \times TRT_i)\log(t_k - 13) + \eta_i$	Markov no intercept p_{11}	7
3	$logit(p_{10,i,k}) = \beta_0 + \beta_1 \times TRT_i + \eta_i$ $logit(p_{11,i,k}) = \beta'_0 + \beta'_1 \times TRT_i + (\beta'_2 + \beta'_3 \times TRT_i)\log(t_k - 13) + \eta_i$	Markov no time p_{10}	7
4	$logit(p_{10,i,k}) = \beta_0 + \beta_1 \times TRT_i + (\beta_2 + \beta_3 \times TRT_i)\log(t_k - 13) + \eta_i$ $logit(p_{11,i,k}) = \beta'_0 + \beta'_1 \times TRT_i + \eta_i$	Markov no time p_{11}	7
5	$logit(p_{10,i,k}) = \beta_0 + \beta_1 \times TRT_i + \eta_i$ $logit(p_{11,i,k}) = (\beta'_2 + \beta'_3 \times TRT_i)\log(t_k - 13) + \eta_i$	Markov no time p_{10} no intercept p_{11}	5
6	$logit(p_{10,i,k}) = \beta_0 + \beta_1 \times TRT_i + (\beta_2 + \beta_3 \times TRT_i)\log(t_k - 13) + \eta_i$ $logit(p_{11,i,k}) = logit(p_{10,i,k}) + \beta_4$	Logistic log-linear	6
7	$logit(p_{10,i,k}) = \beta_0 + \beta_1 \times TRT_i + \eta_i$ $logit(p_{11,i,k}) = logit(p_{10,i,k}) + \beta_4$	Logistic no time	4
8	$logit(p_{10,i,k}) = \beta_0 + \beta_1 \times TRT_i + (\beta_2 + \beta_3 \times TRT_i)\left(\frac{t_k - 14}{7}\right) + \eta_i$ $logit(p_{11,i,k}) = logit(p_{10,i,k}) + \beta_4$	Logistic linear	6
9	$logit(p_{10,i,k}) = \beta_0 + \beta_1 \times TRT_i + (\beta_2 + \beta_3 \times TRT_i)\left(\frac{t_k - 14}{28}\right) + (\beta_5 + \beta_6 \times TRT_i)\left(\frac{t_k - 14}{28}\right)^2 + \eta_i$ $logit(p_{11,i,k}) = logit(p_{10,i,k}) + \beta_4$	Logistic linear and quadratic	8
10	$logit(p_{10,i,k}) = \beta_{0,k} + \beta_{1,k} \times TRT_i + \eta_i$ $logit(p_{11,i,k}) = logit(p_{10,i,k}) + \beta_3$	Discrete time	16

Mod.: Model; N: number of estimated parameters; t_k: visit time in days; TRT=0 for treatment, TRT=1 for reference.

Results

Results will be presented in sequence for each of the three components of this thesis work: residual error modeling (Paper I), parameter uncertainty (Paper II-V) and model prespecification (Paper VI-VII).

Residual error modeling

Performance of dTBS on real data examples

Improvement of model fit using dTBS as judged by the likelihood ratio test was significant for all models, with dOFV ranging from -243 to -7 (Table 9). Most models displayed right-skewed residuals ($\lambda < 1$) and scedasticity between additive and proportional on the *untransformed* scale (Figure 13). dTBS parameters were estimated with satisfactory precision in the six models for which the covariance matrix was available. Results with FOCEI and SAEM were similar except for two models (paclitaxel and pefloxacin), for which λ estimates indicated higher skewness with SAEM than with FOCEI (λ = -0.6 versus 0.15 and -1 versus -0.8 respectively). The direction of the estimated skewness remained identical for the two methods.

Estimating λ and ζ simultaneously was better than estimating only one of these parameters for all investigated models but one. Improvement in model fit over the original model was seen for six out of 10 models with the Box-Cox transformation or the power parameter alone. Estimates of λ and ζ differed depending on whether they were estimated simultaneously or not. This confirmed that both parameters were needed to correct simultaneously for skewness and scedasticity.

Table 9. Estimated error parameters, associated standard errors (SE) and dOFV using the dTBS and the t-distribution approaches for the 10 real data examples

Model	Orig. error model	ε-shk (%)	λ (SE)	ζ (SE)	Sced. 1-λ+ ζ	dOFV dTBS[b]	ν	dOFV t-dist
ACTH/ cortisol[34]	comb[a]	2.9	0 (-) 0 (-)	0.68 (0.27) 0.47 (0.18)	1.68 0.53	-86	3	-28
Cladribine[36]	comb	15.8	-0.65 (1.2)	-0.92 (1.0)	-0.58	-20	5	-36[c]

Cyclophosphamide/ metabolite[37]	add / comb	5.4	0.85(-)
			0.86(-)
				0 (-)
				0 (-)
					0.15
					0.16
						-8.6	9	-2.6

Ethambutol[38]	comb	11.8	0.67 (0.21)	0.67 (0.16)	1	-43	3	-100[c]
Moxonidine PK[39]	add	11.6	1.5 (0.07)	1.6 (0.08)	1.1	-243	3	-400
Moxonidine PD[39]	add	11.4	-0.93 (-)	-1.1 (-)	0.84	-14	9	-25
Paclitaxel[40]	add	19.3	0.15 (-)	-0.25 (-)	0.6	-22	3	-7.4[c]
Pefloxacin[41]	prop	23.2	-0.79 (0.61)	-1.2 (0.58)	0.59	-21	4.7	-20
Phenobarbital[42]	prop	28.9	1.8 (0.44)	0.83 (0.23)	0.03	-7	∞	0
Prazosin[35]	prop	11.2	2.4 (0.17)	2.5 (0.16)	1.1	-100	3	-169

[a]additive component fixed; [b]presented dTBS results are those obtained with the FOCEI method; [c]standard estimation of ν impossible, estimated through likelihood profiling; shk: shrinkage; Sced.: scedasticity.

Other model parameters and related precision could change between the original and the dTBS model on the fixed effects level and/or the random effect level. Changes in goodness-of-fit plots were typically minor, with improvements in the distribution of CWRES, NPDE and IWRES residuals apparent for models with high dOFV drops. Influence diagnostics showed that 64% of individuals benefitted from dTBS within a dataset on average, with 14% contributing to the significant part of the OFV drop. The OFV sum over the cross-validation datasets was lower for dTBS than for the original model for all models but cladribine and pefloxacin, evidencing good prediction properties. dTBS parameter estimates were consistent between the cross-validated datasets. The two models showing worse predictive properties with dTBS were associated with high imprecision on the dTBS parameters. Runtimes were not markedly different between the original and dTBS models under identical estimation methods.

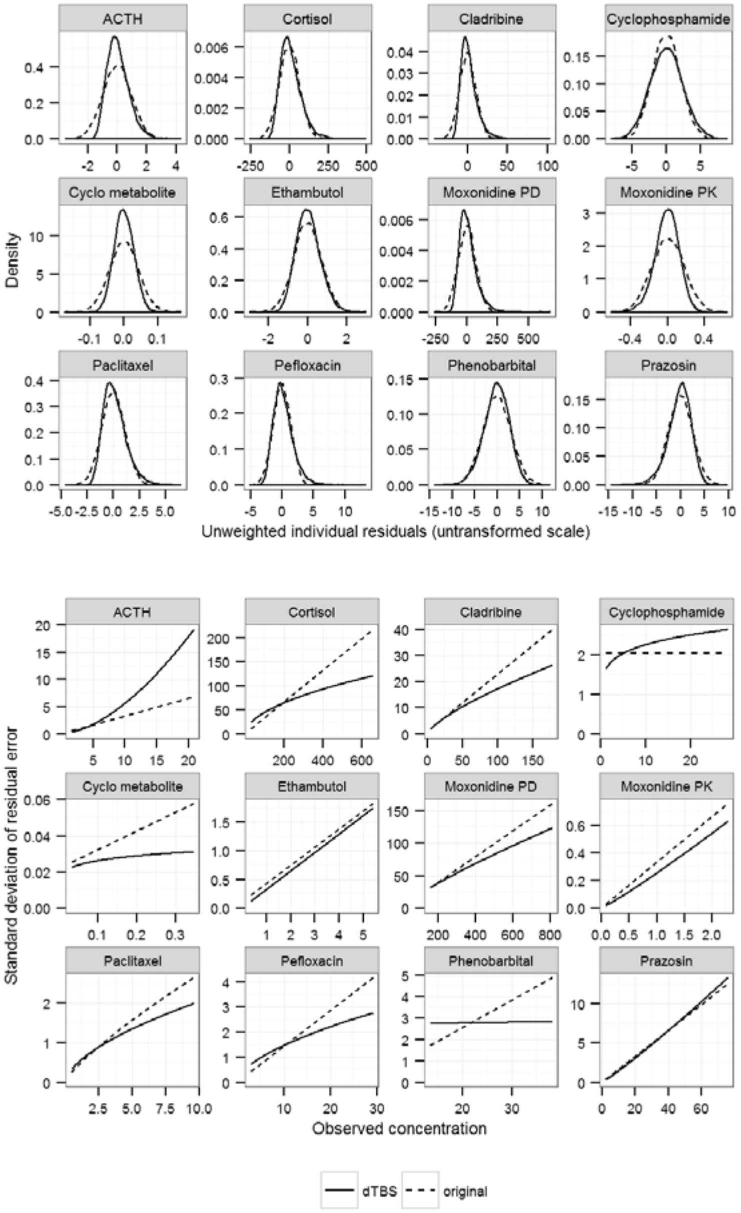

Figure 13. Simulated residual error distributions (top panel) and standard deviation of the residual error variance as a function of the observed data (bottom panel) on the untransformed scale for the original and dTBS error models for the 12 endpoints of the 10 real data examples. Dotted lines correspond to the original error model and solid lines to the dTBS error model. These distributions were obtained through simulations using the final dTBS/original estimates. In the top panel, the standard deviations of the distributions were calculated based on the medians of the observed data.

Performance of the t-distribution on real data examples

Using a t-distribution with estimated degree of freedom instead of a normal distribution for the RUV led to significant improvement in model fit for five models. When significant, dOFV were large (Table 9), ranging from -400 to -20. Estimated degrees of freedom ν ranged from 3 (the lowest possible value, corresponding to the highest heavy tails) to 200 (the highest possible value, very close to a normal distribution). The precision of the degrees of freedom were not available. For three models, the degree of freedom could not be estimated due to instability issues with the LAPLACE method. Changes in other model parameters were observed in a number of cases. The impact on individual plots was marked for some models, such as moxonidine PK. Improvements in the distribution of the residuals could be observed, as exemplified in Figure 14. Influence diagnostics showed that 71% of individuals within a dataset benefitted from the t-distribution on average, with around 29% responsible for the significant part of the OFV drop.

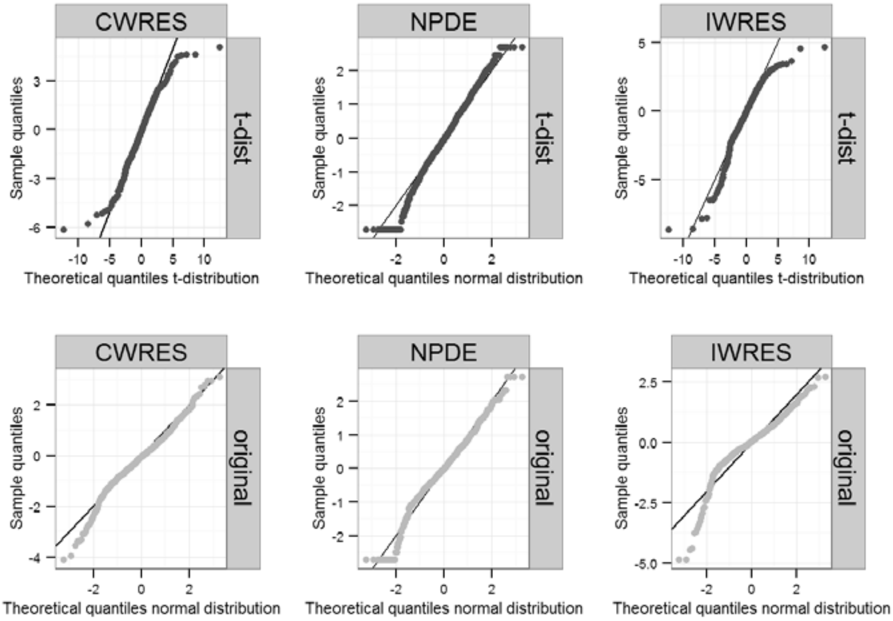

Figure 14 CWRES, NPDE and IWRES QQ-plots for the original and t-distributed error models in the prazosin example. Dark circles correspond to the final t-distribution model, light circles to the original model. Sample quantiles are compared to the theoretical quantiles of a standard normal distribution for the original model and to that of a standard normal distribution (NPDE) or a t-distribution with 3 degrees of freedom (CWRES, IWRES) for the t-distributed error model.

Simulation results

With dTBS, λ estimates were unbiased for the additive-on-log scenario with FOCEI and for the proportional scenario with SAEM. A downward bias of 0.13 remained for the additive scenario. The approximate scedasticity (1 - λ + ζ) showed no bias, even in the presence of bias in λ. Other model parameters were well estimated in all scenarios. Precision on the dTBS parameters was satisfactory (below 0.25) for the additive-on-log and proportional models, but poor (0.75) for the additive model. The type I error rate associated with the estimation of the dTBS parameters was always below the nominal level of 5%.

With the t-distribution, the estimated degrees of freedom of the t-distribution when simulated under normality tended towards the upper bound of 200, and the type I error rate associated with the estimation of ν was close to 0%.

Parameter uncertainty

All results pertaining parameter uncertainty will now be presented: the performance of the developed dOFV diagnostic in assessing uncertainty adequacy, the limitations of bootstrap in NLMEM, and the performance of the SIR methodology with examples of its application.

Performance of the dOFV uncertainty diagnostic

The dOFV diagnostic could detect the departure of an uncertainty estimate from the true uncertainty distribution based on differences in dOFV distributions. The use of the theoretical dOFV distribution as a reference instead of the SSE was appropriate in the two real data and the two simulation examples, as theoretical and SSE dOFV distributions were almost superimposed whichever dataset size (Figure 15 and Figure 16).

Evaluation of bootstrap adequacy in NLMEM

Bootstrap dOFV distributions deviated clearly from the theoretical dOFV distributions in the real data examples (Figure 15, left panel). Deviations were linked in part to sample size: data simulated based on these examples showed deviations from the theoretical dOFV distributions at identical sample size (Figure 15, middle panel), but not at the 8-fold increased sample size (Figure 15, right panel). The estimated degrees of freedom for the different dOFV distributions are provided in Table 10.

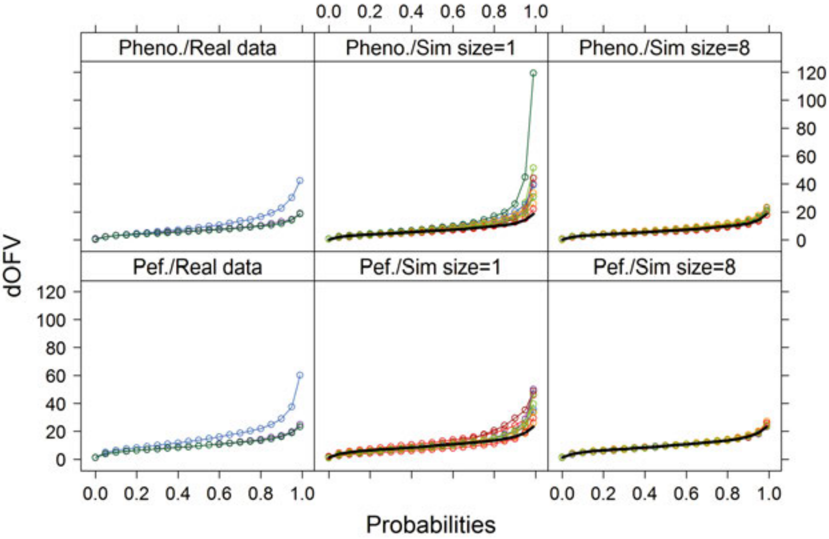

Figure 15. dOFV distribution plots for the two real data examples. Left panels provide bootstrap dOFV distribution for the real data (blue), the theoretical dOFV distribution (green) and the SSE dOFV distribution (pink). Middle and left panels provide bootstrap dOFV distributions for the simulated datasets of equal and 8-fold increased size (colors), as well as the theoretical dOFV distribution (black solid line). Pheno.: phenobarbital, Pef.: pefloxacin.

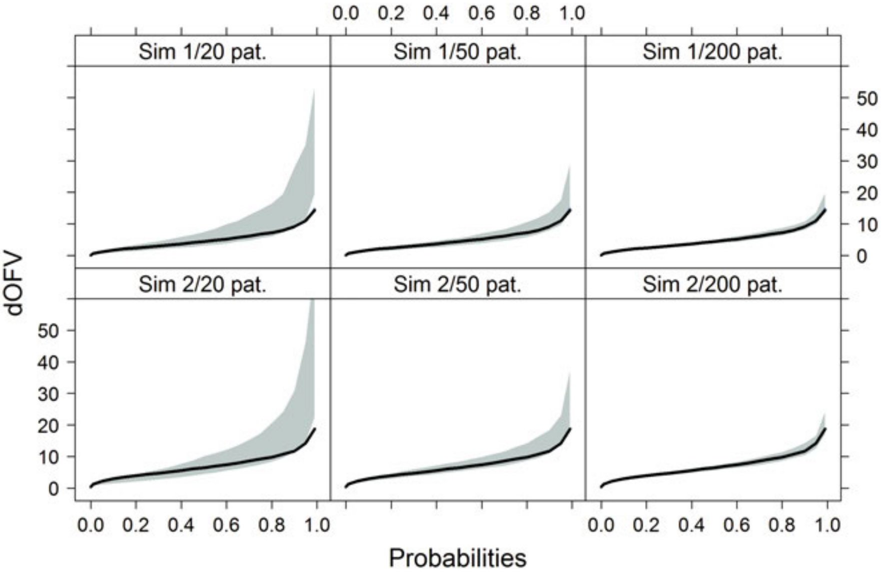

Figure 16. dOFV distribution plots for the two simulation examples. Grey shaded areas represent the range of dOFV curves for n=100 bootstraps, with the theoretical dOFV distribution superimposed (solid black line). One panel corresponds to one simulation example and dataset size. Sim: simulation, pat.: patients.

Similar trends were observed for the simulation examples: the range of bootstrap dOFV distributions decreased with increasing dataset size, and approached more and more the theoretical dOFV distribution (Figure 16). Differences in dOFV distributions could be linked to differences in CI based on the results of the simulations examples. Bootstrap coverage was always satisfactory for fixed effects, but deviations from the expected coverage were observed for random effects at the lowest sample size (coverage between 0.70 and 0.80 instead of 0.90) and to a lesser extent at the middle sample size (coverage around 0.85). IIV distributions with bootstrap appeared shifted down compared to SSE at low sample sizes (Figure 17). Looking at 95% CI, bootstrap underestimated medians by 20-25%, upper confidence bounds by up to 50% and lower confidence bounds by 5%.

Table 10. Degrees of freedom (median [range]) of the dOFV distributions for the real data and simulation examples

Real data	**Npar.**	**Df original**	**Df sim 1x**	**Df sim 8x**
Phenobarb.[42]	7	11.4	8.81 [6.86, 14.3]	7.35 [6.45, 8.45]
Pefloxacin[41]	10	16.5	11.3 [8.48, 14.0]	10.0 [9.78, 10.9]
Simulation	**Npar**	**Df 20-4**	**Df 50-4**	**Df 200-4**
Simulation 1	5	6.25 [4.32, 10.6]	5.48 [4.10, 7.15]	5.07 [4.42, 5.94]
Simulation 2	7	8.39 [5.74, 14.4]	7.55 [6.22, 9.93]	7.15 [6.24, 8.08]

Phenobarb.: Phenobarbital; Npar: Number of estimated parameters.

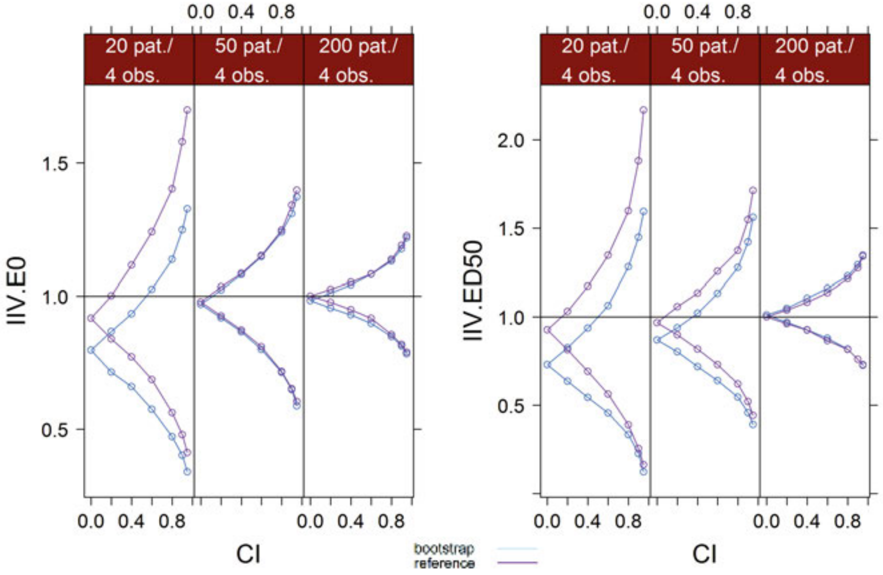

Figure 17. Comparative CI bounds of the bootstrap and of the reference (SSE) at different confidence levels for the IIV of the PD simulation example. Values were normalized by the true simulation value. Pat.: patients; obs.: observations.

Performance of the 1-step SIR on simulated data

In the simulated PK and PD examples, coverage with SIR was similar to coverage with the covariance matrix when using the latter as proposal distribution (Figure 18). Most parameters displayed suboptimal coverage at low sample sizes, especially IIV for which coverage rates were around 85% instead of 95%. Coverage was satisfactory at the highest sample size.

It was however apparent from the diagnostic plots that SIR settings were not fully appropriate: the proposal dOFV distribution was too narrow, i.e. below the reference chi-square, and the *M*/*m* ratio was too low, i.e. the temporal trends plots displayed downward trends. SIR settings using inflations of the proposal distribution by 3, 2 and 1.5 for sample sizes of 20, 50 and 200 respectively proved appropriate. SIR results using these settings were improved, with the coverage of IIV CL and IIV ED50 only remaining below its expected value at the lowest sample size.

Performance of the 1-step SIR on real data

For the three real data examples, the 1-step SIR procedure starting from the covariance matrix with an *M/m* ratio of 5 produced satisfactory results based on the developed diagnostics: the dOFV distributions of the SIR resamples were below the chi-square distribution and the temporal trends plots displayed no trends. Comparative parameter 95% CI obtained with the covariance matrix, SIR, bootstrap and LLP for the moxonidine example are presented in Figure 19. All methods provided similar CI for fixed effects with symmetric CI. The 95% CI of KA and lag-time (TLAG) varied between methods: they were narrowest and asymmetric with SIR and LLP, symmetric with the covariance matrix and widest and most asymmetric with bootstrap. Asymmetry was also marked for IIV and inter-occasion (IOV) parameters with SIR, LLP and bootstrap. RUV uncertainty was lowest for SIR and LLP, with a low degree of asymmetry for all methods. The phenobarbital and pefloxacin examples showed similar trends. In terms of runtime, the covariance matrix was the fastest method, followed by LLP, SIR and bootstrap (e.g. 14s, 15min, 1h and 2h respectively in the moxonidine case).

The real data examples were also used to investigate the impact of SIR settings, i.e. of *M/m* ratio and proposal distribution. The minimum *M/m* ratio necessary for SIR results to be considered final was different in the three investigated examples: it was found to be 6 for moxonidine, 4 for pefloxacin and 2 for phenobarbital. The necessary ratio was lower the closer the proposal distribution was to the chi-square distribution. The proposal distribution was found to have a profound impact on SIR results: inflations of the covariance matrix performed well, while deflations performed badly. Diagnostic plots with the deflated proposals showed proposal dOFV distributions

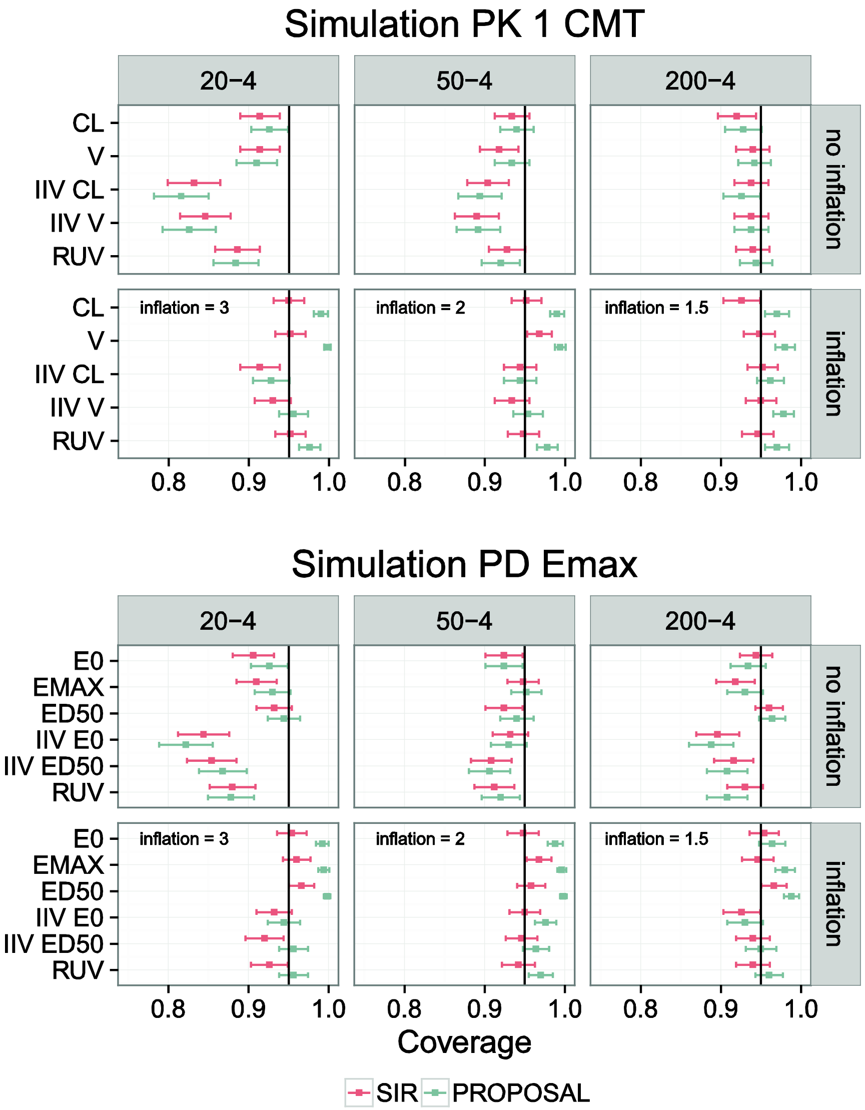

Figure 18. Coverage with SIR is as good as or better than coverage with the covariance matrix. The squares represent the observed 95% coverage for the parameters of the two simulation examples with SIR (red) and with the proposal distribution (green). The horizontal error bars represent the 95% CI around the observed coverage (500 simulated datasets per example and dataset size). SIR was performed both with the default workflow ("no inflation" panels: covariance matrix as proposal distribution and *M*/*m* = 5) and with an optimized workflow ("inflation" panels: covariance matrix inflated by 3, 2 and 1.5 as proposal distributions for the datasets with 20, 50 and 200 individuals respectively and *M*/*m* = 5)

below the reference distribution and an exhaustion of samples in the temporal trends plot even at the highest *M/m* ratio of 10. The developed SIR diagnostic plots were able to distinguish between appropriate and inappropriate settings in all cases. The degree of freedom of the dOFV distribution proved a good indicator of stable SIR results, as similar degrees of freedom corresponded to similar parameter RSE and CI bounds.

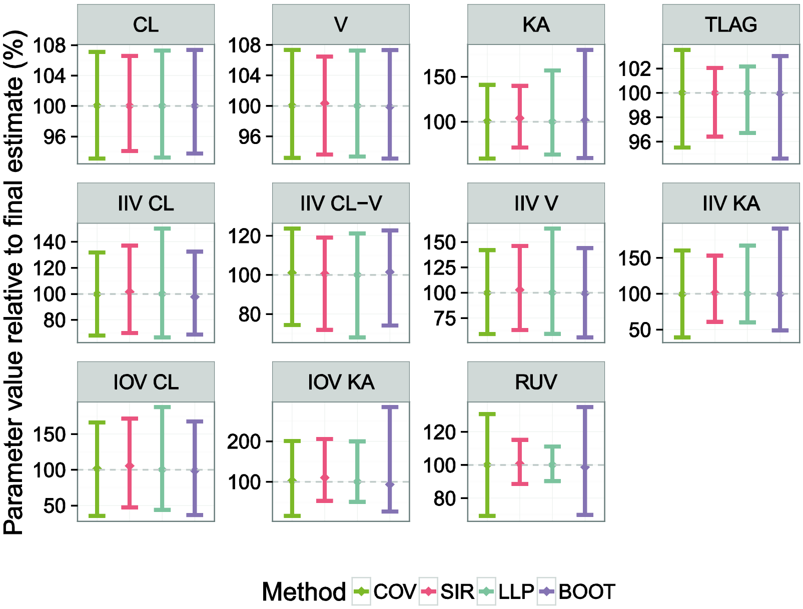

Figure 19. Comparative 95% CI of the moxonidine model parameters between four uncertainty methods: covariance matrix (COV, green), sampling importance resampling (SIR, red), log-likelihood profiling (LLP, blue) and bootstrap (BOOT, violet). Vertical error bars represent the 95% CI and the points represent the median of the uncertainty distributions. All random effects are on the variance scale.

Performance of the 5-step SIR on real data

Based on the results on SIR settings just presented, an improved 5-step SIR procedure was developed and tested on 25 NLMEM. For 20 models, SIR was started from the model's covariance matrix, and for 5 models SIR was started from a limited bootstrap (200 samples). For the models starting from the covariance matrix, inflation by 1.5 or above was required in half the cases, as indicated by dOFV distributions of the initial proposal distribution partly of fully below the reference chi-square. SIR convergence was achieved after 3 iterations on average (Figure 20). Two models (PD8 and PD15) needed 7 and 11 iterations to converge, respectively. One model (PD1) did not converge and displayed a degree of freedom oscillating above the total number of estimated parameters. Another model (PD11) stabilized at a degree of freedom above the total number of estimated parameters. Final degrees of freedom of the SIR dOFV distribution were on average 20% lower than the total number of estimated parameters. The median degrees of freedom of the proposal distributions were 1.3-fold higher than the total number of estimated parameters for the covariance matrices and 4-fold higher for the limited bootstraps.

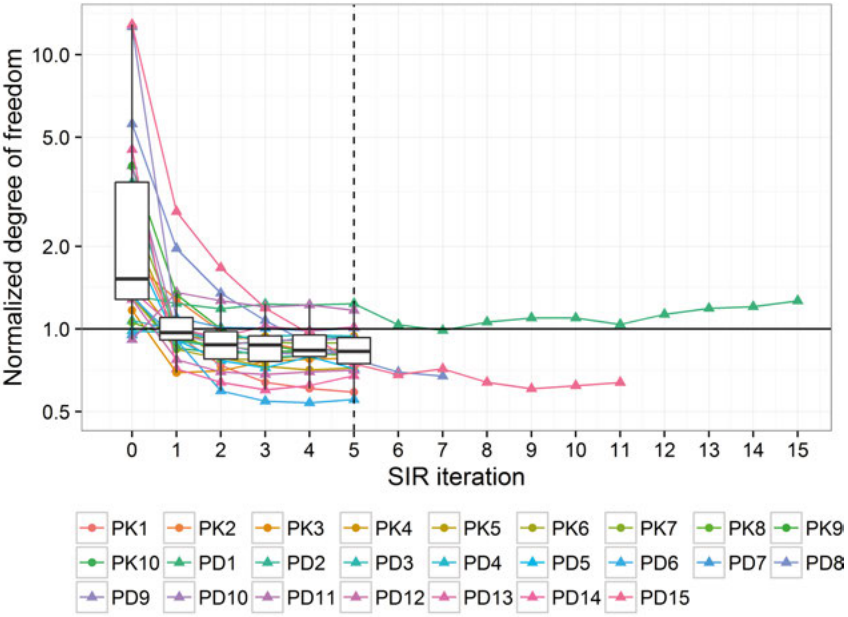

Figure 20. Convergence of the 5-step SIR over the 25 investigated models as represented by the estimated degree of freedom of the SIR resamples distribution at each iteration, normalized by the total number of estimated parameters of each model. The normalized degree of freedom at the 0th iteration is the degree of freedom of the informed proposal distribution (covariance matrix or limited bootstrap). Boxplots represent the median, first and third quartiles of the degree of freedom during the proposed 5-step procedure, of which the 5th and last iteration is indicated by the vertical dashed line. The horizontal line represents a degree of freedom equal to the number of estimated parameters.

SIR was generally robust to the initial proposal distribution: results starting the SIR procedure from a generic covariance matrix were similar to those starting from an informed proposal distribution. RSE and final degrees of freedom differed by less than 5% on average between the informed and the generic SIR. The greatest discrepancy was seen for the PK1 model, which showed an 8-point difference between final degrees of freedom of the generic and informed SIR. On average 8 iterations were needed for the generic SIR to converge.

Comparing the 5-step SIR to other uncertainty methods, SIR was 10 times faster than bootstrap on average. Note that a faster bootstrap implementation has been proposed[90], but was not applied here due to the current lack of experience with this implementation. Differences between uncertainty estimates obtained from SIR, the covariance matrix, bootstrap and SSE were highly model- and parameter-dependent. Median RSE and CI widths over all model parameters were similar between all methods but the bootstrap, which showed greater uncertainty (Figure 21). Regarding the shape of the uncertainty distributions, SIR displayed similar asymmetry to SSE, with CI being on average 20% longer on the side of the parameter distribution corresponding to lower values than on side corresponding to higher values. Asymmetry was highest with bootstrap and lowest with the covariance matrix.

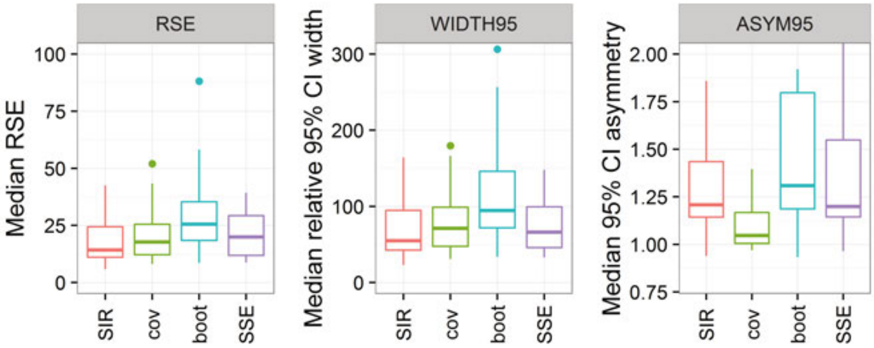

Figure 21. Distribution of the median (over all parameters) RSE, 95% CI width (WIDTH95) and asymmetry (ASYM95) for all models by uncertainty method: SIR, covariance matrix (cov), bootstrap (boot) and SSE.

Performance of SIR for decision-making using a WBPBPK model

The WBPBPK model previously developed in rats was successfully adapted to describe the data from the four other species. The three scaling models described the data fairly well, with differences mostly marked for the mice profiles and in the predicted variability of the data. Table 11 displays clearance-related parameters and their estimated uncertainty obtained with SIR for the three scaling models. Estimated RSE were similar between scaling models and were considered reasonable, so all three models were retained as plausible models to be used for extrapolation to human.

Median CMS and colistin plasma concentration-time profiles were predicted in human using Models A, B and C and were compared to the observed data (Figure 22). CMS predictions were in the same range with the three models, but their precision differed. They described the data well apart from some discrepancies at early time points, where both the maximum concentration and the time at which it occurs were underestimated. Predicted colistin profiles differed more markedly between models and were further away from the observed data. Models A and B overpredicted late concentrations, while Model C underpredicted late concentrations. Early concentrations were best described by Model B.

Table 11. CMS and colistin clearance-related parameter estimates by scaling model

	Model A		**Model B**		**Model C**	
Parameter* (unit)	**Typical value (RSE)**	**IIV% (RSE)**	**Typical value (RSE)**	**IIV% (RSE)**	**Typical value (RSE)**	**IIV% (RSE)**
Slope$_{\text{CMS}}$ (no unit)	1.07 (18)	113 (24)	1.2 (17)	98 (23)	0.88 (14)	130 (12)
Slope$_{\text{hyd-CMS}}$ (h^{-1})	0.153 (13)	-	0.134 (23)	37 (37)	0.135 (31)	93 (23)
EXP$_{\text{hyd-CMS}}$ (no units)	0.835 (4)	-	1.0 (12)	37 (37)	1.07 (18)	93 (23)
Slope$_{\text{nr-coli}}$ (h^{-1})	0.276 (16)	34 (26)	5.09 (19)	16 (34)	*mice* 12.7 (5) *rat* 1.24 (12) *rabbit* 0.514 (16) *baboon* 0.198 (12) *pig* 0.503 (17)	-
EXP$_{\text{nr-coli}}$ (no unit)	0.359 (15)	34 (26)	0.99 (6)	16 (34)	-	-

IIV: inter-individual variability; RSE: relative standard errors in %. *clearances are expressed as the product of a slope multiplied by a constant (volume or filtration rate) to a given exponent.

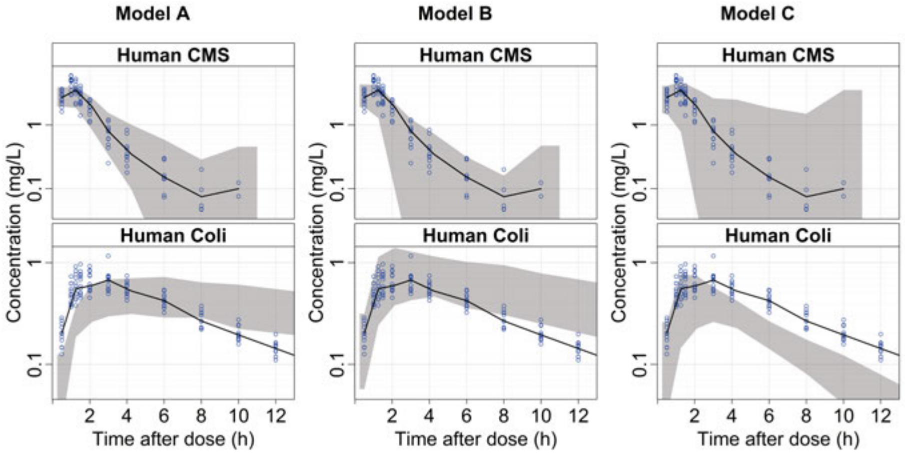

Figure 22. WBPBPK model predictions of CMS and colistin PK profiles in healthy volunteers receiving a single dose of CMS sodium 80 mg through a 1-h i.v. infusion. The blue circles represent the observed data, the black line the median of the observations, the grey shaded area the 95% CI around the median model predictions when simulations include IIV, RUV and uncertainty in the estimated parameters. Predictions and observations at 15 and 18h post-dose were omitted for visualization purposes.

Model prespecification

Results pertaining model uncertainty will now be presented. The performance of model-averaging approaches as fully pre-specified analysis methods for confirmatory studies was investigated first in the context of safety TQT studies and then in the context of efficacy in rheumatoid arthritis.

Model-averaged test for QT prolongation assessment

The proposed model-averaged test for QT prolongation assessment, based on the combination of a parametric linear and a nonparametric I-splines estimators using global MISE weights, led to satisfactory type I error control in the investigated scenarios (Figure 23). The parametric and the nonparametric tests also led to satisfactory type I error control. The nonparametric test displayed a percentage of rejections significantly below the nominal level of 5% in most cases, with the model-averaged test correcting only part of this conservatism. Type I error increases were observed for the model-averaged and/or nonparametric tests in two scenarios (Emax model, middle and low noise, 50 subjects/arm), which may be attributed to the knot selection of the nonparametric estimator and will be discussed later.

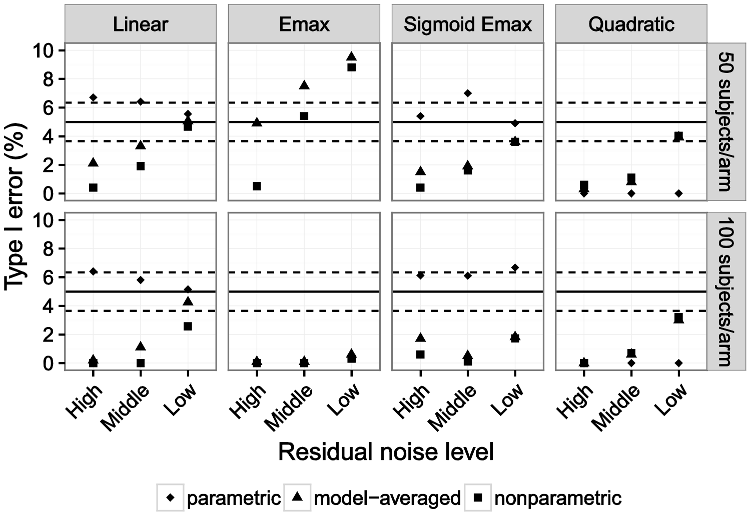

Figure 23. Type I error of the tests based on the parametric estimator, the model-averaged estimator, and the nonparametric estimator in the investigated scenarios. The solid black horizontal line represents the nominal level at 5% and the dashed black lines its 95% prediction interval for 1000 simulations ([3.6%; 6.4%]).

The power of the model-averaged test was at least as high as the power of the nonparametric test, and was considerably higher in some scenarios (Figure 24). Under the linear simulation model, the power using the model-averaged estimator was on average 14% higher than using the nonparametric estimator and 14% lower than using the parametric estimator, across all drug effects. Power gain was 6.2% on average under the nonlinear simulation models.

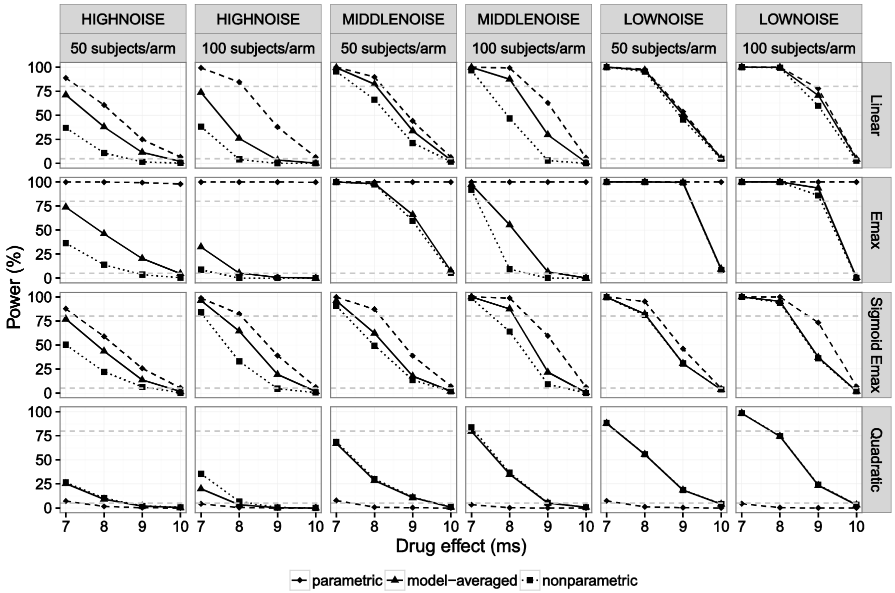

Figure 24. Power of the tests based on the parametric estimator, the model-averaged estimator, and on the nonparametric estimator in the investigated scenarios. Horizontal dotted lines represent 80% power (commonly desired power) and 5% (type I error at 10 ms).

The model-averaged estimator led to small upward, i.e. conservative, bias and good precision of the estimated drug-induced QT prolongation at $\gamma_{max,high}$. MISE weights attributed to the parametric estimator were highest under the linear simulation model (median around 0.5), moderate under the sigmoid Emax model (median around 0.25), and lowest under the Emax and quadratic models (median around 0.05).

The three estimators were also applied to the real data example. The estimated drug-induced QT prolongation was 2.2 ms (two-sided 90% CI [1.5; 2.8]), 2.5 ms (90% CI [1.9; 3.5]) and 2.7 ms (90% CI [2.0; 3.7]) based on the parametric, model-averaged and nonparametric estimators. From *Figure 25*, the adequacy of the linear model was questionable.

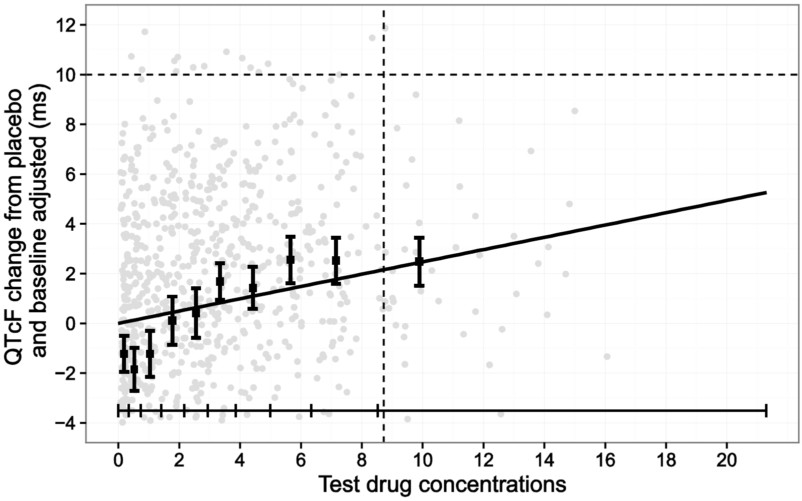

Figure 25. Concentration quantile–ΔΔQTcF plot for the real data example, with estimated QT prolongation using the linear estimator (solid black line). Grey points represent the data. Black squares with vertical bars denote the observed arithmetic means and 90% CI for the baseline and placebo-adjusted QTcF (ΔΔQTcF) within each concentration decile, plotted at the median concentration of the decile. The horizontal solid black line with tick marks shows the range of plasma concentrations divided into deciles. The horizontal dashed black line shows the 10 ms threshold and the vertical dashed black line shows the observed geometric mean $\hat{\gamma}_{max,high}$.

Model-averaged test for rheumatoid arthritis trials

The second model-averaging approach proposed for efficacy trials in rheumatoid arthritis, which consisted of 10 models for ACR20 response weighted using BIC weights, showed acceptable type I error rates overall (Figure 26). Type I error was slightly elevated in two cases (Scenario 5 and 9, 1000 patients/arm), one of which was due to simulation noise. Type I error was also elevated in one case (Scenario 5, 1000 patients/arm) and two cases (Scenario 2, 300 patients/arm and Scenario 9, 100 patients/arm) for the single model and classical analysis, respectively.

Power improvements over the classical analysis were marked for seven out of the 12 scenarios. Figure 27 displays the power for the three estimators under a true drug effect of 5%. The greatest power gains were observed for the model with the lowest number of parameters (Scenario 7), going from 64% with the classical test to 86% with the model-averaged test under a true responder rate difference of 5% and 1000 patients/arm.

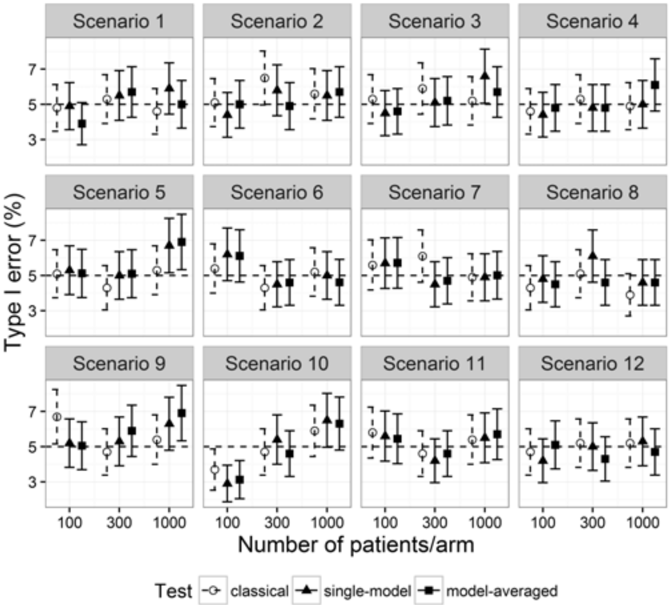

Figure 26. Type I error rates for the classical (empty circles), single-model (full triangles) and model-averaged (full squares) tests for all simulation scenarios. Vertical error bars represent the 95% CI around the type I error rates. The horizontal dotted line corresponds to the nominal level of 5%.

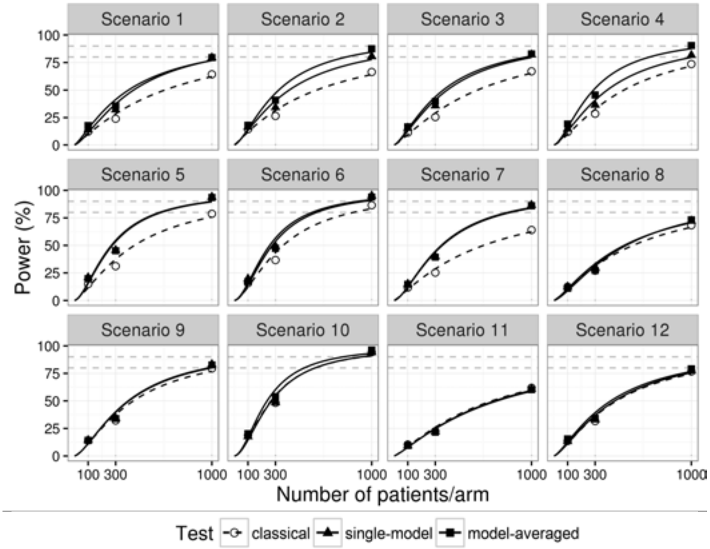

Figure 27. Power for the classical (empty circles), single-model (full triangles) and model-averaged (full squares) tests for all simulation scenarios under a true drug effect of 5%. The horizontal dotted lines correspond to commonly used power targets of 80% and 90%. Lines correspond to logistic regression predictions of the power based on log sample size for the classical (dotted line), single-model and model-averaged (solid lines) analyses.

The estimated responder rate differences at week 24 were unbiased, except for one scenario (Scenario 8) for which differences were conservatively underestimated by 3.3% and 2.2% at 100 and 300 patients/arm under a true difference of 10%. Model-averaging led to biased responder rates in some of the lowest sample size cases, with a bias between -4.2% and 1.1%. The bias was reduced with increasing sample sizes and was below 1% at the highest sample size for all scenarios. No bias was observed with the single model analysis. Model-averaging weights identified the data-generating model, at the latest at the highest sample size, in eight out of the 12 scenarios. In the four remaining scenarios (Scenarios 2, 4, 10 and 12), a simpler model was assigned higher weight. Increases in sample sizes moved the weight distribution towards the data generating model, but convergence speed was highly scenario-dependent.

Discussion

Discussion points will be presented in sequence for each of the three components of this thesis work: residual error modeling (Paper I), parameter uncertainty (Paper II-V) and model prespecification (Paper VI-VII).

Residual error modeling

The implementation of the dTBS and t-distribution residual error models for NLMEM was successful. dTBS is available as a PsN functionality which can easily be applied to models developed using classical error models without any modification of the model file. The t-distribution needs to be manually coded and makes use of the LAPLACE method, which remains a drawback due to the observed instability of this estimation method. These extensions of classical error models led to important improvements in model fit and a better agreement to the assumptions regarding the residual error model. They present the advantage of being capable of handling potentially skewed, heteroscedastic and outlying residuals without the need for predefined and subjective exclusion criteria.

It is recommended to apply dTBS and/or the t-distribution to models obtained through traditional model building as a means to improve the robustness of conclusions drawn from the model. dTBS could be favored under potential skewness in the residuals or trends in the scedasticity relationship, and the t-distribution under extreme outliers. These approaches could however be introduced earlier in model building and retained if leading to significant improvements. The error specific parameters should stay unfixed during model building in order to retain flexibility towards subsequent changes of other parts of the model.

Enhancing residual error model compliance will improve both the estimation of and the simulation from NLMEM. Parameter estimated using maximum likelihood may be biased if the wrong variance model is chosen[29], and inference using the computed likelihood or standard errors of parameter estimates may be invalidated[91,92]. Similarly, if the normality assumption is not verified, maximum likelihood estimation comes back to extended least squares[93] and resulting estimators and their uncertainty may not be appropriate[94]. With regards to simulations, the scedasticity will have a high impact on model predictions, and may be particularly important when simulating data outside of the range of the observed data. Ignoring skewness will often underestimate variability by simulating less extreme values.

Assessing improvement on the error model level

The OFV was used as the main criterion for model improvement under the new strategies. The power to detect improvements in the residual error model was expected to be high, as all observations contribute to this aspect of the model. Observed OFV drops using the dTBS or t-distribution approaches were indeed often important. On the other hand, model improvement was not easily assessed based on commonly used visual goodness-of-fit diagnostics specific or unspecific of the residual error. Specific diagnostics such as IWRES plots were often confounded by shrinkage. Nonspecific diagnostics such as CWRES or VPC were impacted by the interaction of the different levels of variability and often showed little overall change. Further diagnostics were thus investigated to assess individual influence and predictive properties of the models. For both methods, the majority of individuals benefitted from the new model, but often a low proportion of individuals contributed to the significant part of the dOFV drop. This was not unexpected given that only a limited part of these distributions actually deviates from the normal distribution. The superiority of the error model for the entire group of individuals was nevertheless confirmed by cross-validation in the dTBS examples.

Contrarily to other model parameters which often relate to pathophysiological processes, the residual error aggregates a multiplicity of factors such as endpoint type, study design, assay characteristics and model misspecification. As a consequence, while the interpretation of the error parameters is straightforward at the distribution level, they remain difficult to explain or anticipate in a given setting.

Interestingly, real data examples most improved with the new error models were similar between dTBS and the t-distribution. Both approaches allow individual observations to be further away from model predictions, the difference being that dTBS allows this to happen in one direction only (i.e. more negative or more positive residuals) whereas the t-distribution allows both directions (i.e. more negative and more positive residuals). The results showed that allowing some type of outlier was beneficial, even if the symmetry was misspecified. It is interesting to note that both approaches can also be combined. Observed gains at the level of the residual error would however need to be balanced with such added complexity.

dTBS specificities

Even if none of the models investigated were chosen because of a suspected misspecification of the error model, dTBS led to significant improvement in all cases. The impossibility to observe negative endpoints was consistent with most models displaying some degree of right-skewness. Left-skewness was observed in two rich data examples, possibly as a consequence of absorption model misspecification. Investigations keeping one of the dTBS parameters fixed showed that estimating λ alone corrected for scedasticity more than for skewness. This confirmed that the full dTBS approach estimating λ and ζ should always be used. An additional power term, which could account for "combined" error models, could be envisaged. However, the presented dTBS model is believed to provide sufficient flexibility for most applications, which was confirmed by the absence of model improvement when adding a second power term in the examples originally modeled with combined error models.

dTBS parameters could be estimated without bias and with satisfactory precision in the proportional and additive-on-log simulation examples using the SAEM and FOCEI methods respectively. The additive simulation scenario displayed bias on λ, which could not be corrected but which did not impact the type I error or the estimation of other parameter estimates.

t-distribution specificities

The t-distribution proved beneficial for five models. A major limitation of the application of this model remains its implementation using the LAPLACE method, which often led to minimization difficulties. The greatest OFV drops were observed for degrees of freedom close to the lower bound of 3. Individual fits could be largely improved by allowing isolated data points to depart more from the predictions.

In conclusion, the models developed in the first part of this work facilitate model-building decisions by providing unified, flexible RUV models. These models avoid the case-by-case testing of a limited number of traditional models, as well as subjective decisions on how to handle outliers. They can be used to simulate more realistic real-life data. At last, increased compliance to RUV model assumptions is expected to improve NLMEM properties in general.

Parameter uncertainty

After focusing on enhancing model compliance to RUV assumptions, we will now discuss the results concerning the diagnosis of, and improved methods for, parameter uncertainty estimation.

Performance of the dOFV uncertainty diagnostic

The dOFV diagnostic enables to assess whether a given uncertainty estimate can be considered adequate, based on whether its dOFV distribution is at or below the theoretical dOFV distribution. It can be applied to any method for assessing parameter uncertainty, provided parameter vectors can be drawn from the proposed uncertainty distribution. Given the importance of parameter uncertainty in decision-making, scrutiny towards uncertainty estimates should be enhanced by making the dOFV diagnostic an integral part of model assessment.

Two assumptions were made when using the theoretical distribution as a reference: that the dOFV distribution of the true uncertainty follows a chi-square distribution at the investigated sample sizes, and that its degree of freedom corresponds to the total number of estimated parameters in NLMEM. These assumptions were met in the investigated examples, as the SSE distributions overlaid the theoretical distributions in all cases. It remains highly questionable whether the degree of freedom would always be equal to the number of estimated parameters, notably for more nonlinear or more constrained NLMEM (e.g. with parameters bounded by physiological values or inclusion criteria). Lastly, note that the dOFV diagnostic is a global test; it does not indicate for which parameter(s) the uncertainty is not well described. For the bootstrap, parameter-specific diagnostics based on “effective” sample sizes may be useful, as will be detailed in the next subsection.

Evaluation of bootstrap adequacy in NLMEM

Based on the developed dOFV diagnostic, bootstrap proved unsuitable for the two investigated real examples and the simulation examples at low sample size(s). The simulation examples showed that the increase in degree of freedom of the dOFV distribution could be linked to too narrow CI. For example, a 1.25-fold increase in degree of freedom translated into coverage rates of 70% instead of 90% for IIV parameters. As expected, bootstrap adequacy increased with increasing number of individuals. Model misspecification also contributed to the higher than expected degree of freedom: the degree of freedom obtained with data simulated using the same design was increased to a lesser extent than with the real data. Generalization of these results is limited by the investigated models, which were relatively simple models featuring a high proportion of random effects. Stratification on the number of observations per individual could have improved bootstrap results in the real data examples. However, stratification would not have been straightforward due to the heterogeneous distribution of the number of observations per individual, and could have led to too small subgroups. A further limitation of the bootstrap was highlighted in the pefloxacin case, for which half the samples displayed estimation problems. dOFV distributions as well as uncertainty estimates differed when ignoring problematic runs, showing the sensitivity of bootstrap results to the choice of set-up regarding the stratification strategy and the computation of the CI.

The simulation examples enabled to pinpoint the parameters which uncertainty was not well captured by comparing bootstrap CI to SSE CI. IIV uncertainty appeared underestimated at low sample sizes, whereas the uncertainty of fixed effects was well described.

Similar increases in degree of freedom were observed for datasets with 20 individuals and datasets with 70 individuals. Diagnosing *a priori* in which cases bootstrap is inadequate based on sample size proved difficult for NLMEM due to the amount and heterogeneity of information contained in different individuals. Sample size does not reflect heterogeneity, which arises from unbalanced designs or different individual characteristics such as covariates or subject-specific random effects. The number of individuals also does not relate to model complexity: using the same dataset, bootstrap may be adequate for a simple model, but not for a much more complex model. An *a posteriori* method based on parameter-specific "effective" sample sizes was thus developed in this work as a better indicator of bootstrap adequacy. The effective sample size represents how many individuals with perfect information the estimated uncertainty for one parameter corresponds to. Effective sample sizes *N* were calculated based on the formulas for the standard errors of means (Eq. 31) for fixed effects and the standard errors of variances (Eq. 32) for random effects.

$$SE(\bar{X}) = \frac{SD(X)}{\sqrt{N}} \quad \rightarrow \quad N = \frac{VAR(X)}{VAR(\bar{X})} \qquad \text{Eq. 31}$$

$$SE(VAR(X)) = VAR(X)\sqrt{\frac{2}{N-1}} \quad \rightarrow \quad N = 2\left(\frac{VAR(X)}{SE(VAR(X))}\right)^2 + 1 \qquad \text{Eq. 32}$$

Effective sample size for fixed effects and IIV are expected to be at maximum the total number of individuals in the dataset. For IOV, the effective sample size is at maximum the total number of occasions (i.e. the sum of the number of occasions per individual) minus the total number of individuals. For RUV, *N* can be at maximum the total number of observations minus the number of individuals and minus the sum of the number of occasions per individual. Effective samples sizes for fixed effects and IIV in the real data examples were at maximum 30, i.e. less than half the total number of individuals (Figure 28).

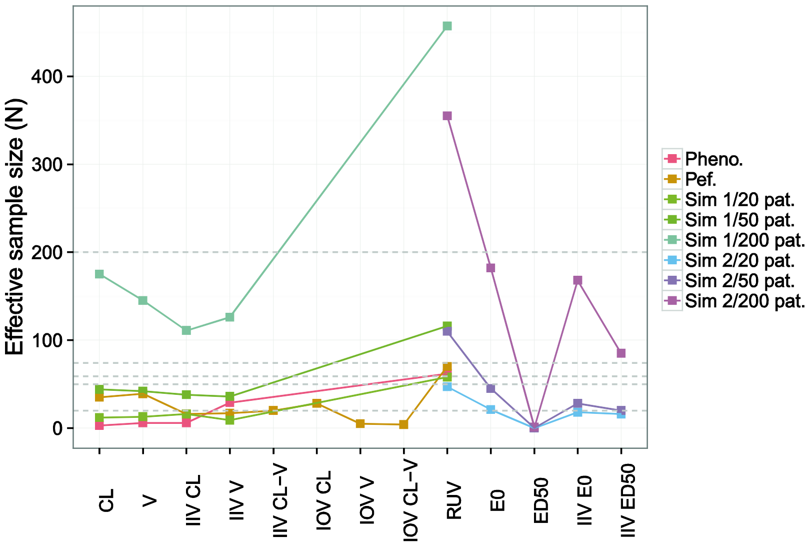

Figure 28. Effective sample sizes calculated for selected model parameters of the real data and simulation examples based on bootstrap uncertainty estimates. Colors correspond to the different examples and dataset sizes. Grey dashed lines correspond to the total number of individuals in the different examples. Pheno.: phenobarbital, Pef.: pefloxacin, Sim: simulation, pat.: patients.

Effective sample sizes were on average equal to 0.75-fold the possible sample size in the simulated PK examples (range: [0.45; 1]), and to 0.59-fold the possible sample size in the simulated PD examples (range: [0; 1]). Based on the corresponding coverage results, this could indicate that the minimum number of effective individuals needs to be at least 45 for bootstrap to be adequate. It is important to point out that the concept of effective sample sizes was developed here in an exploratory manner and more work is needed for this to be used as a decision criterion.

Performance of the 1-step SIR on simulated data

The observed bootstrap inadequacy triggered the development of SIR to improve parameter uncertainty estimation in NLMEM. The initially developed 1-step SIR procedure showed satisfactory coverage when starting from inflations of the covariance matrix. Necessary inflation factors decreased with dataset size, confirming that the adequacy of the covariance matrix increased with increasing sample size. SIR diagnostics evidenced that the covariance matrix often underestimated parameter uncertainty with simulated data, contrarily to the real data examples. Underestimation of parameter uncertainty by the covariance matrix had been observed previously[95]. The final SIR dOFV distributions overlaid the chi-square distributions in the simulation examples, which could indicate that the discrepancies observed between the SIR and the theoretical distributions in the real data examples were linked to model misspecification. However, the lower proportion of random effects in the simulations could also contribute to the SIR degree of freedom being equal to the number of estimated parameters.

Performance of the 1-step SIR on real data

The interpretation of differences between parameter uncertainties obtained with various methods on real data is not straightforward as the truth remains unknown. Thorough comparisons of available methods such as in Donaldson[96] are lacking in NLMEM. SIR was able to detect expected uncertainty asymmetry for variances and nonlinear parameters. SIR results were closest to LLP, which was expected as both are based on the likelihood of different parameter vectors on the original dataset. However, it is difficult to use LLP for simulation as it does not provide full uncertainty distributions. Bootstrap results confirmed the presence of asymmetry. Bootstrap CI were generally wider than SIR. Bootstrap uncertainty has however been shown earlier to be inadequate for the phenobarbital and pefloxacin examples. SIR was thus found to provide an adequate estimate of parameter uncertainty in the investigated real data examples. In terms of runtime, SIR was about twice as fast as bootstrap, but runtime gains are not easily generalizable as they will depend on the adequacy of the proposal distribution and the runtime difference between OFV estimations and OFV evaluations.

Performing SIR on the real data examples varying *M/m* and/or using inflations or deflations of the proposal distribution enabled to understand the impact of SIR settings and improve their selection. A *M/m* ratio of 5 was sufficient in the investigated cases, and could even have been further reduced for two examples, leading to faster runtimes. The developed diagnostics enabled to assess whether *M/m* was sufficient *a posteriori*. However, no robust quantitative relationship could be established to assess the necessary *M/m a priori*, for example using the difference in degrees of freedom between the proposal and the reference dOFV distributions. Starting from too narrow distributions proved problematic for SIR, as the limited number of samples in the tails of the distribution makes the expansion of uncertainty very slow. This issue could however be easily identified in the diagnostics. It is thus recommended to inflate the proposal distribution instead of increasing *M/m* when the proposal appears too narrow. It is important to note that even if only variants of the covariance matrix were used as proposal distributions here, a major advantage of SIR is that it can be used with any multivariate parametric distribution, i.e. also for models for which the covariance matrix is not available. A limited number of bootstrap samples or a generic covariance matrix can easily be used as proposal distribution. Another dimension of any multivariate distribution is the correlation between the distributions. Similarly to the issue observed with too narrow proposal distributions, SIR could be inefficient at reducing misspecified high correlations, and the consequences of such misspecifications remain unclear[97]. This could not be investigated here as high correlations were not observed in the investigated examples. It is however advised to reduce correlations when performing SIR, or ideally to use parametrizations that minimize such issues[98]. Lastly, the ultimate diagnostic to test whether SIR results are final would be to perform a second SIR using the SIR resamples as proposal distribution. Identical proposal and resamples distributions would confirm that SIR results are final. This thought triggered the development of the iterative 5-step procedure discussed below.

Performance of the 5-step SIR on real data

The 5-step SIR procedure starting from the covariance matrix or a limited bootstrap was satisfactory for 22 out of the 25 NLMEM investigated. SIR was globally robust to the choice of initial proposal distribution and thus it is recommended to use an informed proposal distribution for runtime gain. Inflation was needed in about half the cases when starting from the covariance matrix. This was partly due to the use of the multivariate normal distribution for the first iteration, which was expected to lead to a suboptimal description of the uncertainty of random effects. The use of a multivariate Box-Cox distribution allowing for parameter-specific asymmetry relaxed symmetry constraints in subsequent iterations. Limitations of the Box-Cox distribution to approximate nonparametric parameter vectors were sometimes apparent, but it did not seem to hamper SIR efficiency in the investigated examples. It should be noted that multivariate parametric distributions impose constraints on the correlation level. This may be problematic in the case of high correlations, as it might be difficult for SIR to move away from them if they are misspecified. Further refinements of the SIR procedure, including more flexible parametric distributions, correlations structures, or sampling strategies[99] could be envisaged to further improve SIR performance.

SIR proved particularly useful for identifying local minima and estimating uncertainty in the presence of priors. Limitations of SIR were apparent for a model displaying instability in the likelihood estimation and a model featuring an on/off parameter, for which the degree of freedom stabilized above the number of parameters whichever proposal distribution was used. A 31-parameter PK model containing 9 IIV parameters also displayed diagnostic plots which failed to identify the exhaustion of samples for IIV parameters.

SIR provided median RSE and CI widths relatively similar to the covariance matrix and SSE, which supported the validity of the developed procedure for uncertainty estimation. The fact that SIR provided asymmetry estimates close to SSE showed its improvement over the covariance matrix, which performed well in terms of uncertainty magnitude (RSE ad CI width) but not symmetry. Bootstrap also performed well at describing the shape of the uncertainty. However, bootstrap led to uncertainty magnitudes markedly higher than the other methods, potentially overestimating variability due to suboptimal stratification. Estimation issues were relatively common for bootstrap and SSE, which in effect restricts the use of these methods. SIR seemed to perform better than the other methods based on the estimated degrees of freedom. Efforts to link the final degree of freedom with model characteristics such as sample size or the proportion of random effects were not successful.

Performance of SIR for decision-making using a WBPBPK model

WBPBPK models play an increasing role in drug development[100,101] and the developed model was the first interspecies WBPBPK model for a polymyxin antibiotic. Interspecies scaling[102,103] was performed using literature data to account for differences in physiological parameters, and approaches based on allometric scaling were used for the drug-dependent clearance parameters $CL_{r\text{-}CMS}$, $CL_{hyd\text{-}CMS}$ and $CL_{nr\text{-}coli}$. The three evaluated scaling models for $CL_{nr\text{-}coli}$ based on volume (Model A), volume and maximum lifespan potential (Model B), and estimation of species-specific $CL_{nr\text{-}coli}$ (Model C) performed similarly well for the five animal species. The interspecies scaling led to more adequate predictions for CMS than for colistin. Colistin elimination pathways and protein binding in tissue, which remain poorly understood, appeared to involve processes not adequately described by allometry. Parameter uncertainty obtained with SIR was reasonable for all three models, which were thus retained as potential scaling candidates for the extrapolation to human. It is interesting to note that SIR was particularly indicated in the present case, as the covariance matrix was expected to perform badly due to the low amount of information contained in the data. Bootstrap could not be applied as the uncertainty on parameters related to priors would have been underestimated, and LLP would not have been able to simulate parameter vectors needed for the extrapolation to human.

Using the interspecies WBPBPK model to predict human CMS and colistin plasma concentrations taking parameter uncertainty into account stressed the strengths and limitations of the model. Apart from mispredictions in the very early profiles, median CMS profiles in human were relatively well described with the three models, which was not unexpected given the good scalability of CMS between animal species. Median colistin profiles in human were less well described. Prediction differences between Models A, B and C were more marked for colistin than for CMS. Colistin concentrations were predicted in the right range, but changes in concentrations over time were either too rapid or too slow, which was attributed to a misprediction of $CL_{nr\text{-}coli}$. Despite none of the alternatives being fully adequate, Model B presented the best interspecies scaling properties overall. This model could be used to optimize lead selection of new polymyxin-like antibiotics based on PK properties such as plasma and tissue concentrations for example. It could be further refined as additional data becomes available, in particular on colistin elimination.

In conclusion, the work on parameter uncertainty enables to decide how much an uncertainty estimate can be trusted and provides a new method capable of better quantifying the uncertainty linked to a given decision based on NLMEM. SIR is applicable to many situations, including cases where other methods fail such as small datasets, highly nonlinear models or meta-analysis[104]. SIR can support model-building decisions and enable the development of more complex models for which uncertainty could not have been obtained otherwise. SIR can provide better decisions based on the final model and endpoints, and is useful for clinical trial simulations where parameter uncertainty plays a major role.

Model prespecification

The last aspect of uncertainty which was addressed in this work was model uncertainty. Two model-averaging approaches were developed and applied to safety and efficacy settings of QT prolongation assessment and efficacy in rheumatoid arthritis.

Model-averaged test for QT prolongation assessment

The model-averaged test based on a parametric linear and a non-parametric I-splines estimators weighted by global MISE weights showed operating characteristics globally appropriate to be used in the considered TQT confirmatory settings. The type I error was satisfactory in a set of realistic simulation scenarios, and power gains compared to the nonparametric estimator were observed. The behavior of the model-averaged estimator was impacted by three factors: the monotonicity constraint, the selection of knots of the nonparametric estimator, and the weighting.

The nonparametric test appeared very conservative, with type I errors often close to 0. Removing the monotonicity constraint on the nonparametric estimator led to type I errors much closer to the nominal level. However, allowing non monotonically increasing estimators is not possible for real data in the considered TQT safety settings, as the estimated QT prolongation at all concentrations below the expected maximum concentration $\gamma_{max,high}$ has to be smaller than the prolongation estimated at $\gamma_{max,high}$.

The second driver of the results proved to be knot selection, which remains an area of research for I-splines. Other nonparametric estimators such as smoothing splines[105] provide more easily optimized settings, but were not as easy to constrain and to use under the specified correlation structure. The knot strategy used here performed differently across scenarios, as evidenced by the fact that type I error and bias did not always decrease, and power did not always increase, over increasing information level. Suboptimal knot selection was mostly conservative, but could also lead to type I error increases in isolated cases. Observed type I error elevations with the model-averaged test were decreased when optimizing knot selection, from 9.5% to 5.8% in one scenario using 8 knots instead of 5 (Emax, low noise, 50 subjects/arm), and from 7.5% to 4.9% in one other using 7 knots instead of 5 (Emax, middle noise, 50 subjects/arm). Ideally, an automated algorithm for knot selection should be developed, but in real data settings exploratory simulations under plausible models *a priori* combined with sensitivity analyses *a posteriori* could suffice for satisfactory knot selection.

The last factor impacting the behavior of the model-averaged estimator was the weighting. MISE global weights were chosen because of their theoretical properties of converging to 1 in probability when the parametric model is true and to 0 when it is not. Convergence seemed slow, as median weights attributed to the linear model were around 0.50 and did not increase much over sample size. This limited the gain in power that could be achieved by the model-averaged estimator. Faster converging weights such as BIC were investigated, but they failed to disqualify the linear model in some of the nonlinear scenarios, leading to poor type I error control.

A particularity of the method was the prespecified structure of the covariance matrix of the residual error between individual QT measurements. This assumption appeared reasonable for the considered real data example. More complex covariance structures could however be envisaged.

Lastly, it should be stressed that the proposed method is based on concentration-response analysis and as such is not limited to TQT studies. It is also applicable to early phase data such as single and multiple ascending dose studies, which may replace TQT studies for QT prolongation assessment in the near future[106,107].

Model-averaged test for rheumatoid arthritis trials

The model-averaged analysis utilizing a pool of 10 Markov-type models weighted by BIC proved a valid alternative to the classical end-of-trial analysis for superiority testing in rheumatoid arthritis. Type I error was controlled for 34 out of the 36 simulation cases (12 scenarios of three sample sizes each) with the classical and model-averaging approaches, whereas it was controlled for 35 cases for the single-model analysis. This being a simulation-based assessment, we expected one of the 36 tested cases to provide a higher than expected type I error. The 2% elevation in type I error for Scenario 9 at 1000 patients/arm (at 7% instead of 5%) could not be explained after running additional simulations.

The model-averaged test led to important increases in power over the classical analysis for Scenarios 1-7. Relative power gains were 30% on average. They differed heavily between models, the models with the lowest number of parameters typically showing the highest power gains. Power gains were often similar between the single-model and the model-averaged tests. Greater power with the model-averaged test was seen when a model simpler than the data-generating model featured high weights, which is a potential advantage of model-averaged procedures as long as it is not achieved at the expense of detrimental bias. The absence of power gains for Scenarios 10-12 was not surprising, as the data-generating models were not continuous functions of time. However, the reasons for the absence of power gains for Scenarios 8 and 9 remain unclear.

At high sample sizes, BIC weights identified the true data-generating model (by attributing the highest weights to this model) in eight out of 12 scenarios. Both model structure and model parameter values influenced the attributed weight. The inclusion of Model 10 (categorical time) as a safeguard mechanism against model misspecification was not always guaranteed. This model came with many parameters and thus a high penalty, which could only be overcome if the fit of all other models was very poor. The choice of BIC was motivated by its link with the probability of the model being correct given the observed data. Model averaging can also be performed using other weights, such as the Akaike Information Criterion (AIC)[80]. AIC weights may lead to greater power gains due to the selection of simpler models but do not converge to the data-generating model, which was why BIC was preferred here.

Together with the weighting strategy, the selection of the type and the number of models to include in the pool plays a major role in model-averaging. Published models of ACR20 response consist of Markov models[87], logistic regression[108] and latent-variable approaches[109]. First order Markov models, some falling back to logistic regression, were used here. Extensions of the model pool to second order Markov elements or latent-variable models could be envisaged to span a greater range of models and further reduce the risk of model misspecification. Regarding the number of models to include, a balance need to be struck between spanning a wide range of possible models to avoid model misspecification and limiting the number of models to maximize power[110] and ease of use. Published model-averaging analyses typically include 2-20 nested or non-nested models[111-114]. Unfortunately, few investigations on the impact of model pool exist in literature[113]. As a consequence, the set-up of model-averaging methods remains a case-by-case matter. The models included here were thought to cover a wide enough range of both plausible and stress-case structures while avoiding overparametrization.

In conclusion, the fully prespecified model-averaged analyses developed in the last part of this thesis enable appropriate hypothesis testing for two types of confirmatory trials, for which a need for more efficient methods is present. Observed power gains translate into a reduction in the number of patients needed to perform such trials. Because type I error control often cannot be theoretically demonstrated for NLMEM, it is understood that from a regulatory perspective the acceptance of such methodologies will require case-by-case extensive simulations. The results presented here can nevertheless inform the development of similar approaches in other settings with regards to the key questions of the selection of models and the choice of weights.

Conclusion

In this thesis, different aspects of pharmacometric model-based analysis were scrutinized in order to enhance the use of pharmacometric models for decision-making in clinical drug development. Methods improving the description of the residual error model and the evaluation of parameter uncertainty were developed. These methods extended currently available tools with flexible models particularly suited for NLMEM. Model-averaging approaches were developed to address model uncertainty and comply with regulatory requirements regarding the prespecification of analysis methods in confirmatory settings. The performed work facilitates the application of pharmacometric analysis to key decision points such as confirmatory trials, and could hence improve the efficiency of drug development.

In particular:

- Residual error models were extended to be able to account for skewed, heteroscedastic and heavy-tailed residuals. The dTBS and/or t-distribution strategies are easy to use and increase model compliance to distributional assumptions, thus improving model appropriateness.

- The dOFV distribution plot was developed as a diagnostic for the adequacy of parameter uncertainty estimates. Based on this diagnostic, the performance of bootstrap appeared limited at sample sizes common in NLMEM.

- An alternative method for estimating parameter uncertainty, SIR, was developed. SIR was applied to a wide range of models and data and proved a powerful and easy-to-use method for uncertainty estimation in NLMEM.

- Two fully prespecified methods based on model-averaging were developed, one for the assessment of QT prolongation and one for efficacy in rheumatoid arthritis. The proposed methods provided satisfactory type I error control and higher power than the standard methodologies.

Acknowledgments

The work presented in this thesis was carried out at the *Department of Pharmaceutical Biosciences, Uppsala, Sweden* and at the *Pharmacometrics Group, Novartis, Basel, Switzerland*. I am grateful for support from the *DDMoRe* and *IDeAl* initiatives, which funded major parts of the research in this thesis. I would also like to thank *Apotekarsocieten* and *Småland nation* for traveling grants to various courses and conferences around the world.

These 5 years were not only full of work, but also, and very importantly, full of great people whom I feel incredibly lucky to have met:

Prof. Mats Karlsson, my main supervisor. It was an honor to work with you, from the early weekly meetings trying to write down whatever you said so that I could understand it after having thought about it for some days, to the later sessions brainstorming new ideas and projects. What you built and continue building in pharmacometrics is truly impressive, and I am very thankful that you enabled me to be a part of it.

Dr. Martin Bergstrand, my co-supervisor. You got on board for a project that never happened in the end, thank you so much for staying nevertheless! Thank you for keeping things focused, I learned many things from you.

Dr. Didier Renard, my "unofficial" supervisor at Novartis. Thank you for bringing great projects, for your availability despite all your other responsibilities and for your patience towards my pharmacist mind trying to keep up with statistics. You are a type of scientist we should strive to be.

Dr. Günter Heimann, thank you for bringing me in for the QT project. I know you think we disagreed many times, but I learned a lot from you questioning my work. It was great working with you.

I would also like to thank: *Prof. Margareta Hammerlund-Udenaes* for supporting the creation of the AAPS Student Chapter, *Prof. Lena Friberg and Prof. Ulrika Simonsson* for providing a great work environment, *Assoc. Prof. Andy Hooker* for reminding us science needs to be challenged, *Siv* for your quiet insight and being so cheerful about everything from pharmacometrics to sports, *Mia* for fun times with common friends, *Bettan* for always being nice to talk to, *Kajsa*, I am so impressed by your knowledge and how you manage do everything you do. SIR wouldn't have been the same without you! *Rikard* for being such a great addition to the PsN team, *Magnus,*

Tobias & Jerker for IT support, *Agneta Ekholm* for formatting the manuscripts in this thesis, and *Marina*, *Ulrica* & *Karin* for admin support.

From past and present Novartis, I would like to thank *Bruno* for always bringing jokes to meetings, *Oliver* for much help and nice afterworks, *Paolo* for being the most enthousiastic pharmacometrician I know, *Ulrica* and *Hannah* for answering all my questions about Uppsala, and all the people who make this group a very inspiring place to work.

I also want to thank friends and colleagues at Pharmetheus, especially *Marie* for being such a kind and enabling leader, *Arnaud* for IT enthousiasm and helping me live in a flat with furniture, *Sofia* from my time back in Mölndal, and *Jakob* for fun discussions.

Thank you to my half-time opponents *Sylvie Retout*, *Karl Brendel* and *Sebastian Ueckert* for valuable feedback and putting things into perspective.

To the "old" Uppsala team: *Elin* (yes you are part of it now!), working with you is a pleasure, thanks for many great times and (few) French-Swedish evenings, *Brendan,* thanks not for your science but for your parties (you said it), *Ron* for your help on dTBS and nice times whenever we meet again, *Akash* for being down to earth and welcoming me in Aachen, *Jan-Stefan* for saving us from a flat crisis in London, *Paul* for discussions we don't remember, *Brigitte* for a great model, *Åsa* for your laugh and showing me one could survive methodology, *Camille* for convincing me Uppsala was a great place at a breakfast meeting in Paris, *Elodie* for leading the way for the Frenchies, *Sebastian* for being such an interesting person to interact with at and outside of work, my co-author *Ronald* for your good spirit, *Jocke* for being a great and helping mind from my master's thesis to Pharmetheus, *Anders K* for having entertaining ideas about so many things, *Kristin* for being an IDeAl travelling companion, *Annika* for your kindness, *Yasunori* for your genuine interest in science and how to pass it on, *Julia* for all your energy, *Elke* because it's not racist if it's true, and *Winn* for great pictures.

To the "in between" Uppsala team: *Chunli*, the best roomie!, *Ari* for loud nation discussions, *Sofia* for good moods, *Thomas* for calm humor, *Henrik* for not saying half the puns you think of, *Erik* for nice social evening and some political debate, *Robin* for your help as neighbor and good times out, *Eric* for your help back in master student times, *Steve* for great pics and fun conversations, *Nebojsa* for being a great DJ, *Gustaf* for best actor, *Jörgen*, *Oskar* and *Oskar* for indulging me for the relay race, *Marina*, *Ida*, *Siti*, *Jesmin*, *Shijun,* and *Yang* for making life in the department so nice.

To the "new" Uppsala team: *Eva* for crying at the movies, *Margreke* for being so easy going, *Anders T* for nice company during non Anders-only week-ends, *João* for making us challenge our comfort zone, *Moustafa* and *Gunnar* for enjoying the methodology group, *Sebastian* for very interesting thoughts about many things, *Philippe* for continuing the French influence.

To all the "short term" Uppsala people: thank you *Rasmus*, *Trine*, *Coen,* *Anke*, *Stein, Aurelia* & *all* for shaking Uppsala up!

To the Uppsala “all-timers”: *Salim* for leading me here and making me think different, *Ana* for having the world’s best “répartie” and being reliable from hard times to party times, *Emilie* for sharing all my Sweden time, being great at time management and being always there for others, *Anna* for being crazy and having the most wonderful wedding, *Benjamin* foRRR being a geek with a storytelling gift, *Jihane* and *Mirmax* for having me over so many times, *Irena* for having the biggest heart and being so humble about it, *David* for great meals and great times during and outside of the lunch club, *Waqas* for long runs and many laughs, *Vijay* for being an amazing friend, *Chayan* for being so inappropriate and loving food, *Lani* for always having new traveling/running/reading projects, *Gopi* for great times and sorry you didn’t get a Tesla.

To my dear friends from Basel: *Ivan*, it has been so great having you as a friend, *Aziz* for being multitalented and eating ketchup pasta, *Vanessa, Kalo, Typhaine, Emma, Jeremy, Arthur, Marion, Damien* for adopting me in your group and living so many awesome adventures together.

To my girls from Pharmacy school *Maïlys*, *Daphné*, *Mouna*, *Joanne* and *Léa* for so many memories and coming to visit me in Sweden, and to *Julie*, my best friend since forever, for being there always.

Lastly, I would like to thank my family, without whom I certainly would not be here today. In particular:

Papa et *Maman*, vous avez protégé et illuminé mon enfance. Vous m’avez donné le goût du voyage, et ouvert toutes les portes de la vie d’adulte, en me laissant le choix de celles que j’allais choisir. Papa, j’aurais aimé que tu sois là pour m’aider à les franchir et continuer à me guider. Tu me manques aujourd’hui comme tu me manques tous les jours. Maman, tu es là et je suis heureuse que malgré la distance nous ayons toujours réussi à passer de nombreux week-ends et vacances tous ensemble. Je t’admire pour ta force de caractère et espère toujours m’en inspirer.

F-X, le meilleur frère du monde, OK j’en ai qu’un seul mais heureusement parce que je sais pas comment feraient les autres pour rivaliser. De New-York, à Nice, tu nous fais rire, tu gères et tu care, je sais pas comment je ferais sans toi. Et à ma belle-sœur, *Virginie*, qui sait le rendre heureux!

A ma *mamie*, pour toutes les vacances magiques de notre enfance et avoir eu la chance de te connaitre en tant qu’adulte.

A mon cousin *Benji* pour sa joie de vivre à toute épreuve, et à son frère *Clément*, personne extraordinaire que je porterai toujours dans mon cœur.

An meinen Mann *Thorsten*, danke, Schatz, dass du all diese Jahre ausgehalten hast : die Distanzbeziehung, tausende Flüge, die immer begrenzte Zeit zusammen, schwierige Momente… Danke, dass du mir immer die Freiheit gegeben hast, und mich dabei unterstüzt hast, meinen Weg zu gehen. Jetzt fängt es endlich an – ich freue mich so auf abendliche Diskussionen, Reisen durch die Welt, und noch viel mehr. Ich liebe dich !

Anne-Gaëlle

References

1 International Conference on Harmonization, E9: Statistical Principles for Clinical Trials. (Fed Regist., 1998).

2 Sheiner, L. B. Learning versus confirming in clinical drug development. *Clin Pharmacol Ther* 61, 275-291, doi:10.1016/S0009-9236(97)90160-0 (1997).

3 Marshall, S. F. *et al.* Good Practices in Model-Informed Drug Discovery and Development (MID3): Practice, Application and Documentation. *CPT: Pharmacometrics & Systems Pharmacology*, n/a-n/a, doi:10.1002/psp4.12049 (2015).

4 Lalonde, R. L. *et al.* Model-based drug development. *Clin Pharmacol Ther* 82, 21-32, doi:10.1038/sj.clpt.6100235 (2007).

5 Miller, R. *et al.* How Modeling and Simulation Have Enhanced Decision Making in New Drug Development. *Journal of pharmacokinetics and pharmacodynamics* 32, 185-197, doi:10.1007/s10928-005-0074-7 (2005).

6 US Department of Health and Human Services, Food. and. Drug. Administration. *Innovation or stagnation? Challenge and opportunity on the critical path to new medical products*, <http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm113411.pdf> (2004).

7 Ette, E. I. & Williams, P. J. Pharmacometrics: The Science of Quantitative Pharmacology. (2007).

8 Holford, N. H. G., Kimko, H. C., Monteleone, J. P. R. & Peck, C. C. Simulation of Clinical Trials. *Annual Review of Pharmacology and Toxicology* 40, 209-234, doi:doi:10.1146/annurev.pharmtox.40.1.209 (2000).

9 Suryawanshi, S., Zhang, L., Pfister, M. & Meibohm, B. The current role of model-based drug development. *Expert Opinion on Drug Discovery* 5, 311-321, doi:10.1517/17460441003713470 (2010).

10 Demin, I., Hamren, B., Luttringer, O., Pillai, G. & Jung, T. Longitudinal model-based meta-analysis in rheumatoid arthritis: an application toward model-based drug development. *Clin Pharmacol Ther* 92, 352-359, doi:10.1038/clpt.2012.69 (2012).

11 Marshall, S. F. *et al.* Modeling and Simulation to Optimize the Design and Analysis of Confirmatory Trials, Characterize Risk–Benefit, and Support Label Claims. *CPT: Pharmacometrics & Systems Pharmacology* 2, 1-3, doi:10.1038/psp.2013.4 (2013).

12 Bretz, F., Pinheiro, J. C. & Branson, M. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* 61, 738-748, doi:10.1111/j.1541-0420.2005.00344.x (2005).

13 Lee, J. Y. *et al.* Impact of pharmacometric analyses on new drug approval and labelling decisions: a review of 198 submissions between 2000 and 2008. *Clinical pharmacokinetics* 50, 627-635, doi:10.2165/11593210-000000000-00000 (2011).

14 Food and Drug Administration, *Food and Drug Administration Modernization Act of 1997*,<http://www.fda.gov/RegulatoryInformation/Legislation/ Federal-FoodDrugandCosmeticActFDCAct/SignificantAmendmentstotheFDCAct/ FDAMA/FullTextofFDAMAlaw/default.htm> (1997).

15 Rajman, I. PK/PD modelling and simulations: utility in drug development. *Drug Discovery Today* 13, 341-346, doi:http://dx.doi.org/10.1016/j.drudis. 2008.01.003 (2008).

16 Vong, C., Bergstrand, M., Nyberg, J. & Karlsson, M. O. Rapid sample size calculations for a defined likelihood ratio test-based power in mixed-effects models. *The AAPS journal* 14, 176-186, doi:10.1208/s12248-012-9327-8 (2012).

17 Jonsson, E. N. & Sheiner, L. B. More efficient clinical trials through use of scientific model-based statistical tests. *Clin Pharmacol Ther* 72, 603-614, doi:10.1067/mcp.2002.129307 (2002).

18 Sampson, M. R., Benjamin, D. K. & Cohen-Wolkowiez, M. Evidence-based guidelines for pediatric clinical trials: focus on StaR Child Health. *Expert review of clinical pharmacology* 5, 525-531, doi:10.1586/ecp.12.52 (2012).

19 Schork, N. Personalized medicine: Time for one-person trials. *Nature News* (2015).

20 Hilgers, R. D. *Integrated Design and Analysis of Small Population Group Trials*, <http://www.ideal.rwth-aachen.de/> (2013).

21 Hu, C. & Zhou, H. An improved approach for confirmatory phase III population pharmacokinetic analysis. *Journal of clinical pharmacology* 48, 812-822, doi:10.1177/0091270008318670 (2008).

22 Chow, A. T. *et al.* Utility of population pharmacokinetic modeling in the assessment of therapeutic protein-drug interactions. *Journal of clinical pharmacology*, doi:10.1002/jcph.240 (2013).

23 Orloff, J. *et al.* The future of drug development: advancing clinical trial design. *Nat Rev Drug Discov* 8, 949-957,doi:http://www.nature.com/ nrd/journal/v8/n12/suppinfo/nrd3025_S1.html (2009).

24 Kimko, H. & Pinheiro, J. Model-based clinical drug development in the past, present and future: a commentary. *British journal of clinical pharmacology* 79, 108-116, doi:10.1111/bcp.12341 (2015).

25 Hu, C., Zhang, J. & Zhou, H. Confirmatory analysis for phase III population pharmacokinetics. *Pharmaceutical statistics* 10, 14-26, doi:10.1002/pst.403 (2011).

26 Kang, D. W., Schwartz, J. B. & Verotta, D. A sample size computation method for non-linear mixed effects models with applications to pharmacokinetics models. *Statistics in medicine* 23, 2551-2566, doi:Doi 10.1002/Sim.1695 (2004).

27 Ueckert, S., Karlsson, M. O. & Hooker, A. C. Accelerating Monte Carlo power studies through parametric power estimation. *Journal of pharmacokinetics and pharmacodynamics* 43, 223-234, doi:10.1007/s10928-016-9468-y (2016).

28 Petersson, K. F., Hanze, E., Savic, R. & Karlsson, M. Semiparametric Distributions With Estimated Shape Parameters. *Pharm Res* 26, 2174-2185, doi:10.1007/s11095-009-9931-1 (2009).

29 Karlsson, M. O., Beal, S. L. & Sheiner, L. B. Three new residual error models for population PK/PD analyses. *Journal of pharmacokinetics and biopharmaceutics* 23, 651-672 (1995).

30 Lindsey, J. K., Byrom, W. D., Wang, J., Jarvis, P. & Jones, B. Generalized Nonlinear Models for Pharmacokinetic Data. *Biometrics* 56, 81-88, doi:10.2307/2677106 (2000).

31 Lindsey, J. K. & Jones, B. Modeling Pharmacokinetic Data Using Heavy-Tailed Multivariate Distributions. *Journal of Biopharmaceutical Statistics* 10, 369-381, doi:10.1081/bip-100102500 (2000).

32 Hu, C., Moore, K. H., Kim, Y. H. & Sale, M. E. Statistical issues in a modeling approach to assessing bioequivalence or PK similarity with presence of sparsely sampled subjects. *Journal of pharmacokinetics and pharmacodynamics* 31, 321-339 (2004).

33 Bieth, B. *et al. Longitudinal model-based test as primary analysis in phase III*, <http://www.ema.europa.eu/docs/en_GB/document_library/Presentation/2011/11/WC500118295.pdf> (2011).

34 Lonnebo, A., Grahnen, A. & Karlsson, M. O. An integrated model for the effect of budesonide on ACTH and cortisol in healthy volunteers. *British journal of clinical pharmacology* 64, 125-132, doi:10.1111/j.1365-2125.2007.02867.x (2007).

35 Karlsson, M. O. & Sheiner, L. B. The importance of modeling interoccasion variability in population pharmacokinetic analyses. *Journal of pharmacokinetics and biopharmaceutics* 21, 735-750, doi:10.1007/bf01113502 (1993).

36 Lindemalm, S. *et al.* Application of population pharmacokinetics to cladribine. *BMC pharmacology* 5, 4, doi:10.1186/1471-2210-5-4 (2005).

37 Hassan, M. *et al.* A mechanism-based pharmacokinetic-enzyme model for cyclophosphamide autoinduction in breast cancer patients. *British journal of clinical pharmacology* 48, 669-677 (1999).

38 Jonsson, S. *et al.* Population pharmacokinetics of ethambutol in South African tuberculosis patients. *Antimicrobial agents and chemotherapy* 55, 4230-4237, doi:10.1128/aac.00274-11 (2011).

39 Hempel, G. *et al.* Population pharmacokinetic-pharmacodynamic modeling of moxonidine using 24-hour ambulatory blood pressure measurements. *Clin Pharmacol Ther* 64, 622-635, doi:10.1016/s0009-9236(98)90053-4 (1998).

40 Friberg, L. E., Henningsson, A., Maas, H., Nguyen, L. & Karlsson, M. O. Model of chemotherapy-induced myelosuppression with parameter consistency across drugs. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 20, 4713-4721 (2002).

41 Wahlby, U., Thomson, A. H., Milligan, P. A. & Karlsson, M. O. Models for time-varying covariates in population pharmacokinetic-pharmacodynamic analysis. *British journal of clinical pharmacology* 58, 367-377, doi:10.1111/j.1365-2125.2004.02170.x (2004).

42 Grasela, T. H., Jr. & Donn, S. M. Neonatal population pharmacokinetics of phenobarbital derived from routine clinical data. *Developmental pharmacology and therapeutics* 8, 374-383 (1985).

43 Lindbom, L., Pihlgren, P. & Jonsson, E. N. PsN-Toolkit--a collection of computer intensive statistical methods for non-linear mixed effect modeling using NONMEM. *Computer methods and programs in biomedicine* 79, 241-257, doi:10.1016/j.cmpb.2005.04.005 (2005).

44 Davidian M, G. D. *Nonlinear Models for Repeated Measurement Data*.(1995).

45 Newey, W. K. & McFadden, D. in *Handbook of Econometrics*
(ed Elsevier) (1994).

46 NONMEM User's Guides. (1989-2009) (Ellicott City, MD, USA, 2009).

47 Efron, B. Bootstrap Methods: Another Look at the Jackknife. 1-26, doi:10.1214/aos/1176344552 (1979).

48 Ette, E. I. & Onyiah, L. C. Estimating inestimable standard errors in population pharmacokinetic studies: the bootstrap with Winsorization. *European journal of drug metabolism and pharmacokinetics* 27, 213-224 (2002).

49 Thai, H. T., Mentre, F., Holford, N. H., Veyrat-Follet, C. & Comets, E. Evaluation of bootstrap methods for estimating uncertainty of parameters in nonlinear mixed-effects models: a simulation study in population pharmacokinetics. *Journal of pharmacokinetics and pharmacodynamics* 41, 15-33, doi:10.1007/s10928-013-9343-z (2014).

50 Efron, B. in *The Jackknife, the Bootstrap and Other Resampling Plans CBMS-NSF Regional Conference Series in Applied Mathematics* i-xi (Society for Industrial and Applied Mathematics, 1982).

51 Sheiner, L. B. Analysis of pharmacokinetic data using parametric models. III. Hypothesis tests and confidence intervals. *Journal of pharmacokinetics and biopharmaceutics* 14, 539-555 (1986).

52 Denney, W. N-dimensional Likelihood Profiling: An Efficient Alternative to Bootstrap *PAGE. Abstracts of the Annual Meeting of the Population Approach Group in Europe. PAGE 21*, Abstr 2594 (2012).

53 Wilks, S. The Large-Sample Distribution of the likelihood ratio for testing composite hypotheses. *American Mathematical Society* (1937).

54 Wahlby, U., Jonsson, E. N. & Karlsson, M. O. Assessment of actual significance levels for covariate effects in NONMEM. *Journal of pharmacokinetics and pharmacodynamics* 28, 231-252 (2001).

55 Rubin, D. in *Bayesian Statistics 3* (eds JM. Bernardo, MH. DeGroot, DV. Lindley, & AFM. Smith) 395-402 (Oxford University Press, 1988).

56 Svensson, E. M., Dooley, K. E. & Karlsson, M. O. Impact of lopinavir-ritonavir or nevirapine on bedaquiline exposures and potential implications for patients with tuberculosis-HIV coinfection. *Antimicrobial agents and chemotherapy* 58, 6406-6412, doi:10.1128/aac.03246-14 (2014).

57 Gordi, T. *et al.* A semiphysiological pharmacokinetic model for artemisinin in healthy subjects incorporating autoinduction of metabolism and saturable first-pass hepatic extraction. *British journal of clinical pharmacology* 59, 189-198, doi:10.1111/j.1365-2125.2004.02321.x (2005).

58 Bogason, A. *et al.* Inverse relationship between leukaemic cell burden and plasma concentrations of daunorubicin in patients with acute myeloid leukaemia. *British journal of clinical pharmacology* 71, 514-521, doi:10.1111/j.1365-2125.2010.03894.x (2011).

59 Brill, M. J. *et al.* The Pharmacokinetics of the CYP3A Substrate Midazolam in Morbidly Obese Patients Before and One Year After Bariatric Surgery. *Pharm Res* 32, 3927-3936, doi:10.1007/s11095-015-1752-9 (2015).

60 Nielsen, E. I., Sandstrom, M., Honore, P. H., Ewald, U. & Friberg, L. E. Developmental pharmacokinetics of gentamicin in preterm and term neonates: population modelling of a prospective study. *Clinical pharmacokinetics* 48, 253-263, doi:10.2165/00003088-200948040-00003 (2009).

61 Karlsson, M. O., Jonsson, E. N., Wiltse, C. G. & Wade, J. R. Assumption testing in population pharmacokinetic models: illustrated with an analysis of moxonidine data from congestive heart failure patients. *Journal of pharmacokinetics and biopharmaceutics* 26, 207-246 (1998).

62 Bouchene, S. *et al. Development of a Whole-Body Physiologically Based Pharmacokinetic Model for Colistin and Colistin methanesulfonate in Rat* PhD thesis, Uppsala University, (2016).

63 Jonsson, S., Simonsson, U. S., Miller, R. & Karlsson, M. O. Population pharmacokinetics of edoxaban and its main metabolite in a dedicated renal impairment study. *Journal of clinical pharmacology* 55, 1268-1279, doi:10.1002/jcph.541 (2015).

64 Dorlo, T. P. *et al.* Pharmacokinetics of miltefosine in Old World cutaneous leishmaniasis patients. *Antimicrobial agents and chemotherapy* 52, 2855-2860, doi:10.1128/aac.00014-08 (2008).

65 Plan, E. L., Elshoff, J. P., Stockis, A., Sargentini-Maier, M. L. & Karlsson, M. O. Likert pain score modeling: a Markov integer model and an autoregressive continuous model. *Clin Pharmacol Ther* 91, 820-828, doi:10.1038/clpt.2011.301 (2012).

66 Zingmark, P. H., Edenius, C. & Karlsson, M. O. Pharmacokinetic/ pharmacodynamic models for the depletion of Vbeta5.2/5.3 T cells by the monoclonal antibody ATM-027 in patients with multiple sclerosis, as measured by FACS. *British journal of clinical pharmacology* 58, 378-389, doi:10.1111/j.1365-2125.2004.02177.x (2004).

67 Claret, L. *et al.* Model-based prediction of phase III overall survival in colorectal cancer on the basis of phase II tumor dynamics. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 27, 4103-4108, doi:10.1200/jco.2008.21.0807 (2009).

68 Troconiz, I. F., Plan, E. L., Miller, R. & Karlsson, M. O. Modelling overdispersion and Markovian features in count data. *Journal of pharmacokinetics and pharmacodynamics* 36, 461-477, doi:10.1007/s10928-009-9131-y (2009).

69 Hamren, B., Bjork, E., Sunzel, M. & Karlsson, M. Models for plasma glucose, HbA1c, and hemoglobin interrelationships in patients with type 2 diabetes following tesaglitazar treatment. *Clin Pharmacol Ther* 84, 228-235, doi:10.1038/clpt.2008.2 (2008).

70 Abrantes, J. A., Almeida, A., Sales, F., Falcao, A. & Jonsson, S. A repeated time-to-event model for epileptic seizures in patients undergoing antiepileptic drug withdrawal during video-electroencephalography monitoring. *PAGE. Abstracts of the Annual Meeting of the Population Approach Group in Europe* 23, Abstr 3180 (2014).

71 Karlsson, K. E., Eriksson, R., Karlsson, M. O. & Nyberg, J. Estimating a Cox proportional hazard model in NONMEM. *PAGE. Abstracts of the Annual Meeting of the Population Approach Group in Europe.* 23 (2014).

72 Silber, H. E. *et al.* An integrated model for glucose and insulin regulation in healthy volunteers and type 2 diabetic patients following intravenous glucose provocations. *Journal of clinical pharmacology* 47, 1159-1171, doi:10.1177/0091270007304457 (2007).

73 Savic, R. M., Jonker, D. M., Kerbusch, T. & Karlsson, M. O. Implementation of a transit compartment model for describing drug absorption in pharmacokinetic studies. *Journal of pharmacokinetics and pharmacodynamics* 34, 711-726, doi:10.1007/s10928-007-9066-0 (2007).

74 Svensson, R. J. & Simonsson, U. Application of the Multistate Tuberculosis Pharmacometric Model in Patients With Rifampicin-Treated Pulmonary Tuberculosis. *CPT Pharmacometrics Syst Pharmacol* 5, 264-273, doi:10.1002/psp4.12079 (2016).

75 Hennig, S., Friberg, L. E. & Karlsson, M. O. Characterizing time to conversion to sinus rhythm under digoxin and placebo in acute atrial fibrillation. *PAGE. Abstracts of the Annual Meeting of the Population Approach Group in Europe* 18, Abstr 1504 (2009).

76 Jauslin, P. M. *et al.* An integrated glucose-insulin model to describe oral glucose tolerance test data in type 2 diabetics. *Journal of clinical pharmacology* 47, 1244-1255, doi:10.1177/0091270007302168 (2007).

77 Hornestam, B., Jerling, M., Karlsson, M. O. & Held, P. Intravenously administered digoxin in patients with acute atrial fibrillation: a population pharmacokinetic/pharmacodynamic analysis based on the Digitalis in Acute Atrial Fibrillation trial. *European journal of clinical pharmacology* 58, 747-755, doi:10.1007/s00228-002-0553-3 (2003).

78 Chen, C. *et al.* Population pharmacokinetic-pharmacodynamic modelling of rifampicin treatment response in a tuberculosis acute mouse model. *PAGE. Abstracts of the Annual Meeting of the Population Approach Group in Europe* 23, Abstr 3224 (2014).

79 Choy, S., de Winter, W., Karlsson, M. O. & Kjellsson, M. C. Modeling the Disease Progression from Healthy to Overt Diabetes in ZDSD Rats. *The AAPS journal* 18, 1203-1212, doi:10.1208/s12248-016-9931-0 (2016).

80 Hjort, N. L. & Claeskens, G. Frequentist Model Average Estimators. *Journal of the American Statistical Association* 98, 879-899, doi:10.1198/016214503000000828 (2003).

81 Fridericia, L. S. The duration of systole in an electrocardiogram in normal humans and in patients with heart disease. 1920. *Annals of noninvasive electrocardiology : the official journal of the International Society for Holter and Noninvasive Electrocardiology, Inc* 8, 343-351 (2003).

82 International Conference on Harmonisation, E14: Guidance on clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs. (Fed Regist., 2005).

83 Ramsay, J. O. Monotone Regression Splines in Action. 425-441, doi:10.1214/ss/1177012761 (1988).

84 Yuan, Y. & Yin, G. Dose–Response Curve Estimation: A Semiparametric Mixture Approach. *Biometrics* 67, 1543-1554, doi:10.1111/j.1541-0420.2011.01620.x (2011).

85 Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics* 6, 461-464 (1978).

86 Felson, D. T. *et al.* American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. *Arthritis and rheumatism* 38, 727-735 (1995).

87 Lacroix, B. D. *et al.* A pharmacodynamic Markov mixed-effects model for determining the effect of exposure to certolizumab pegol on the ACR20 score in patients with rheumatoid arthritis. *Clin Pharmacol Ther* 86, 387-395, doi:10.1038/clpt.2009.136 (2009).

88 R: A language and environment for statistical computing. (R Core Team. Foundation for Statistical Computing, Vienna, Austria., 2015).

89 Jonsson, E. N. & Karlsson, M. O. Xpose—an S-PLUS based population pharmacokinetic/pharmacodynamic model building aid for NONMEM. *Computer methods and programs in biomedicine* 58, 51-64, doi:http://dx.doi.org/10.1016/S0169-2607(98)00067-4 (1998).

90 Leary, R., Dunlavey, M. R. & Chittenden, J. A Fast Bootstrap Method Using EM Posteriors. *PAGE. Abstracts of the Annual Meeting of the Population Approach Group in Europe* 22, Abstr 2797 (2013).

91 Silber, H. E., Kjellsson, M. C. & Karlsson, M. O. The impact of misspecification of residual error or correlation structure on the type I error rate for covariate inclusion. *Journal of pharmacokinetics and pharmacodynamics* 36, 81-99, doi:10.1007/s10928-009-9112-1 (2009).

92 Long, J. S. & Ervin, L. H. Using heteroscedasticity consistent standard errors in the linear regression model. *Am Stat* 54, 217-224, doi:Doi 10.2307/2685594 (2000).

93 Sheiner, L. B. & Beal, S. L. Pharmacokinetic parameter estimates from several least squares procedures: superiority of extended least squares. *Journal of pharmacokinetics and biopharmaceutics* 13, 185-201 (1985).

94 Maas, A. & Cora, J. M. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis* 46, 427-440 (2004).

95 Yafune, A. & Ishiguro, M. Bootstrap approach for constructing confidence intervals for population pharmacokinetic parameters. I: A use of bootstrap standard error. *Statistics in medicine* 18, 581-599 (1999).

96 Donaldson, J. R. & Schnabel, R. B. Computational Experience with Confidence Regions and Confidence Intervals for Nonlinear Least Squares. *Technometrics* 29, 67-82, doi:10.2307/1269884 (1987).

97 Garnett, C. & Holford, N. The relative importance of between-subject correlation of population parameters compared with estimation correlation when applied to pharmacokinetic simulation. *Clinical Pharmacology & Therapeutics* 75, P.Abstracts (2004).

98 Reeve, R. & Turner, J. R. Pharmacodynamic Models: Parameterizing the Hill Equation, Michaelis-Menten, the Logistic Curve, and Relationships Among These Models. *Journal of Biopharmaceutical Statistics* 23, 648-661, doi:10.1080/10543406.2012.756496 (2013).

99 Owen, A. B. Controlling Correlations in Latin Hypercube Samples. *Journal of the American Statistical Association* 89, 1517-1522, doi:10.2307/2291014 (1994).

100 Zhao, P. *et al.* Applications of physiologically based pharmacokinetic (PBPK) modeling and simulation during regulatory review. *Clin Pharmacol Ther* 89, 259-267, doi:10.1038/clpt.2010.298 (2011).

101 Edginton, A. N., Theil, F. P., Schmitt, W. & Willmann, S. Whole body physiologically-based pharmacokinetic models: their use in clinical drug development. *Expert opinion on drug metabolism & toxicology* 4, 1143-1152, doi:10.1517/17425255.4.9.1143 (2008).

102 Hu, T. M. & Hayton, W. L. Allometric scaling of xenobiotic clearance: uncertainty versus universality. *AAPS pharmSci* 3, E29, doi:10.1208/ps030429 (2001).

103 Kenyon, E. M. Interspecies extrapolation. *Methods in molecular biology (Clifton, N.J.)* 929, 501-520, doi:10.1007/978-1-62703-050-2_19 (2012).

104 Korell, J., Martin, S. W., Karlsson, M. O. & Ribbing, J. A model-based longitudinal meta-analysis of FEV1 in randomized COPD trials. *Clinical Pharmacology & Therapeutics* 99, 315-324, doi:10.1002/cpt.249 (2016).

105 Wood, S. N. Monotonic smoothing splines fitted by cross validation. *SIAM J. Sci. Comput.* 15, 1126-1133, doi:10.1137/0915069 (1994).

106 Darpo, B. *et al.* Results from the IQ-CSRC prospective study support replacement of the thorough QT study by QT assessment in the early clinical phase. *Clin Pharmacol Ther* 97, 326-335, doi:10.1002/cpt.60 (2015).

107 Harmonisation, I. C. o. E14: Guidance on clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs. Questions & Answers (R3). (2015).

108 Lee, H. *et al.* Population pharmacokinetic and pharmacodynamic modeling of etanercept using logistic regression analysis. *Clin Pharmacol Ther* 73, 348-365 (2003).

109 Hu, C., Xu, Z., Rahman, M. U., Davis, H. M. & Zhou, H. A latent variable approach for modeling categorical endpoints among patients with rheumatoid arthritis treated with golimumab plus methotrexate. *Journal of pharmacokinetics and pharmacodynamics* 37, 309-321, doi:10.1007/s10928-010-9162-4 (2010).

110 Klingenberg, B. Proof of concept and dose estimation with binary responses under model uncertainty. *Statistics in medicine* 28, 274-292, doi:10.1002/sim.3477 (2009).

111 Kang, S. H., Kodell, R. L. & Chen, J. J. Incorporating model uncertainties along with data uncertainties in microbial risk assessment. *Regulatory toxicology and pharmacology : RTP* 32, 68-72, doi:10.1006/rtph.2000.1404 (2000).

112 Wheeler, M. W. & Bailer, A. J. Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environmental and Ecological Statistics* 16, 37-51, doi:10.1007/s10651-007-0071-7 (2009).

113 Morales, K. H., Ibrahim, J. G., Chen, C.-J. & Ryan, L. M. Bayesian Model Averaging with Applications to Benchmark Dose Estimation for Arsenic in Drinking Water. *Journal of the American Statistical Association* 101, 9-17 (2006).

114 Moon, H., Kim, H. J., Chen, J. J. & Kodell, R. L. Model averaging using the Kullback information criterion in estimating effective doses for microbial infection and illness. *Risk analysis : an official publication of the Society for Risk Analysis* 25, 1147-1159, doi:10.1111/j.1539-6924.2005.00676.x (2005).

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy 223

Editor: The Dean of the Faculty of Pharmacy

Distribution: publications.uu.se
urn:nbn:se:uu:diva-305697

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2016