Covariate Model Building in Nonlinear Mixed Effects Models

JAKOB RIBBING
Dissertation presented at Uppsala University to be publicly examined in B41, BMC, Husarg. 3, Uppsala, Tuesday, June 5, 2007 at 09:15 for the degree of Doctor of Philosophy (Faculty of Pharmacy). The examination will be conducted in English.

Abstract

Population pharmacokinetic-pharmacodynamic (PK-PD) models can be fitted using nonlinear mixed effects modelling (NONMEM). This is an efficient way of learning about drugs and diseases from data collected in clinical trials. Identifying covariates which explain differences between patients is important to discover patient subpopulations at risk of sub-therapeutic or toxic effects and for treatment individualization. Stepwise covariate modelling (SCM) is commonly used to this end. The aim of the current thesis work was to evaluate SCM and to develop alternative approaches. A further aim was to develop a mechanistic PK-PD model describing fasting plasma glucose, fasting insulin, insulin sensitivity and beta-cell mass.

The lasso is a penalized estimation method performing covariate selection simultaneously to shrinkage estimation. The lasso was implemented within NONMEM as an alternative to SCM and is discussed in comparison with that method. Further, various ways of incorporating information and propagating knowledge from previous studies into an analysis were investigated. In order to compare the different approaches, investigations were made under varying, replicated conditions. In the course of the investigations, more than one million NONMEM analyses were performed on simulated data. Due to selection bias the use of SCM performed poorly when analysing small datasets or rare subgroups. In these situations, the lasso method in NONMEM performed better, was faster, and additionally validated the covariate model. Alternatively, the performance of SCM can be improved by propagating knowledge or incorporating information from previously analysed studies and by population optimal design.

A model was also developed on a physiological/mechanistic basis to fit data from three phase II/III studies on the investigational drug, tesaglitazar. This model described fasting glucose and insulin levels well, despite heterogeneous patient groups ranging from non-diabetic insulin resistant subjects to patients with advanced diabetes. The model predictions of beta-cell mass and insulin sensitivity were well in agreement with values in the literature.

Keywords: Pharmacokinetics, Pharmacodynamics, Modeling, Covariate selection, Stepwise selection, Covariate analysis, Methodology, Model validation, Model evaluation, Type-2 diabetes, Beta-cell function, Meta analysis, Cross-validation, Least absolute shrinkage and selection operator, Pharmacometrics, ED optimization

Jakob Ribbing, Department of Pharmaceutical Biosciences, Box 591, Uppsala University, SE-75124 Uppsala, Sweden

© Jakob Ribbing 2007

ISSN 1651-6192
urn:nbn:se:uu:diva-7923 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7923)
We are drowning in information and starving for knowledge.

*Rutherford D. Roger*
Papers Discussed

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


Copyright Springer. Reproduced with permission


Copyright Springer. Manuscript reproduced with permission.

III  **Ribbing J**, Hooker AC and Jonsson EN. Non-Bayesian Knowledge Propagation using Model-Based Analysis of Data from Multiple Clinical Studies. Submitted

IV  **Ribbing J**, Hamrén B, Svensson MK and Karlsson MO. A Model for Glucose, Insulin, Beta-Cell and HbA1c Dynamics in Subjects with Insulin Resistance and Patients with Type 2 Diabetes. In Manuscript
Contents

1. Introduction ............................................................................................................. 11
   1.1. Background ..................................................................................................... 11
   1.2. Nonlinear Mixed Effects Models .................................................................... 13
   1.3. The Covariate Model .................................................................................... 15
       1.3.1. What is a covariate? .................................................................................. 15
       1.3.2. What is a true covariate? .......................................................................... 16
       1.3.3. Multiplicative and additive parameterisation ............................................ 16
       1.3.4. Categorical covariates .............................................................................. 17
       1.3.5. Continuous covariates .............................................................................. 18
       1.3.6. Covariate transformation .......................................................................... 19
   1.4. Software for Model Fitting ............................................................................. 19
   1.5. Estimation Methods in NONMEM ............................................................... 20
       1.5.1. Individual empirical bayes estimates and shrinkage .................................. 21
   1.6. Model Development and Selection .............................................................. 21
       1.6.1. Stepwise selection ..................................................................................... 23
   1.7. The Base Model ............................................................................................ 23
   1.8. Selection of the Covariate Model .................................................................. 24
       1.8.1. Covariate selection outside NONMEM ..................................................... 24
       1.8.2. Covariate selection within NONMEM ....................................................... 26
   1.9. The Lasso in Ordinary Multiple Regression ................................................ 29
   1.10. Model Validation and Evaluation .................................................................. 29
       1.10.1. Cross-validation ..................................................................................... 30
   1.11. Modelling of Type 2 Diabetes ...................................................................... 32

2. Aims ......................................................................................................................... 34

3. Methods .................................................................................................................. 35
   3.1. Software ......................................................................................................... 35
   3.2. Analysed Data ................................................................................................ 35
       3.2.1. Simulated data (Papers I-III) ................................................................. 36
       3.2.2. Real data (Papers II and IV) ................................................................. 37
   3.3. Development of the Lasso for Covariate Selection within NONMEM (Paper II) ................................................................. 37
   3.4. Procedure for Analysis of nlme Models (Papers I-III) ..................................... 38
   3.5. Development of a Model for FPG and FI (Paper IV) ........................................ 40
   3.6. Adaptive Learning (Paper III) ....................................................................... 42
   3.7. Calculation of Power and Inclusion Rates (Papers I and III) ......................... 43
   3.8. Evaluation of Selection Bias (Papers I and III) .............................................. 43
   3.9. Evaluation of Predictive Performance (Papers I-III) ...................................... 44
   3.10. Evaluation of Design Optimization (Paper III) ............................................ 44
   3.11. Evaluation of the Model for FPG and FI (Paper IV) ..................................... 45
Abbreviations

AIC  Akaike’s Information Criterion; A criterion for model selection
ATAR  covariate indicating concomitant medication with ataractic drugs
BCM  beta-cell mass
CL  drug clearance, unit of flow, e.g. l/h
cov  denotes a covariate
cov_ij  value of cov for the jth observation of in subject i
CRCL  creatinine clearance, marker of renal function
CV  coefficient of variation, variability relative to the typical value
df  degrees of freedom
DGR  disease groups, to classify subjects at different stages of T2DM
DP  data pooling, analysing a dataset consisting of several studies
DV  dependent variable
EBE  empirical-Bayes estimate of individual parameter, i.e. \( \tilde{P}_i \) or \( \tilde{\eta}_i \)
FO  first-order method; estimation method in NONMEM
FOCE  first-order conditional estimation method
FOCE-I  FOCE with interaction; takes \( \eta \cdot P \) interaction into account
fu  ratio of unbound and total drug in plasma
GA  genetic algorithm; Can be used as a procedure for model selection
FFAs  free fatty acids
FI  fasting insulin level, measured in serum or plasma
FPG  fasting plasma glucose; marker of short-term glycemic control
GAM  generalised-additive modelling
HbA1c  fraction glycosylated haemoglobin A1c
IIV  random inter-individual variability, estimated as \( \tilde{\sigma} \)
IOV  random inter-occasion variability
\( k_a \)  absorption rate constant, parameter in first-order absorption models
\( lasso \)  least absolute shrinkage and selection operator
mae  mean absolute error, a measure of precision
mle  maximum-likelihood estimates
MM  model merge, combining results from analyses of separate datasets
nlme  nonlinear mixed effects
NI  naïve independent, analysing study without knowledge propagation
OFV  extended least squares objective function value
\( P \)  denotes any structural-model parameter
PD  pharmacodynamics
\( \tilde{P}_i \)  parameter value for subject i
PI  prediction interval, the model-predicted confidence interval of DVs
PK  pharmacokinetics
PK-PD  pharmacokinetic-pharmacodynamic
pOFV  predictive OFV, representing the likelihood of external data
PPAR  peroxisome proliferator-activated receptor
PPV  population-parameter variability (random and predictable)
PS  pre-specified, fitting a pre-specified model without model selection
PSP  fitting the PS model on all available data
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>PaN</td>
<td>Perl-speaks-NONMEM</td>
</tr>
<tr>
<td>r</td>
<td>Pearson correlation coefficient</td>
</tr>
<tr>
<td>rmse</td>
<td>root mean squared error, a measure of precision</td>
</tr>
<tr>
<td>S</td>
<td>insulin sensitivity, model parameter describing insulin resistance</td>
</tr>
<tr>
<td>SBC</td>
<td>Schwartz’s Bayesian criterion; A criterion for model selection</td>
</tr>
<tr>
<td>SCM</td>
<td>stepwise covariate modelling; A procedure for covariate selection</td>
</tr>
<tr>
<td>SEX</td>
<td>sex, often confounded with gender. Gender is the socio-economic aspects that follows with the (biological) sex</td>
</tr>
<tr>
<td>STS</td>
<td>standard two-stage approach</td>
</tr>
<tr>
<td>T2DM</td>
<td>type 2 diabetes mellitus</td>
</tr>
<tr>
<td>WAM</td>
<td>Wald Approximation Method; A procedure for covariate selection</td>
</tr>
<tr>
<td>TVP</td>
<td>typical value of parameter P, given covariate values for subject i</td>
</tr>
<tr>
<td>V</td>
<td>volume of drug distribution (apparent)</td>
</tr>
<tr>
<td>VPC</td>
<td>visual predictive check, used for evaluation of nlme models</td>
</tr>
<tr>
<td>WT</td>
<td>body weight</td>
</tr>
<tr>
<td>$y_{ij}$</td>
<td>dependent variable, $j^{th}$ observation of in subject $i$</td>
</tr>
<tr>
<td>$f_{ij}$</td>
<td>model prediction of dependent variable, $j^{th}$ observation of in subject $i$</td>
</tr>
<tr>
<td>$e_{ij}$</td>
<td>residual error, weighted for the $j^{th}$ observation of in subject $i$</td>
</tr>
<tr>
<td>$\eta_{ij}$</td>
<td>the deviation from TVP$_i$ in subject $i$</td>
</tr>
<tr>
<td>$\omega_{P}$</td>
<td>magnitude of the IIV in parameter $P$, random-effects parameter</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>magnitude of the intra-individual error</td>
</tr>
<tr>
<td>$\theta$</td>
<td>denotes a fixed-effect that is estimated in the model</td>
</tr>
<tr>
<td>$\theta_{cov}$</td>
<td>the covariate coefficient for covariate $cov$ on parameter $P$</td>
</tr>
<tr>
<td>$\theta_{Ppop}$</td>
<td>the population typical value of parameter $P$</td>
</tr>
</tbody>
</table>
1. Introduction

1.1. Background

When a drug is administered to a patient, a chain of events takes place, eventually leading to a treatment response. Normally, the drug molecules must first be absorbed into the body and pass the liver before being distributed into the body tissues. The drug molecules reaching their site of action exert an effect by binding to specific target receptors. The treatment response for some drugs can appear almost immediately upon interaction of drug and receptor, but can also often be delayed (minutes to years). The delay may be due to a cascade of intermediate events and/or accumulation of the effect following the drug-receptor interaction.

Thus, at times it can be more practical to measure a marker of a drug-treatment effect rather than the so-called clinical endpoint, the hoped-for therapeutic outcome. For example, the treatment of high blood pressure is often evaluated by the reduction of blood pressure although the clinical endpoint is the absence of heart attack or stroke, or even survival.

Eventually, the effect of a single drug dose wears off, normally because the drug disappears from the body via metabolism and/or excretion into the urine. All events following drug administration can be divided into one of two processes.

1. Pharmacokinetics (PK) describes the fate of the drug molecules in the body after administration of a dosage regimen, i.e. what the body does to the drug.

2. Pharmacodynamics (PD), on the other hand, describes the action of the drug molecules in the body, i.e. what the drug does to the body.

The PK and PD of a drug are often described in the form of a model summarizing the important concepts in both qualitative and quantitative terms. The model is often longitudinal, meaning that it describes repeated measurements over time. Nonlinear models are used because the dependent variables (e.g. drug concentrations or drug effects) vary nonlinearly with time. The pharmacokinetic-pharmacodynamic (PK-PD) model can also be coupled with a disease model that describes the disease and its development over time. This is illustrated in Figure 1.
Figure 1. Illustration of a PK-PD model. The PK model describes systemic drug exposure. The PD model describes the response to the drug, measured in terms of changes to a clinical endpoint or a biomarker relevant for that endpoint. The endpoint, in turn, reflects desired or adverse effects. Whereas the dosage regimen is the input to the PK model, the resulting exposure may be used as the input to the PD model. The collective model is called a PK-PD model.

The parameter in the model may have an associated physiological or mechanistic meaning, e.g. the volume of blood or plasma that is completely cleared of the drug per unit of time, or the maximum effect that the drug could produce if it were to occupy all its receptors. Highly mechanistic models can include processes of cell ageing or differentiation, or drug distribution into various organs based on blood flow and organ size, or known chemical or molecular interactions. The opposite of a mechanistic model is an empirical model, which describes the observations of response/exposure but has no scientific basis for its structure.

A population model, also called a nonlinear mixed-effects (nlme) model quantifies the variability between individuals in the model parameters. It is vitally important to take inter-individual variability (IIV) into account in the model, rather than treating it as an independent error, both in order to describe the variability itself and to obtain an accurate description of the typical individual. The exposure and response to a drug varies substantially among patients. As a consequence, evidence that a treatment is safe and effective in the typical patient does not warrant its use in the whole patient population. Decisions in drug development are increasingly being made on model-based population analyses of the available information. This approach is encouraged by many of the regulatory agencies, e.g. the Food and Drug Administration in the USA and the Medical Products Agency in Sweden.

It is often useful to explain the variability in a parameter using a covariate model that describes the relations between covariates and parameters. Whereas a parameter is a fixed quantity estimated according to the model, a covariate is an independent variable that contains information on a parameter, i.e. the parameter depends on the covariate. For example, small patients have smaller volume into which the drug is distributed (V) and lower clearance of the drug from the body (CL). This is because CL and V are parameters that depends on the body size. This is the main reason to that children often receive lower doses that adults. Another example of a covariate-parameter relation is a positive correlation between markers of renal (i.e. kidney) function and the renal drug clearance. A common marker of the renal function is the so-called creatinine clearance (CRCL) which can easily be calculated from the level of serum creatinine and patient demographic characteristics. A covariate model can be
used for identification of patient subpopulations at risk for sub-therapeutic or toxic effects and to subsequently individualise the treatment and the initial dosage regimen. Furthermore, such a model is useful for identifying the need for and aiding the design of new studies in the drug development process. If the identified covariate relations are in line with the literature or prior expectations, the covariate model supports the structure of the other parts of the model. Thus, the development of a covariate model may also be viewed as a component of the model evaluation.

1.2. Nonlinear Mixed Effects Models

A PK-PD model contains so-called mixed effects if it has two or more levels of random effects, i.e. where the random variability in the parameters is treated separately from the residual (intra-individual) variability. The term “mixed” is used because the model estimates both fixed and random effects simultaneously. Typically, the random effects form a hierarchy in the sense that the random components in the parameters are constant within an individual whereas the residual variability may change for each observation.

This hierarchical modelling approach offers several advantages over a so-called marginal-modelling approach, which is merely descriptive and only applicable to a rich design with observations at the same times in all subjects. As seen in Figure 1, the structural PK-PD model can be divided into several sub-models. The population (mixed-effects) model can also be subdivided as shown in Figure 2. The structural model describes the general makeup of the PK-PD model, which is usually the same for all individuals. The stochastic model describes the random component of the population-parameter variability (PPV) in the structural model, e.g. IIV, and the distribution of residual errors. The covariate model explains the predictable aspects of the PPV.

Figure 2. Illustration of a population model. The structural model may be a PK-PD model describing the profile of exposure/response to a drug over time for a typical individual. The stochastic model describes the random variability between and within individuals. The covariate model describes the relation between the covariates and the structural model parameters.
A wide variety of structural models are relevant to the area of PK-PD modelling. A few of these are described in the Methods section but no attempt is made to provide an overview of different structural models in this thesis. For this introduction, it is sufficient to describe the prediction of the dependent variable (exposure or outcome) in an individual according to

$$\hat{y}_i = f(x_{ij}, \hat{P}_j)$$  

(1)

where $x_{ij}$ are design variables (e.g. regimen and time) and $\hat{P}_j$ is the individual parameter vector for the $j$th observation in individual $i$. $f_j$ is a function of population parameters, random-effects parameters and covariates as described below. The subscript $j$ can be dropped if individual parameters are constant over time. This is normally how parameters are viewed: as constants that are estimated in the model. The independent variables that may be input to the model include, for example, dose, time and covariates. The dependent variables are predicted by the model. These may be drug exposure, response or outcomes that may not be measured directly, such as the sensitivity to insulin of a diabetic patient. The remainder of this subsection explains the concept of the stochastic model and the next subsection explains the covariate model. For brevity, the explanation is essentially limited to concepts relevant to the work in this thesis.

The random components are most often assumed to be derived from a parametric distribution. An individual parameter ($P$) is commonly distributed according to

$$P_i = TVP_i \cdot e^\eta_i$$  

(2)

where, regarding the parameter $P$ for subject $i$, $TVP_i$ is the typical parameter value and $\eta_i$ is the random effect that is normally distributed around zero, with standard deviation $\omega$, reflecting the IIV. This parameterisation assumes that the random variability around the typical value has a log-normal distribution, which is often reasonable for parameters that have a lower (physiological) boundary at zero. If the parameter also has a higher boundary at one, a logit transformation can be used to reshape the distribution. This will effectively restrict all individual parameter values between zero and one and is parameterised according to

$$P_i = \frac{e^{\eta_i(TVP_i/(1-TVP_i))}}{1 + e^{\eta_i(TVP_i/(1-TVP_i))}}$$  

(3)

Further, if some individuals have been observed several times on multiple occasions and there is a random variability in $P_i$ over time, Equation (2) can be expanded to include inter-occasion variability (IOV) according to

$$P_{ik} = TVP_i \cdot e^{\kappa \sigma_{\eta_i}}$$  

(4)

where, regarding parameter $P$ on occasion $k$ in subject $i$, $\kappa_{\sigma_{\eta_i}}$ is the random effect normally distributed around zero with standard deviation $\pi$, reflecting the IOV. The IOV can only be separated from the residual error if some individuals have been observed several times on the same occasion.

The model predictions ($\hat{y}_i$) of the observed values ($y_{ij}$) are associated with a residual error which can be additive, proportional/exponential, or a combination of the two.
Due to approximations in the estimation procedures and for numerical stability, the error may be assessed on log-transformed values. The three error models used in this thesis are described by

\[ y_j = \hat{y}_j \cdot (1 + \epsilon_j) \quad (5) \]

\[ \ln (y_j) = \ln (\hat{y}_j) + \epsilon_j \quad (6) \]

\[ \ln (y_j) = \ln (\hat{y}_j) + \theta^{\text{chl}} \cdot \epsilon_j \quad (7) \]

where, for the \( j^{\text{th}} \) observation in subject \( i \), \( \epsilon_j \) is the individually weighted residual error that has a normal distribution centred around zero and a standard deviation \( \sigma \). \( \theta \) represents the IIV in the residual-error magnitude\(^{15} \) and if \( \theta \) is not zero \( \sigma \) represents the typical residual-error magnitude within an individual. The residual error can represent assay/measurement error and errors in the recorded time and dose as well as model misspecification, e.g. non-adherence to therapy\(^{16-18} \) that is not accounted for. Usually, even for well controlled clinical trials, assay error does not constitute the major part of \( \sigma \).\(^{15} \)

Correlations between individual parameters or between residual errors can be included in the model. Individual parameter correlations not associated with covariates in the model are included as covariance between different random effects (\( \eta \)). Correlations between the intra-individual errors (\( \epsilon \)) may either be included as a function of time within the same dependent variable and individual (auto-correlation) or as a covariance of \( \epsilon \) between different dependent variables measured at the same time, e.g. observations of drug and metabolite or repeated assays of the same sample.\(^{15} \)

Unless otherwise noted, the random effects are assumed to be independently and identically distributed, i.e. without any correlation and not according to Equation (7) above.

### 1.3. The Covariate Model

A covariate model describes the relations between covariates and parameters. This introduction merely discusses the relation to structural model parameters. However, covariates may also influence the random effects distribution, e.g. the magnitude of the IIV or residual error\(^{15} \) or the probability of belonging to a certain mixture.\(^{19,20} \) A mixture could for example consist of poor and extensive drug metabolizers.

#### 1.3.1. What is a covariate?

Covariates are characteristics describing the patient, the conditions of the drug treatment or other factors potentially influencing the outcome. The covariates may be constant within an individual (e.g. sex) or changing over time (e.g. age). Potential covariates in a PK-PD model include:
Jakob Ribbing

- Demographics: age, weight, sex.
- Markers of organ function: CRCL, alanine aminotransferase, bilirubin.
- Environmental indicators: concomitant medication, smoking, season, gender.
- Others: disease state or progression, quantified on an assessment scale or by a biomarker; genotype of a metabolising enzyme or a binding site; other phenotype, e.g. the ratio of unbound and total drug in plasma ($f_u$); or other laboratory measurement.

In nlme models, dose and time are normally not considered as covariates but rather as variables that are an integrated part of the structural model. This may also be the case for a biomarker which is on the mechanistic pathway to the actual treatment response. The biomarker may be included as part of the structural model and this variable would then be observed with a residual error as a dependent variable. The covariates, on the other hand, are included in the model as if measured without any error, as independent variables.

### 1.3.2. What is a true covariate?

We can view all PPV as being predictable by a (large) set of latent variables. However, these latent variables may be unknown and immeasurable. On the other hand, the covariates that we actually do measure can be partially correlated with these and may either be causative to (e.g. sex) or affected by (e.g. creatinine clearance) latent variables. In this thesis, the pragmatic view is taken that a true covariate is one that, among the investigated covariates, carries unique information on a structural model parameter.

If data are generated by simulation from a model, the true covariates are those used in the simulation model, whereas all others are defined as false. However, in the event that one of the true covariates is not available for investigation, any other covariate that carries unique information on the unavailable covariate can instead be defined as true. Regarding real data, the distinction of true/false becomes pointless, since it is impossible to reject any covariate as false. Instead, one can discuss whether a covariate relation is important or clinically relevant, or if it in any other way supports the model.

### 1.3.3. Multiplicative and additive parameterisation

The contributions of different covariate effects on a structural model parameter are often combined as a multiplicative covariate model according to

$$TVP_{ij} = \theta_{\text{pop}} \cdot \prod_{\text{cov}=1}^{N_{\text{cov}}} (1 + \text{CovEffect}_{i,j})$$

where $\text{CovEffect}_{i,j}$ is the fractional change in the parameter due to the covariate $(\text{cov})$ in individual $i$ at time point $j$. $\theta_{\text{pop}}$ is the population typical value of parameter $P$ (i.e. $TVP_{ij}$) for a subject $i$ with all $N_{\text{cov}}$ covariates equal to the median at time $j$. The subscript $j$ can be dropped if the level of the covariate is constant over time, as for e.g. sex and race. An additive covariate model is described according to
Multiplicative parameterisation contains a fixed-interaction component. This fixed interaction can sometimes be useful for mechanistic reasons. One example of this is the two covariates body size and genotype of a metabolizing enzyme, and their relation to drug clearance ($CL$). A poorly metabolizing genotype that reduces a parameter by 50% combined with a small body size that reduces the parameter by the same degree will not together result in a 100% reduction of $CL$, but rather in a 75% reduction. However, the main rationale for using multiplicative parameterisation is that it is practical to set boundaries on $\theta_{P_{cov}}$ so that $TVP_{ij}$ is confined to positive values (zero often being a physiological boundary for the structural model parameters). This can also be applied to additive parameterisation. However, in the additive case it is only possible to avoid negative values of the structural model parameters for the combinations of covariate values that are in the dataset - not for the patient population as a whole. In this thesis, multiplicative parameterisation is used without exception. For brevity, a covariate model will sometimes be referred to as linear or piece-wise linear without mentioning the fact that it is multiplicative.

1.3.4. Categorical covariates

A categorical covariate can only attain two or more discrete values, or levels. If these can be ordered, the covariate is said to be measured on the ordinal scale. Examples of such covariates are tumour stages and categorization of a continuous variable. If the variable cannot be ordered, it is said to be measured on the nominal scale. An example of such a covariate is race. However, the distinction between ordered and non-ordered covariates is only important if the covariate has more than two levels. The effect of a non-ordered categorical covariate on a parameter $P$ can be expressed relative to the reference level according to

$$CovEffect_{P_{cov},ij} = \begin{cases} 0 & \text{if } cov_{ij} \text{ at reference level} \\ \theta_{P_{cov1}} & \text{if } cov_{ij} \text{ at level 1} \\ \theta_{P_{cov2}} & \text{if } cov_{ij} \text{ at level 2} \\ \theta_{P_{cov3}} & \text{if } cov_{ij} \text{ at level 3} \\ \vdots & \end{cases}$$

(10)

the coefficient $\theta_{P_{covL}}$ is the fixed effect that establishes the impact of belonging to level $L$ relative to the reference level. As an example, with the reference level at 0, an ordered categorical covariate can instead be parameterised according to

$$CovEffect_{P_{cov},ij} = \begin{cases} 0 & \text{if } cov \text{ at 0} \\ \theta_{P_{cov1}} & \text{if } cov \text{ at 1} \\ \theta_{P_{cov1}} \cdot (1+\theta_{P_{cov2}}) & \text{if } cov \text{ at 2} \\ \theta_{P_{cov1}} \cdot (1+\theta_{P_{cov2}}) \cdot (1+\theta_{P_{cov3}}) & \text{if } cov \text{ at 3} \\ \vdots & \end{cases}$$

(11)
where $\theta_{\text{cov},i}$ is now a coefficient describing the effect relative to the previous level, i.e. $L-1$. Restricting $\theta_{\text{cov},2}$, $\theta_{\text{cov},3}$, etc. to positive values confines the effect to be either increasing or decreasing as the level of the covariate increases. If this is not a reasonable assumption Equation (10) can be used, although the covariate is measured on the ordinal scale.

1.3.5. Continuous covariates

The effect of a continuous covariate (i.e. interval or ratio scale) on a parameter ($P_i$) is often expressed relative to its median in a relevant patient population, denoted $\text{cov}$. The most common functional forms of covariate relations are linear, piece-wise linear, power and exponential. The linear relation is parameterised according to

$$\text{CovEffect}_{\text{cov},ij} = \theta_{\text{cov}} \cdot (\text{cov}_i - \text{cov})$$

where $\theta_{\text{cov}}$ is the covariate coefficient, i.e. a fixed-effects parameter. The piece-wise linear relation with a break point at the median covariate value is parameterised according to

$$\text{CovEffect}_{\text{cov},ij} = \begin{cases} 
\theta_{\text{cov}1} \cdot (\text{cov}_i - \text{cov}) & \text{if } \text{cov}_i < \text{cov} \\
\theta_{\text{cov}2} \cdot (\text{cov}_i - \text{cov}) & \text{if } \text{cov}_i \geq \text{cov}
\end{cases}$$

The break point(s) can be estimated instead of pre-specified as the median. However, such an estimation may be unstable and consumes extra degree(s) of freedom ($df$). The functional forms for the power and exponential covariate models are

$$\text{CovEffect}_{\text{cov},ij} = \left(\frac{\text{cov}_i}{\text{cov}}\right)^{\theta_{\text{cov}}} - 1$$

$$\text{CovEffect}_{\text{cov},ij} = \theta_{\text{cov}2} \cdot e^{\theta_{\text{cov1}} (\text{cov}_i - \text{cov})}$$

Examples of possible shapes describing the above functional forms are shown in Figure 3.
Figure 3. Examples of shapes that can be formed by four common functional forms of a covariate relation. The piece-wise linear and exponential relations are more flexible and consume an extra degree of freedom (df) whereas the power and linear relations rely more on assumptions. The choice of which parameterisation to use depends e.g. on the amount and range of the data, the signal in the data and mechanistic preferences in combination with the purpose of the model.

1.3.6. Covariate transformation

Many nonlinear covariate relations can also be investigated in the framework of linear models by the introduction of dummy covariates. These include piece-wise linear models with pre-defined breakpoints, log-linear models, and ordered or non-ordered categorical covariates. For example, a non-ordered categorical covariate with three levels can be replaced by two dummy covariates, the first dummy being one if the original covariate is level 1 and zero otherwise, the second dummy being one if the original covariate is level 2 and zero otherwise. Instead of the categorical covariate, the two dummy covariates can be included as two linear models according to Equation (12).

1.4. Software for Model Fitting

Fitting the model means estimating the model parameter values resulting in the best fit to the available information (e.g. data). The traditional approach to estimating the parameter values for a population model is to perform the analysis in two stages, the
standard two-stage (STS) approach. First, the individual parameters are estimated separately for each individual. Then, the parameter variability is calculated based on the individual estimates obtained in the first stage. This approach requires a substantial number of observations in each individual, usually three to five observations per subject for each structural model parameter. However, this number is highly dependent on how informative the data are. For practical reasons and because of the often invasive nature of the observations (blood samples), the data possible to obtain from patients are often sparse, i.e., a few samples per individual. Therefore, the STS approach results in inflated (upwardly biased) estimates of IIV and is also limited to models with few parameters. In addition, the STS may provide biased estimates of the typical profile, since individuals with too few observations have to be omitted; for example, subjects with fast elimination will have few measurements above the limit of quantification. These problems taken together was the motivation for development of the NONMEM software that was introduced in 1980 as the first software performing hierarchical nlme modelling. For nlme models describing clinical data, it is still the most widely used regression software both in academia and industry.

All fitting of nlme models in this thesis has been performed using NONMEM software. This software estimates the parameters of a parametric nlme model, according to an approximate maximum likelihood. The same is true for the SAS-procedures NLMIXED (NLINMIX/NLMEM) using (non-adaptive) Gaussian quadrature, in S-Plus. Exact maximum-likelihood estimation is performed by NLMIXED using the adaptive-Gaussian quadrature in SAS, SAEM, MCPEM and PEM. All of the above software packages use a Bayesian (i.e., hierarchical) approach in the sense that the population parameters inform the individual parameters; cf., for example, Equation (2). NONMEM can also make use of a so-called frequentist prior to propagate the prior knowledge of the parameters into the current analysis. Full Bayesian methods make use of the prior parameter distribution without parametric assumptions. Software packages for this approach include WinBUGS and MCSim. Software packages for nonparametric maximum-likelihood estimation include NPEM and NPML.

Despite the fact that the developers have already run out of unique abbreviations and acronyms for the words Nonlinear Mixed-Effects Models, there is still continuous development of new software for nlme models. Further, many of the old software packages are being upgraded with new functionalities and new estimation methods. In the next subsection we focus on the NONMEM estimation methods that have been used in this thesis.

1.5. Estimation Methods in NONMEM

Parameter estimation in NONMEM is based on maximizing the likelihood of the data, given the model. An iterative search within the parameter space terminates at a maximum likelihood and the parameter estimates obtained in this manner are called the maximum-likelihood estimates (mle). An approximation of the likelihood is obtained via the extended least squares objective function value (OFV). Assuming that the random effects are normally distributed, the OFV is “up to a constant” equal to minus twice the natural logarithm of the likelihood.
The random effects on the nonlinear structural model parameters often preclude the obtainment of an analytical solution to the objective function. In NONMEM, this is approached by approximating the OFV by a linearization of the nonlinear model. The first-order method (FO) uses a first-order Taylor-series expansion at \( \eta = 0 \) whereas the first-order conditional estimation method (FOCE) performs the same linearization with respect to \( \eta \). Interactions between \( \eta \) and \( \varepsilon \) can be accounted for by using the FOCE with interaction (FOCE-I). In this thesis FOCE-I was used for application of the error models in Equations (5) and (7). In the former, the interaction is implicit since \( \varepsilon \) is multiplied by the model prediction which in turn is a function of the \( \eta \), cf. Equations (1) and (2).

1.5.1. Individual empirical bayes estimates and shrinkage

NONMEM provides the individual parameter estimates, e.g. \( P_i \) or \( \eta_{iP} \), as empirical-Bayes estimates (EBE) generated during the estimation procedure (e.g. FOCE and FOCE-I) or posthoc (FO). These individual parameter estimates may tend to shrink towards the population typical value (\( \eta_{iP} = 0 \)) in comparison with the true individual parameter values. This reduction in the variability of an EBE is called \( \eta \)-shrinkage. An extreme example of shrinkage is PK and drug-specific PD parameters in patients who receive placebo instead of active treatment. Since, for these parameters and in this context, the data from the placebo recipients contain no information, the individual estimates rely solely on the typical parameter value, \( TVP_i \). This means that all available information comes from the total population, resulting in complete shrinkage in this group. However, the shrinkage induced by sparseness of data (information about a parameter) is more harmful than a complete lack of information. This will be discussed further in subsection 1.8.1.

The degree of \( \eta \)-shrinkage can be assessed as the standard deviation of the \( \eta_{iP} \) relative to the IIV estimated for the same parameter in the same dataset. Alternatively, if \( \eta_{iP} \) is biased, i.e. not centred around zero, the root mean square can be evaluated instead of the standard deviation, according to

\[
\text{shrinkage}_{\eta_{iP}} = 1 - \frac{\sum_{i=1}^{N} \eta_{iP}^2}{\sqrt{N \cdot \omega_p}}
\]  

(16)

where \( N \) is the number of subjects in the dataset. Bias in the \( \eta \) is of interest for diagnostic purposes. Equation (16) does not indicate any shrinkage due to bias, since in this thesis, the degree of shrinkage and bias in the \( \eta \) are treated separately.

1.6. Model Development and Selection

In drug development, model-based population analysis is often used in an exploratory manner for “learning”, but model-based confirmatory analysis is also increasingly applied. Whereas a confirmatory analysis is focused on testing a pre-specified hypothesis, an exploratory population analysis can take new concepts and ideas into account during the process. This is achieved by investigating a large number of hypotheses based on, for example, informative graphics as well as statistical testing.
Often, the model is developed in several stages. In the first stage, the structural and stochastic models are developed, starting with a simple model and expanding the complexity when supported by the data. Highly influential or pre-specified covariate relations can be included in the model even at this early stage. Estimation of IIV for all structural model parameters and correlations between all random components is seldom possible in NONMEM because of lack of information in the data. Thus, there is often considerable work involved at this stage to develop the appropriate structural, IIV, and residual-error models. The first-stage model is called the base (or basic) model in this thesis.

In the second stage, the explanatory value of the covariates is investigated on the parameters of the base model. Covariate relations that explain part of the IIV in a parameter can be included in the model at this stage. In the third and last stage, the stochastic model is re-evaluated and refined. Preferably, other parts of the model are constantly re-evaluated throughout the developmental process. Even so, there is a problem with this stepwise approach to model selection in that the selection of one model feature may be conditional on another model feature. This interaction between different parts of the model can be handled by investigating all combinations of all model features. An attempt to achieve this using a genetic algorithm (GA) has been made. However, since the number of combinations is often vast; this approach is currently only possible in rather limited modelling exercises. Furthermore, the criteria for selection that can be used in an automated procedure are also limited at present. The most common approach is to increase (or decrease) the model complexity in a stepwise manner. This is explained further in the next subsection.

The current standard analysis results in a single nlme model (the final model) rather than a set of possible models. This final model summarizes and quantifies the knowledge gained from the analysed data, possibly also including information (data) or knowledge (parameter distributions or values) from analyses of other studies. Inferences are drawn from the final model as if no model selection had been performed based on the information (data) that was used to estimate the model parameters.

Depending on the purpose of the model, the criteria for selecting one model over another can include

- mechanistic plausibility and prior beliefs: consideration of model structure and parameter estimates, the former also setting a limit to what is investigated in the first place
- plots displaying goodness-of-fit and other graphics
- successful convergence or obtainable covariance matrix of the estimates
- statistical significance: p-value, Akaike’s information criterion (AIC), Schwartz’s Bayesian criterion (SBC), etc.
- relevance: clinically unimportant features may be ignored in favour of model parsimony
- simulation properties
- predictive performance: in internal or external validation
- parameter precision: e.g. avoiding collinearity due to the data/model
- influential individuals: is the model feature promoted by only one or a few individuals?
- other model fit as diagnostics
- other practical matters: e.g. computer run-time
1.6.1. Stepwise selection

Stepwise-selection procedures currently dominate model selection in nlme modelling. This approach can be exemplified by a system of selection based solely on p-values. It should, however, be noted that the criteria can include other more subjective conditions or terms, as mentioned above. If FOCE (or FOCE-I) is used, the difference in OFV between two hierarchical models is approximately chi-square distributed with the appropriate df.73, 74

In forward selection, the model complexity (model size) is increased from a simple model. The features of interest that can be fitted are evaluated by including them one at a time into the nlme model. The feature that performs best according to the p-value is included into the model if it is statistically significant. Subsequently, all other model features are re-evaluated in the new model and a second feature is included if it is significant. The procedure stops at the full-forward model when no further model features are statistically significant. An alternative approach is backward elimination. The starting point is a model that initially includes all features of interest, called the full model. Elimination is performed until only statistically significant features remain.

When this approach is possible, the end result is often as good as an all-subset selection, which investigates all combinations of the different model features of interest. This is not always the case with forward selection.75 However, the presence of features that are mutually exclusive or inestimable in combination due to correlations between the estimates often prevents the use of backward elimination techniques. Further, many ideas that are generated during an exploratory analysis cannot be included in the initial model. Because of this, backward elimination can only be used for part of the model-building procedure. Another alternative is to use forward selection with a less strict (i.e. a higher) p-value. The full forward model obtained in this manner is then refined by backward elimination using the same or a stricter p-value. This procedure is called “forward inclusion and backward elimination” and is commonly applied in the selection of a covariate model.10

1.7. The Base Model

The base model is the starting model in a covariate investigation. The model can contain covariates, however only for pre-specified covariate relations included in the model without testing. The pre-defined covariate model can be fixed so that no covariate parameters are estimated or the structure of the covariate model can be fixed while the parameters are estimated from data. An example of the former is assuming an allometric relation between PK parameters and a measure of body size.76 Estimating the covariate model parameters is necessary when information about the covariate relationship is qualitative but insufficiently quantitative. An example of such a situation is the use of CRCL77 as a covariate to explain variability in the clearance of a drug which is known to be mainly eliminated via glomerular filtration.12
1.8. Selection of the Covariate Model

In an exploratory analysis, there are often a large number of covariates available which would be interesting to test on one or more structural model parameters. There are several reasons for selecting a subset of these instead of including all of them in the final model. The first is that a model containing all the potential covariate-parameter relations may not be estimable, i.e. the estimation would not converge in NONMEM. Second, most of these potential covariate effects are unimportant. Consequently, to obtain an overview of which covariates are important, the less important or uncertain covariate relations should be removed. Also, estimation of uncertain relations is imprecise and removing these improves the performance of the model for predicting external data, i.e. improves the predictive performance. Naturally, predicting the outcome in new patients or future events is often an important objective of the model.

Another reason for reducing model complexity is that potential covariates are often correlated. If all of them are included a high correlation between the estimates of the covariate coefficients arises. These coefficients are then estimated with high imprecision, making it impossible to decide which of them are important or clinically relevant. Extrapolation of such a model outside the current range of the data can therefore be unfortunate. One solution to this problem is to reduce the number of investigated covariates by removing redundant (correlated) covariates based on the relevant current literature and to retain the rest in the model, i.e. a pre-specified model. This is a good approach for obtaining correct confidence intervals and for summarizing current knowledge of clinical relevance. However, this approach does not explore different functional relations or which of the correlated covariates are important.

For the reasons mentioned above, a subset of the potential covariate relations is often selected for the final model. The actual selection of which relations to include can be made after investigating the results of including each of them in the nlme model. These procedures are collectively called selection within NONMEM. Alternatively, the outcomes of the different relations can be investigated outside NONMEM, after which one or several selected models are investigated within NONMEM. The latter alternative, which is called selection outside NONMEM, has not been investigated in this thesis work but is extensively discussed in next subsection.

1.8.1. Covariate selection outside NONMEM

The main advantage of selecting the covariate model outside NONMEM is that an investigation within NONMEM is computer intensive, resulting in long computer run-times and/or high demands on a computer grid. Therefore, a seemingly appealing approach is to perform graphical inspections of the relations between covariates and the EBE of individual parameters to find the relevant relations. However, if data are sparse, this may lead to shrinkage of the EBE towards the typical parameter value so that a clinically relevant relation may become distorted in its shape or appear as unimportant or falsely important.

A statistical evaluation of a relation can be used to pick up even weak trends that may be invisible in a graphical inspection. This is commonly used for identifying the covariate model using generalized-additive modelling (GAM). This approach is not
appropriate for handling time-varying covariates or randomly time-varying parameters, i.e. those with IOV. Investigating covariates on parameters that do not have any IIV according to the base model is not possible at all. Further, the issue of shrinkage creates additional problems unless the data contain rich information on the investigated individual parameters. A modest correlation between the structural model parameters can result in a false correlation between the individual parameters. Thus, a relation between a covariate and one parameter may induce a false relation between the covariate and several other parameters. Further, shrinkage is more pronounced for the individual parameter values that are implausible according to the nlme base model. This changes the apparent shape of the relation between covariate and parameter. This is illustrated in Figure 4. Due to shrinkage, covariate relations may be hidden, induced or distorted when using GAM.

![Figure 4](image)

**Figure 4.** Illustration of the effects of shrinkage. The left panel displays the individual estimates of CL from the model including an important relation between CL and CRCL. The right panel displays the same estimates from a model not including the covariate relation. Each circle represents a separate individual. The solid line is a smooth of the data whereas the broken line represents the expected CL according to the model used for simulation. Including the covariate relation, the two lines agree well. However, if not including the important relation, the shrinkage increases from 0.37 to 0.77 and the quality of the individual estimates is reduced. As a result, the relation between CL and CRCL is distorted and appears to be nonlinear. The data was simulated so that information on CL is independent of the value of the parameter, meaning that this is a pure effect of shrinkage.

When GAM or graphical analysis is used for covariate selection, it is advisable to evaluate the covariate relation in the parameter space and not the relation with $\eta_i$. A linear relation in GAM will otherwise translate to an exponential relation in the nlme model, given that IIV is distributed according to Equation (2). If any covari-
ate is already included in the nlme model, this should be accounted for by regressing \( P_{-TVP} \) on \( cov \). Alternatively, for additive-covariate parameterisation, \( P_{-TVP} \) should be used, cf. subsection 1.3.3. If shrinkage is a problem for any of the investigated individual parameters, two further actions can be taken to slightly reduce these effects. The first is to evaluate other functional relations (shapes) in addition to the one selected by GAM. More importantly, only the most significant relation in GAM should be included in the nlme model. Then, after fitting the new nlme model in NONMEM, a new GAM analysis can be made based on the new individual parameters. These two steps are performed until no relation is found by GAM or until inclusion in NONMEM is no longer supported. Alternatively, all relations identified by GAM can be included in the nlme model initially and then subsequently re-evaluated using backward elimination within NONMEM.

A completely different approach performing the covariate selection outside NONMEM is to use Wald's approximation to the likelihood ratio test (WAM).\(^59\) This method requires the point estimates and the covariance matrix of the estimates from a full model fit including all covariate relations of interest. If these can be obtained, WAM may be used to quickly perform an all-subset selection, investigating all possible combinations of covariates and parameters. The most promising models are identified outside NONMEM by this approach and subsequently evaluated within NONMEM. The main criticism of this approach is that it is often impossible to obtain the variance-covariance matrix of the parameter estimates for the full model from NONMEM. It may also be difficult to obtain these by non-parametric bootstrapping due to long run times, parameters ending up at the boundaries, or other failed convergences in NONMEM. These problems are partly due to a tendency of pharmacometricians to investigate more relations than supported by the data.\(^79\)\(^80\) These problems seem to be the main reason for the less wide-spread use of WAM, compared to alternative methods. Another potential objection to WAM is that it only selects covariate-parameter combinations whereas the functional form, the originator suggests, should be determined using graphical inspection.\(^59\) As mentioned above, this is inappropriate for individual parameters with extensive shrinkage. Further, the graphical approach has additional complications when investigating the functional form in a multivariate setting. The WAM originator further concludes that the approximation may work less well for models with a high degree of nonlinearity. Unfortunately, these include many PD models and nonlinear PK models (i.e. nonlinearity in the differential equations). The WAM is an appealing method for covariate selection where the degree of nonlinearity is low and either investigation of different functional forms is unwished or possible to investigate without problems of shrinkage or covariate correlations.

1.8.2. Covariate selection within NONMEM

Covariates are often selected within NONMEM in a stepwise manner, e.g. using the procedure stepwise covariate modelling (SCM).\(^10\) The stepwise procedures have been extensively investigated in traditional statistics and a few of the associated problems are outlined below.

SCM often uses a p-value as an indicator of when to halt inclusion or deletion of further covariate coefficients, i.e. as a stopping rule. In general, several covariates
are investigated, possibly on a number of structural model parameters and in several different functional forms. Therefore, the overall type-I error rate (i.e. the probability of including one or more false covariate coefficients into the model) is much higher than indicated by the required p-value. This is a problem of multiple comparisons. To correct for this, a stricter p-value is often used in the selection, although this can in turn result in omitting relations that are actually important. Further, correction of the p-value is only approximate or even arbitrary. Because of correlations, finding the value that corresponds to the overall type-I error rate is very computer intensive. Thus, although the p-value is used as a criterion for selection based on the ideas of hypothesis testing, the actual strength by which the null hypothesis has been rejected is unknown in the case of multiple comparisons. When predictive performance is the main objective, a stopping criterion focusing on this can be expected to perform better. Cross-validation is such a criterion.

The coefficients selected by SCM are exaggerated because of selection bias. A relation that seems important is often statistically significant whereas one which by random chance seems less important is left out. In this manner, the selected relations are on average more important than they would have been if the full model had been estimated without selection.

A systematic difference is called bias. In this context it is called selection bias since it is caused by two elements in the selection procedure; the requirement of statistical significance and the competition between correlated covariates. An illustration of extreme selection bias is presented in Figure 5. Selection bias can decrease the predictive performance of a model. This is a problem for all selection procedures which do not include shrinkage of the selected covariate coefficients, so that the magnitudes of the coefficients are reduced compared to the mle. We use the term “shrinkage” for both penalized estimation of covariate coefficients and the η-shrinkage in that arise due to sparse data (cf. subsection 1.5.1). Although the two kinds of shrinkage are sound for the same reasons they are treated as separate issues in this thesis.

The problem of selection bias due to SCM has been investigated for a typical PK analysis. The conclusion was that selection bias seems to be a minor problem. However, the simulations in this investigation were based on a model that was obtained by applying SCM to a typical PK dataset. Thus, the coefficients in the model are expected to be biased to a larger magnitude, increasing the inclusion rates and decreasing the selection bias when the selection was subsequently investigated on the simulated data. Except for this case study, the problem of selection bias has not been investigated in nlme models. The degree to which the investigation of several covariates can further inflate the selection bias due to competition is also unclear.
Jakob Ribbing

Figure 5. Illustration of the distribution of estimates from 7400 replicated studies simulated with a weak covariate coefficient. The grey bars represent all 7400 estimates whereas the black bars represent a subset of 611 estimates where the covariate was retained in the model due to statistical significance. The selection bias is very clear in this example. All estimates are exaggerated when selected from a dataset of such a small size (20 subjects). The distribution of all estimates (grey bars) is centred around the true value of the covariate coefficient.

Associated with selection bias, SCM is very categorical when selecting covariates. A relation is either included according to the mle or completely excluded from the model. This categorical selection leads to highly variable estimates and reduces the predictive performance of the model. A selection procedure such as the least absolute shrinkage and selection operator (lasso) method, which shrinks the coefficients for uncertain relations has been found to perform better in this context.84

Compared to all-subset selection, the stepwise approach may not come across the optimal combination of predictors in its search path. This can especially occur when forward selection is used on a set of predictors that perform well together but poorly alone.85 This problem can be overcome by starting a backward elimination of the full covariate model.75 However, with many covariates or different functional relations to investigate, this approach becomes impossible or at least very time consuming.

SCM allows excessively many relations to be tested. The final covariate model may not show any signs of overfitting, even though too many hypotheses were tested before arriving at the final model. Translating the general advice in traditional statistics79, 80 to covariate selection in mixed-effects modelling indicates that it often harms the predictive performance of a model if more than one covariate parameter per 10-20 individuals in the dataset is investigated on each structural model parameter. Fewer covariate parameters should be investigated for parameters on which information is sparse or when categorical covariates are investigated. However, this rule-of-thumb is highly dependent on the quality of the hypothesised relations and thus highly dependent on the drug, the investigated patient population, and the therapeutic area. The number of investigated covariate parameters includes, e.g. different functional forms for a relation that has been considered in a graphical analysis prior to SCM. Because of this, the actual number of investigated covariate parameters is often unknown.
1.9. The Lasso in Ordinary Multiple Regression

The lasso method is a penalized estimation technique for linear models. However, many non-linear covariate relations can also be investigated in the lasso framework, by the creation of dummy covariates as described in subsection 1.3.6. Before using the lasso, the covariates must be standardized to zero mean and standard deviation one. Then, the lasso estimates of the regression coefficients are as in ordinary least-squares regression but subject to restriction on the magnitude of the coefficients according to

\[ \sum_{k=1}^{N} |\beta_k'| \leq t \]  

(17)

where \( \beta_k' \) is the regression coefficient operating on the standardized covariate \( k \), and the amount of shrinkage is determined by the value of \( t \), the tuning parameter. In an nlme model, \( \beta_k' \) corresponds to \( \theta_{cov} \) in Equation (12).

For low values of \( t \), the implication of this restriction is that some covariate coefficients are slightly shrunk compared to the maximum likelihood estimate, whereas others are shrunk all the way to zero. The latter situation is the same as eliminating the covariate relation from the model. For the lasso, the value of \( t \) determines the model size. The value of \( t \) yielding the model with the best predictive performance can be estimated using cross-validation as outlined below. Illustrations of the effects of applying different degrees of shrinkage in the lasso estimation are shown in section 3.3 where the lasso in NONMEM is described.

The lasso is closely related to the ridge regression method, where the squared coefficients, rather than the absolute coefficients, are subject to restriction. The rationale for using absolute coefficients is that model selection and estimation shrinking occur at the same time, thus resulting in a parsimonious model. However, the lasso offers no advantages over ridge regression with respect to predictive performance.

1.10. Model Validation and Evaluation

A population model is often evaluated with respect to how simulations reflect observations or a relevant statistic or parameter calculated from these. The predictive performance of a model can be evaluated more simply when comparing differences in the covariate model. The prediction error is often evaluated on the observations or on the individual parameter estimates obtained after setting all random effects to zero, i.e. the first-order approximation PRED in NONMEM. Another measure of the predictive performance can be derived from evaluating the likelihood of the data given the model, estimated with fixed population parameters. In this work, the term validation is sometimes used instead of evaluation. This is done in analogy with the term "cross-validation". A distinction between evaluation and validation is not intended; Validation would otherwise imply an evaluation of the model for its purpose.

The predictive performance of a model is called a model error if evaluated on a parameter with a known value. The value is only known when the data have been generated via simulation from a model. Otherwise, the predictive performance is called a prediction error and is evaluated on, for example, \( y_i \) or \( P_i \) for all observations or
individuals, respectively. A prediction error can never reach zero since part of the PPV is random, i.e. IIV cannot be explained by the collected covariates. For the prediction error on $y_{ij}$, random IIV contributes further to the part of the prediction error which cannot be predicted by the model.

In this thesis, model validation means external validation, i.e. the use of external data for estimating the actual prediction error. Data were considered external if they had not been used for model selection or parameter estimation. Depending on the purpose of the modelling, it may also be necessary for the external data to come from a separate study or to represent an extrapolation in some other manner, e.g. the prediction of observations of long-term treatment based on a model built on data from short-term treatment. Evaluating a model on the same data that were used for model development yields only an apparent prediction error. This is biased low compared to the actual prediction error and always favours increased model complexity. As an example, if a parameter is added to the structural part of a model, the likelihood of the data given the model will increase and the difference between observation and prediction will decrease, since the model is fitted to these data.

1.10.1. Cross-validation

Cross-validation is a procedure for estimating and comparing the prediction errors of one or several models of different sizes. Typically, it is used to select a model of appropriate size, i.e. the model size that results in the closest prediction of external data. If the main goal is predictive modelling, cross-validation can be used in model selection, instead of a p-value, AIC or another stopping criterion. In this thesis, cross-validation is only used in the lasso method. In this subsection, however, the cross-validation procedure is described in a general manner which is valid for various subset-selection procedures. Consequently, model size here denotes the t-value for the lasso method. However, if cross-validation were to be run in SCM, model size should be interpreted as the number of covariate parameters.

Cross-validation is similar to data splitting but uses the data more efficiently to produce a more precise estimate of predictive performance and the most appropriate model size. However, unlike data splitting, cross-validation requires the an automated model selection procedure and validation is conditioned on any model-selection decision that has been made outside of this procedure. An example of this is developing the base model using a certain dataset and subsequently using the same dataset for developing the covariate model with cross-validation. In this case, the cross-validation estimate of prediction error may be too optimistic in case the development of the base model was highly driven by the data at hand.

The procedure for five-fold cross-validation of model size is as follows. The dataset is split into five parts containing roughly equal numbers of subjects. A training dataset is constructed by pooling four of the five parts. Model selection (to a certain size) and estimation is made on the training data and the prediction error is calculated on the fifth part that was omitted, the test dataset. The procedure is repeated until the five training datasets have been used to obtain five different models of the same size, each evaluated on the corresponding complementary/external test dataset. The prediction errors in the five test datasets are combined (e.g. added or averaged) to obtain the cross-validation estimate of prediction error for that specific model size.
The appropriate model size is determined by comparison of the estimated prediction error for different model sizes, choosing, e.g. the size with the lowest prediction error. The five training and test datasets are the same for all investigated model sizes. After cross-validation, the original dataset is used to select and estimate a final model of the appropriate size. The model selection procedure is exactly the same for selection in the five training datasets and the original dataset. Compared with data splitting five-fold cross-validation achieves both a better estimate of prediction error and a better selection of the model size, however at the cost of approximately five times longer computer run-times. Data splitting requires more data to be withheld for validation and is also a less efficient use of the data.67, 75, 87

There is another procedure which can also be called cross-validation. Apart from the procedure described above, “cross-validation the right way”, which is used in this thesis, there is a less computer-intensive alternative that has been called “partial cross-validation”91 or “cross-validation the wrong way”94 but other naming conventions also exist.93 This kind of cross-validation is only a validation of the parameter estimation conditioned on the model selection whereas the effects of model selection are not taken into account. The benefits are shorter computer run-times and no necessity for automated model selection. However, the estimate of prediction error is overly optimistic and the method is inconsistent in the sense that when a large number of nonsense (random) covariates are investigated, the selected model size becomes larger than when only a small number are investigated.

Since the total model selection, including structural and variance components, is seldom completely automated in NONMEM (cf. however Bies et al.50), cross-validation the right way has never been completely performed in the development of a nlme model. However, there is one example where the covariate selection has been cross-validated the right way within NONMEM.82 In the lasso, covariate-model selection is performed simultaneously with estimation of the model parameters, and cross-validation of the covariate model can only be performed in the correct manner. However, the outcome is still conditioned on the development of the base model, unless the structure has been pre-specified.

The data split made for creating the training and test datasets does not necessarily have to be random. Stratification can be used to improve estimation of model components where, e.g. sampling times or covariates make information provided by some subjects more useful than others. On the other hand, if prediction across studies is the main goal and several studies are available in the dataset the split can be performed between studies instead of between subjects.

With regard to model validation in general a new external dataset that has not been used in any way to build the nlme model is termed a validation dataset in this thesis. This is not the same as a test dataset which is used in cross-validation and is available during the model development. Further, if a model is successfully validated on the validation dataset, the model is often updated by re-estimating the parameters on the pooled data (including the validation data). Although this is a sensible strategy, the argument can be made that the model containing the updated parameter estimates is left unvalidated.88 Using the same argument, cross-validation performed the right way and across studies delivers an external validation of the model size. However, the final model obtained using this size is merely considered to be internally validated.
1.11. Modelling of Type 2 Diabetes

Insulin plays a major role in the regulation of blood glucose. This hormone is produced in the pancreatic beta cells. The response (sensitivity) of obese subjects to insulin is often impaired, so that a higher level of insulin is needed to maintain normal glucose levels. This so-called insulin resistance has been linked to increased levels of free fatty acids (FFAs) in the blood. Glucose regulation is normally maintained by increased secretion of insulin. However, along with increased FFA levels, reduced elimination of insulin also acts as negative feedback in subjects with insulin resistance.95, 96 For the increased insulin secretion to be maintained over an extended period, the increased activity of the available beta cells is not sufficient, the total mass of beta cells must also adapt to meet the higher demand for insulin. If this adaptation fails, the blood glucose levels become too high (hyperglycaemia). This gradual development of disease, called type 2 diabetes mellitus (T2DM), is the most common form of diabetes.

Fasting plasma glucose (FPG) and the fraction glycosylated haemoglobin A1c (HbA1c) are used as biomarkers to assess short-term and long-term glycemic control, respectively. Additionally, the endogenous fasting insulin (FI) level can be used to obtain an estimate of insulin sensitivity and beta-cell function. Population PK-PD modelling has been used to characterize relationships between drug exposure and biomarkers in T2DM.97-99 Applying a mechanistic PK-PD model allows the use of data from a variety of studies with heterogeneous patients and various experimental conditions. A mechanistic model that describes these highly variable conditions would also be able to capture all the available information in one model. Such a model would also be expected to better predict the outcome of new types of studies. One such prediction that is of interest concerns the effects of different anti-diabetic therapies on the 10-year progression of the disease, based on data from patient studies of one year’s duration or less. This approach is appealing not only because the shorter studies finish more quickly, are better controlled and do not suffer from the same degree of (informative) dropout/discontinuation or changes in therapy,100-102 but also because of less complex ethical considerations and lower costs.

Recently, the development of mechanism-based models for T2DM has accelerated. The starting point was a rather empirical model provided by Frey et al.97 describing the relation between gliclazide exposure and the effect on FPG over time in T2DM patients by incorporating an effect compartment to characterize the pharmacodynamic delay. A more mechanistic approach was presented by de Winter et al.98, who developed a population PD model focusing on disease progression, describing the interplay among FI, FPG and HbA1c. This model was based on two large phase III studies in drug-naïve T2DM patients that investigated the effects of one year’s treatment with pioglitazone, metformin or gliclazide. The model incorporated components for beta-cell function and insulin sensitivity, distinguishing immediate treatment effects from effects on long-term disease progression. Hamrén et al.99 incorporated the effect of red blood cell ageing and glycosylation of Hb into a population PK-PD model based on a phase II study investigating the effects of 12 weeks’ treatment with tesaglitazar in both drug-naïve and previously treated patients.

Tesaglitazar is a dual peroxisome proliferator-activated receptor (PPAR) α-γ agonist, previously in development for treatment of T2DM. Clinical development was discontinued in May 2006 when results from phase III studies indicated that the overall benefit-risk profile was unlikely to give patients an advantage over currently avail-
Tesaglitazar activates PPAR\(\alpha\), which increases insulin sensitivity in liver, fat and skeletal muscle cells, increases peripheral glucose uptake and decreases hepatic glucose output, similar to the effects of other PPAR\(\gamma\) agonists such as rosiglitazone and pioglitazone\(^{103}\).

Whereas the model proposed by Winter et al. includes FI, the other models mentioned do not. Further, none of the population models takes into account the possible adaptation of the beta-cell mass (BCM)\(^{104}\). When treatment with PPAR agonists is initiated, the response in FI is relatively rapid and a pseudo-steady state is reached within weeks after treatment initiation. However, the decline in FPG is slower, with a pseudo-steady state reached only within months\(^{98}\). This pattern is probably the result of an upward adaptation of the BCM simultaneously with the faster improvement in insulin sensitivity following improved lipid metabolism\(^{105, 106}\). For accurate predictions of long term treatment effects, this adaptation has to be separated from the underlying disease progression that leads to reduction of BCM and insulin sensitivity. Topp et al. suggested a mechanistic model integrating BCM, insulin and glucose dynamics\(^{104}\). To our knowledge, the entire model, derived from different sources in the literature, has never been fitted simultaneously to data. Furthermore, Topp et al. highlighted that the model incorporates neither the effects of anti-diabetic treatment nor all known physiological effects.


2. Aims

The aims of this thesis were to evaluate presently used methods for covariate model building within nonlinear mixed effects modelling, with a special focus on selection bias and predictive performance, and to develop and evaluate alternative procedures and methods for situations where current methods perform inadequately. Furthermore, the thesis also aimed to implement a mechanistic model better describing the dynamics of fasting plasma glucose, fasting insulin and beta-cell mass on and off treatment in a heterogeneous population, ranging from insulin resistant patients to those with advanced diabetic disease.
3. Methods

In paper I, problems with the use of SCM in population pharmacokinetics were investigated. Paper II describes the development of a new method for covariate selection within NONMEM: the lasso. This method can remedy many of the identified problems associated with SCM and other methods not including shrinkage in the estimation procedure. Subsequently, the lasso is compared with SCM. In paper III, different methods of incorporating information from multiple studies into PK-PD models (propagating knowledge) were investigated, as another way of reducing the problems with SCM. The conclusions from these three methodological papers were derived by generating replicated datasets (replicated clinical studies) under different conditions. This was repeated for covariate analysis using different methods. Inferences can be drawn based on the systematic differences between the methods. The fourth paper is different from the others in that it is an applied modelling example with the primary aim of developing a more mechanistic model for describing long-term glucose and insulin dynamics in normal and diabetic patients. The covariate model is developed for handling major differences in the heterogeneous patient population in the data obtained from several clinical trials.

3.1. Software

NONMEM version VI-beta and version VI were used for simulation and estimation of nlme models. The simulation of a multivariate covariate distribution was performed in MATLAB (paper I). Graphics and non-trivial statistical calculations were made in R and Splus. Automated administration and other data manipulation were performed in Perl. Automated interaction with the NONMEM software was achieved using Perl and the Perl-speaks-NONMEM (PsN) application-programming interface. Computer-intensive statistical methods using NONMEM repeatedly were run using the PsN toolkit. Design optimization of population studies was achieved using PopED for MATLAB.

3.2. Analysed Data

The data used in this thesis work was real data originating from clinical trials in drug development as well as simulated data obtained from an nlme model or by resampling from a real dataset. By “simulation from a model” is meant generating data, given the input variables, model and model parameters.
3.2.1. Simulated data (Papers I-III)

In paper I, the covariate values were obtained by simulation from a multivariate standard normal distribution, truncated at ±4 (standard deviations). Covariate 1 was regarded as a true covariate and the other four as false, but with varying degrees of correlation with the true covariate. Drug concentrations were simulated using a one-compartment intravenous bolus model. The structural model included only two parameters describing a mono-exponential decline. The influence of the true covariate was varied including no influence. Studies were simulated with different numbers of subjects and observations per subject, different magnitudes of the effect of the true covariate, different residual-error magnitudes, and IIV. Each of the above scenarios or simulation setups were simulated with 7 400 replicates. This means that for a certain scenario, 7400 studies of identical design were generated. These only differed due to random chance.

In paper II, six covariates were sampled with replacement from an empirical covariate distribution. The sampling was performed as individual row vectors so that the correlation structure between the covariates was preserved. Datasets for covariate analysis, analysis datasets, were generated by sampling individuals with replacement from a large real dataset. Since the data were resampled in chunks of all or none of the data records for each individual, all structures of the original real data were preserved. However, these structures but need not necessarily be found in each analysis dataset, since these were much smaller. Since the data was not generated via simulation from a model, there was no “true model” associated with the data simulated in this manner. The result from analysing the analysis datasets was evaluated on external data. For each analysis dataset, a corresponding validation dataset was sampled. The validation dataset consisted of all patients in the large real dataset that had not been used in the corresponding analysis dataset. In this manner, the validation dataset was slightly different for each analysis dataset. Further, analysis datasets and their corresponding validation datasets were generated with different numbers of subjects, each with 100 replicate datasets.

In paper III, two consecutive studies were simulated according to several scenarios. The simulation was performed with different numbers of true covariates and different magnitudes in the covariate coefficients, in conjunction with different numbers of subjects in the first and second study. In one scenario, the second study was designed based on what had been learnt from analysing the first study. The second study was then optimized for patient stratification and observation times according to an ED-optimal design experiment based on the model from the first study. Loosely stated, a D-optimal design is the design that yields the most information on the model parameters given a set of parameter values, i.e. point estimates. An ED optimal design is the design that is expected to yield the most information on the parameters given the parameter values and the associated parameter uncertainty.

Four observations per subject were used except in the optimized design, where two or three observations were used. Drug concentrations were simulated from a one-compartment model with first-order absorption and one or more covariates affecting CL. Each scenario was simulated with 500-10 000 replicates and each replicate contained one unique realisation of both the first and second studies.
3.2.2. Real data (Papers II and IV)

In paper II, the lasso is illustrated by application to a real dataset. This dataset was not used to draw inferences on this method compared to the SCM, but was merely used as an illustration. This dataset had frequently been investigated in the past.\textsuperscript{11, 14, 74}

In paper IV, three completed clinical phase II/III trials with tesaglitazar were available for analysis: the Study in Insulin Resistance (SIR), a dose-finding study in non-diabetic individuals with hypertriglyceridaemia and abdominal obesity, i.e. signs of insulin resistance;\textsuperscript{114} the Glucose and Lipid Assessment in Diabetes (GLAD) trial,\textsuperscript{115} the phase II study previously used to develop the model by Hamrén et al.;\textsuperscript{99} and the phase III GALLANT6, which had a longer treatment duration than SIR and GLAD. GLAD and GALLANT6 included both drug-naïve and previously treated patients. The total number of subjects was 1460. The observations included tesaglitazar plasma concentrations (C), FPG, and FI.

\[ 3.3. \quad \text{Development of the Lasso for Covariate Selection within NONMEM (Paper II)} \]

An automatic procedure for using the lasso in NONMEM was developed and tested. This procedure transformed the basic model file (i.e. the control stream for the NONMEM software) into a lasso model file, and subsequently ran the lasso. Estimating the lasso model is the same as estimating the full-covariate model, including all investigated covariate relations in the model, however with a restriction on the absolute sum of the covariate coefficients according to

\[
\sum_{p=1}^{P_{\text{cov}}} \sum_{u=1}^{X_{p,u}} | \theta_{p,u} | \leq t
\]

(18)

where the model size is determined by the value of \( t \). The value of \( t \) is estimated by cross-validation on the OFV. One example of the result of a cross-validation is presented in Figure 6. Further, an illustration of how the covariate coefficients grew with \( t \) is found in Figure 7. However, in the actual lasso procedure, a single model fit on the original dataset using the appropriate \( t \) was sufficient after cross-validation had been performed. Figure 7 is provided merely to illustrate the lasso shrinkage and the results within are not needed for running the lasso. The shrinkage applied via the lasso restriction resulted in covariate selection and simultaneous estimation of all parameters in a single NONMEM model fit.
Figure 6. Illustration of how the cross-validation objective function values change with $t$ in a five-fold cross-validation. All OFV values are given as the difference from the base model ($t=0$). The predictions on the five external test datasets are represented by the solid lines with symbols 1-5 (scale on left y-axis). The sum (pOFV) of the prediction on the five test datasets is given by the broken line (scale on right y-axis). Variability among the five sets is high, just as in validation using a small validation dataset obtained by e.g. data splitting. However, the sum clearly indicates a minimum at $0 < t < 0.15$. The optimal $t$ can be investigated in more detail for this range (see Figure 7) but, normally, a covariate model of this small size would not be considered as clinically relevant.

3.4. Procedure for Analysis of nlme Models (Papers I-III)

In papers I and III, the models used for estimation were the same as those used in the simulations that generated the data, with the exception of the covariate model. The linear covariate relations that were eligible for investigation included the ones used in the simulation, but false covariate relations were also investigated. In paper I, the first step of SCM was run with several p-values. In papers II and III, forward inclusion of covariate relations at $p < 0.05$ was followed by backward elimination. In paper II, three different p-values were investigated for the backward elimination whereas in paper III, $p < 0.01$ was used in this procedure.

In paper III, five different approaches to incorporating information or knowledge from the first study into the analysis of the second were investigated. The data was viewed as information and analysis of this information provided qualitative and quan-
Figure 7. Illustration of the covariate model using a finer grid in the relevant range of \( t \)-values. The predictive objective function value (pOFV) is represented by the broken line and the scale on right y-axis. It has its minimum at \( t=0.06 \). At this \( t \), two covariate relations are selected by the lasso: \( CLCAGE \) and \( VHCTZ0 \). The former denotes the effect on CL of age above 60 (otherwise zero, as defined in the original publication) and the latter the effect on V of having no concomitant treatment with hydrochlorothiazide; the typical subject was on this treatment. Covariate relations that are not included in the lasso model for any of the investigated \( t \)-values are not plotted in the graph, but all the 20 investigated covariate relations are presented in the legend (to the right).

In paper II, SCM and the lasso were compared on datasets of different sizes, i.e. including different numbers of subjects. The comparisons were made according to different scenarios, including two different base models and either only investigating linear models or additionally investigating piece-wise linear models, cf. Equation (13). The base models were both parsimonious models identified on the large original dataset whereas the stochastic model was slightly simplified for better estimability on small datasets. Reasonable simplifications were identified by the use of bootstrapping with a small sample size. The two basic models were identical except that one was a basic model without any covariates and the other was an allometric model based on body weight and calculated fat-free mass.
3.5. Development of a Model for FPG and FI (Paper IV)

Topp et al. proposed a model for BCM, (fasting) insulin and glucose dynamics in normal subjects. The structure of this mechanistic model was used as the starting point for the current analysis, and components for treatment and disease progression in different groups of patients were added. Topp et al. presented mean values for the parameters for healthy individuals. However, it is likely that several of these will be changed in patients with diabetes. In the available data, five different disease groups (DGR) were defined as subjects in SIR (DGR 1), naïve and pre-treated GLAD patients (DGR 2 and 4, respectively) and naïve and pre-treated GALLANT6 patients (DGR 3 and 5, respectively). These DGR can at least partly be ordered into different disease stages as described in Figure 8. Further, a population approach was applied to the modelling to account for the heterogeneity between patients within the same DGR.

Figure 8. Illustration of the disease stage in healthy volunteers and the five different disease groups (DGR). The typical drug-naïve patients in the SIR, GLAD and GALLANT6 studies can be assumed to have an increasing degree of diabetes. However, it is not obvious how to rank the pre-treated patients in relation to the drug naïve in the same study.

Topp et al. described the net changes for FPG, FI and BCM in three differential equations. All parameters in these equations cannot be identified from the data available in this work. Therefore, disease progression was described as a change from the healthy state in two parameters which were estimated separately for each of the five DGR. The first parameter was insulin sensitivity ($S$). Reduced insulin sensitivity alone does not cause diabetes, since beta-cell adaptation acts as a negative feedback to eventually bring the FPG back to the set-point. The second parameter ($OFFSET$) offset beta-cell adaptation so that this strove towards a higher set-point FPG. This effectively reduced
the BCM as the OFFSET increased. The rate of beta-cell adaptation at different FPG levels is illustrated for a healthy and an offset (i.e. type 2 diabetic) individual in Figure 9.

Figure 9. Beta-cell adaptation rate versus FPG. The grey curve represents the change in a healthy individual and the black in an offset (diabetic) individual. At higher FPG values the two curves join at a high (positive) adaptation rate. The dotted vertical lines mark the physiological fixed points in each of the two individuals. This is a point of attraction and the beta-cell adaptation acts with a negative feedback to bring the FPG back to this point.

The initial modelling attempts were not successful in describing the treatment effects because a higher decline in FI than expected from the decline in FPG was seen. A relation between insulin elimination and insulin sensitivity \(^95, 96\) was therefore incorporated into the model. Treatment effects were incorporated directly on OFFSET where the effect delay was accounted for by beta-cell adaptation but as an indirect effect on insulin sensitivity. Differences in the degree of disease for different DGR were included in the model. However, since none of the studies lasted longer than six months, no attempt was made to identify individual disease progression during the study period. The adaptation of BCM is seen as a symptomatic effect of treatment, much like the effect on insulin sensitivity. These should be separated from any underlying disease progression in OFFSET and S. Finally, the fact that only fasting observations of glucose and insulin were available allowed the assumption that these variables were at steady state with respect to each other, given the current level of BCM and insulin sensitivity. The interdependencies between the different variables in the model are illustrated in Figure 10.
Figure 10. Schematic illustration of the FPG-FI model. Changes from the steady state are illustrated as indirect effects for all four variables, indicated by solid or broken arrows. However, responses in FPG and FI are relatively fast and therefore assumed at steady state relative to one another, at the given level of S and BCM, indicated by the broken arrows. Drug treatment exerts an indirect effect on S and BCM which explains the delay in the response on FI and FPG. The indirect effect of drug treatment and the effect of S on the elimination of FI are additions to the model as originally suggested by Topp et al.

3.6. Adaptive Learning (Paper III)

In the scenario where the optimal design was used, the second study was performed with an ED optimal design. This was based on the knowledge gained from analysing the first study and on the knowledge on the distribution of the covariate values in the patient population. The number of observations per subject in the second study was substantially lower than in the default scenario but the number of subjects was the same. The study was optimized on the observation times as well as on which individuals to include (stratification).

Special subpopulations are often investigated to rule out the possibility that a covariate is clinically relevant or to confirm the findings of a previous study. However, here design optimization was performed using the full-forward covariate model obtained from analysing the first study. Therefore, only the subpopulations representing the covariates which were in the first model were eligible for stratification. An error model assuming proportional error distribution was used, cf. Equation
Covariate Model Building in NONMEM

(5) which also assumes the same proportional error magnitude during the absorption and elimination phases. Because of this, the sample times were restricted to the same interval as for the non-optimized designs (0.2-4 typical half-lives of the drug).

To account for uncertainty in both the individual covariate values and the parameter estimates, a modified ED optimization was performed. It was modified in the sense that both the model parameters and the model covariates were assumed to have a distribution associated with them.

3.7. Calculation of Power and Inclusion Rates (Papers I and III)

The inclusion rate was calculated for each scenario and for each covariate relation as the percentage of the replicates from which the covariate was selected. Since the data were generated by simulation from a model in papers I and III, the true covariates and those that were only correlated with (or unrelated to) the parameter were known. Because of this, we used the term power in paper I. Power is the percentage of replicate datasets from which all true covariates were selected. In this paper there was only one true covariate.

3.8. Evaluation of Selection Bias (Papers I and III)

Bias denotes the presence of a systematic error causing the expected value (averaged over many repeats) to deviate from the true value. The (conditional) selection bias can only be calculated from the replicates where all true covariates were selected, because it is otherwise confounded with omission bias. This has two implications. The first is that selection bias is best calculated from repeated analyses of simulated data, as it is otherwise not known which covariates are true. The second is that the calculation of selection bias may require analysis of a large number of replicates if the power of finding all true covariate relations is low. This was the reason for the high number of replicates used in papers I and III. Selection bias was calculated on a relative scale according to

$$\text{Selection bias}_{P\text{cov}} = \frac{\sum_{n=1}^{N_{rep}} (s_n \cdot \hat{\theta}_{n\text{cov}} - \theta_{P\text{cov}})}{\sum_{n=1}^{N_{rep}} s_n} \cdot 100\%$$

(19)

where $s_n$ equals one if all true covariates were selected and zero otherwise, $\hat{\theta}_{n\text{cov}}$ is the estimated covariate coefficient for parameter $P$ and covariate $\text{cov}$ from dataset replicate $n$, and $\theta_{P\text{cov}}$ is the covariate coefficient $\text{cov}$ without selection bias. In paper III, $\theta_{n\text{cov}}$ was calculated as the average estimate over all replicates when estimating the true covariate model. In paper I, the value used in simulation was used. There is no difference between these two approaches unless the estimation method provides biased estimates (estimation bias).
3.9. Evaluation of Predictive Performance (Papers I-III)

The predictive performance was used to summarize how well a model can predict new data or some aspect of it. In paper I, the model error of the true covariate parameter was calculated for the replicates where the covariate was statistically significant; this is the same as being selected if we let no other covariates compete for selection. This approach was taken since the initial investigation showed that selection bias was not much inflated due to competition. The model error was calculated on a relative scale as the mean absolute error \( (mae) \) according to

\[
mae_{Pconv} = \frac{\sum_{n=1}^{N_{repl}} s_n \left( \hat{\theta}_{Pconv} - \hat{\theta}_{Pconv} \right)}{N_{nsubj} \sum_{n=1}^{N_{repl}} s_n}
\]  

(20)

where \( N_{repl} \) equals 7,400 and the remaining notation is as explained in subsection 3.8 above.

In paper II, the prediction error was evaluated on the observations of the dependent variable \( (y) \) in the external validation datasets according to:

\[
mae_{DV} = \frac{1}{N_{repl}} \sum_{v=1}^{N_{rep}} \left( \frac{1}{N_{tot,v}} \sum_{i=1}^{N_{tot,v}} \sum_{j} \left( |y_{ijv} - \hat{y}_{ijv}| / y_{ijv} \right) \right) \times 100\%
\]  

(21)

Where for validation dataset \( v \), \( N_{tot,v} \) and \( N_{subj,v} \) are the total number of observations and the number of subjects, respectively. For subject \( i \) in validation dataset \( v \), \( N_{obs,iv} \) is the number of observations and \( y_{ijv} \) and \( \hat{y}_{ijv} \) are the \( j \)th observation and prediction, respectively. The number of replicates \( (N_{repl}) \) for each of 16 scenarios in this paper was 100.

In paper III, the model error was calculated on the estimated typical value of clearance as the root mean squared error

\[
rmse_{TVCL} = \frac{100\%}{N_{repl}} \sum_{n=1}^{N_{repl}} \left( \frac{1}{N_{subj}} \sum_{i=1}^{N_{subj}} \left( \frac{TVCL_{in} - \overline{TVCL}_{in}}{\min [TVCL_{in}, \overline{TVCL}_{in}]} \right)^2 \right)
\]  

(22)

where \( TVCL_{in} \) and \( \overline{TVCL}_{in} \) are the true and predicted TVCL values for individual \( i \) in replicate \( n \), respectively. The min function in the denominator returns the smallest of \( TVCL_{in} \) and \( \overline{TVCL}_{in} \).

3.10. Evaluation of Design Optimization (Paper III)

In addition to bias and predictive performance, the performance of the optimized and sparse design was compared with that of the default design by assessing the distributions of the relative estimation error according to
\[ \text{REE}_{nk} = \frac{\hat{\theta}_{nk} - \theta_k}{\theta_k} \times 100\% \]  

(23)

where \( \theta_k \) is the true parameter value and \( \hat{\theta}_{nk} \) and \( \text{REE}_{nk} \) are the estimate and estimation error for parameter \( k \) in replicate \( n \), respectively. The distribution of the relative estimation error was used to assess the estimation bias and precision of all parameters when the true model was fitted to the second study.

Further, each optimal design calculation provided not only a design but also an estimate of the lower boundary of the parameter precision that would result if the optimized design would be realized and used as the only information source. Although not in the primary scope of paper III, these estimates were also evaluated. The estimate of the coefficient of variation for each parameter \( (\text{CV}_{nk}) \) was compared to \( \text{REE}_{nk} \) when the same model used to optimize the design was re-estimated using the second study with the realized optimal design. If \( \text{CV}_{nk} \) were to be unbiased, the ratio of \( \text{REE}_{nk} \) and \( \text{CV}_{nk} \) would be distributed according to the standard normal distribution, given that parameter estimates are normally distributed around the true value.

### 3.11. Evaluation of the Model for FPG and FI (Paper IV)

In paper IV, the capability of the model to describe the biomarkers without any (clinically) relevant misfit was evaluated graphically. Thus, the model was evaluated by a visual predictive check (VPC) that included both the median profile over time and the distribution of observations around the median. The 95% prediction interval (95%-PI; i.e. the 95%-confidence interval predicted by the model) was included in the graphs. The results were displayed separated for different subgroups, i.e. DSG and dose groups. In the VPC, simulations from the model were compared with the distribution that was actually observed at different points in time and in different subgroups. This was an internal evaluation. As an external evaluation of the model, the model predicted BCM in the different DSG and the relations between insulin sensitivity and insulin elimination were compared with reports from the literature.
4. Results

4.1. Power and Inclusion Rates (Papers I and III)

Paper I describes the power, selection bias, and predictive performance in covariate modelling. Figure 11 displays the power of selecting the true covariate and the type-I error rate when covariate 1 was tested without competition from the other false covariates. The default simulations cover a wide range, from low to high power, and the type-I error rate is only slightly higher than the nominal value. Figure 12 depicts the relative change in the power of finding the true covariate when one of the false covariates was allowed to compete for selection. The selection was more resistant to correlation between the covariates when the dataset was large or when the true covariate coefficient was large, i.e. when a strong covariate effect was present in the investigated population.

![Figure 11. The power of selecting the true covariate without competition is shown for different dataset sizes. Selection was made if \( p < 0.05 \) on datasets that were simulated with a covariate coefficient exerting a strong (S), moderate (M) or weak (W) influence on CL. The dotted horizontal line is drawn at the nominal significance level, 0.05. The actual significance level represented by simulation from the null hypothesis (N) is shown for small datasets (20, 50 and 100 subjects).](image-url)
Figure 12. The relative power of selecting the true covariate vs. correlation with the competing covariate. The relative power is the ratio of power values with and without competition from a false covariate. Results are from one of many simulation set-ups and given for different dataset sizes and a weak (left column), moderate (middle column) and strong (right column) covariate. The relative power is given for $p < 0.05, 0.01$ and $0.001$ in the top, middle and bottom rows, respectively. As expected, weak covariates and small datasets made the selection less resistant to correlation (left column compared to right column and plotting character 1 compared to plotting character 5). Also, a stricter p-value made the selection more resistant to correlation (top row compared to bottom row).

Paper III describes knowledge propagation in model-based analyses. The inclusion rates of all six investigated covariates are shown in Figure 13 for the different simulation scenarios and knowledge propagation approaches. The false covariates are easily identified since their inclusion rates are much lower. The different covariates in this comparison can be said to compete for selection. However, for SEX and WT there is actually a synergy in the selection. These covariates mask the effect of each other and selection of one encourages selection of the other. Not surprisingly, pooling all available data is superior for identifying the true covariates. However, this may also encourage inclusion of false (but correlated) covariates to a higher extent. From a qualitative perspective, the merge approach to knowledge propagation is more inclusive than other approaches.
Figure 13. Inclusion rate for the different approaches to knowledge propagation, given simulation set-up and covariate. Both pre-specified approaches (PS & PSP) result in the same qualitative estimation of knowledge and are therefore not separated in this figure. The actual approaches are symbolized by white bars. For the Model Merge (MM) approach, if a covariate was selected from both the first and the second study, this is symbolised by an overlaid black bar. Analysing the pooled dataset (DP) provides the highest inclusion rate of the true covariates. The MM approach is more inclusive than the approaches selecting the model from either the first or the second study (NI, PS, and PSP).

4.2. Selection Bias (Papers I and III)

For paper I, Figure 14 shows the selection bias in the true covariate coefficient when the true covariate was investigated alone or in competition with the four false covariates. The selection bias was very high (>100%) when a weak covariate relation was selected from a small dataset (≤50 individuals). Surprisingly, however, competition of false covariates added little to the selection bias when statistical significance was required. This can be seen in that the selection bias in the left panel column was only marginally higher than that in the right column. Focusing only on the selection bias without competition and using the results from all available scenarios and covariates, Figure 15 shows that without competition the selection bias was a function of the power. If the power was less than 20% when using a p-value of 0.05 or stricter, the bias was more than 100%. This means that the identified covariate relation can be expected to worsen (external) predictions from the model. Further, evaluating the importance of the biased covariate coefficient may falsely lead to the conclusion that
this is a clinically relevant relation. Attempts to assess the power from the biased estimate would result in an overestimation.

Figure 14. Selection bias in a true covariate coefficient vs. dataset size. Results were derived from one of many simulation set-ups and given for a weak (W), moderate (M) and strong (S) covariate. The bias is given for $p < 0.05$, 0.01 and 0.001 in the top, middle and bottom rows, respectively. In the left column, the true covariate model has been selected in competition with the four false covariates whereas, in the right column, no competing covariate has been allowed to affect selection. As can be seen, the selection bias is severe for weak covariates selected from small datasets. The selection bias caused by requiring statistical significance is not increased much by competition from the four false covariates (left column compared to right). Also, as can be expected, the requirement for a stricter $p$-value increases the selection bias (top row compared to bottom row).
Figure 15. Bias in the estimated covariate coefficient vs. power. Results are derived from different scenarios (plotting characters A-H) given $p < 0.05$, 0.01 and 0.001 in the top, middle and bottom rows, respectively.

In paper III, where bias was only investigated in the presence of competition, two more aspects of selection bias can be seen (Figure 16). The first is that unevenly distributed covariates (e.g. ATAR) require larger datasets. This is because such covariates are less informative. The other aspect is that the masking covariates (SEX and WT), which often end up in the model together, inflate the selection biases of each other. This is due to correlations of estimate, or colinearity. If the correlation is not preserved, e.g. in another patient population, predictions from this model would perform particularly poorly. Further, the two covariates may falsely be regarded as clinically relevant due to the bias. The two may seem clinically relevant if assessing each covariate separately even though the contributions from the two covariates in combination are modest in most individuals. Regarding the performance of the different approaches to knowledge propagation, the naïve-independent approach resulted in very high selection bias. The pre-specified approach was unbiased, the merge approach was almost so, whereas the remaining two approaches were in-between.
Figure 16. Selection bias in the estimated covariate coefficient for the “true” covariates in the different simulation set-ups. To indicate uncertainty, an 80% confidence interval has been drawn as a horizontal line wherever the line expands outside the dot marking the point estimate of bias. As was expected, the Pre-Specified (PS) approach had no selection bias. The Data-Pool (DP) approach shows a moderate selection bias in most set-ups. The Naïve-Independent (NI) approach, however, shows severe selection bias in the estimated effects of ATAR, WT and in all four covariates with the Small Second Study. The Pre-Specified-Data Pool (PSP) had a selection bias which was between those of the NI and the PS approaches. This is seen from the two simulation set-ups where this approach has been evaluated, in the two lowest panels. The Model-Merge (MM) approach had low selection bias.

4.3. Predictive Performance (Papers I-III)

For paper I, Figure 17 shows that the predictive performance is a function of power without competition. Covariate relations selected at $p < 0.05$ with a power $< 25\%$ reduce the predictive performance compared to the null hypothesis of no covariate effect.
Figure 17. Graph showing log mae of the estimated covariate coefficient vs. power. Results were derived from different simulation scenarios (plotting characters A-H), given p < 0.05, 0.01 and 0.001 in the top, middle and bottom rows, respectively. The dotted vertical and horizontal lines intersect at the point where the selected covariate model is expected to have no influence on the predictive performance. This point was estimated using the Splus-super smoother. The lower right quadrant represents the high powered covariates which are expected to improve the predictive performance. The upper left quadrant represents low-powered covariates expected to worsen the predictive performance.

For paper III, Figure 18 displays the predictive performance for the different approaches to knowledge propagation and for different scenarios. The pooled analysis performed the best in all scenarios. The naïve-independent analysis of the second dataset alone was the worst. When the second dataset was small (50 individuals), this approach performed worse than the base model. Fitting a pre-specified model to the pooled dataset was better than the unbiased alternative of only fitting the second study. The optimal design performed about as well as the empirical less sparse design. These are further compared to the default scenario in subsection 4.5.
Figure 18. Root mean squared error (rmse) in the estimated typical values of clearance provided for the different approaches and different simulation set-ups. As references, two models without data-driven model selection are included: the True and the Basic covariate models fitted to the second dataset are represented by the broken and dotted lines, respectively. Despite the selection bias when using the Data Pool (DP) (cf. Figure 16), this approach is superior not only to the Naïve-Independent (NI) approach but also to the Pre-Specified approaches (PS and PSP). The PS performed better than the NI in all set-ups and especially for the Small Second Study set-ups where better prior information on which covariates are important were available from the first study. Where investigated, the Pre-Specified Data Pool (PSP) predicted better than the PS. This could be expected since the pre-specified model selection was not excessively data-driven. The x-axis is cut on the left side at the lowest possible rmse in any of the simulation set-ups (obtained by fitting the true covariate model to the pooled dataset).

For paper II, Figure 19 displays the prediction error for SCM with three different p-values. For predictive-covariate modelling, the optimal p-value for selection was not primarily dependent on the number of covariate relations investigated. Rather it depended on the information content in the data on the investigated relations, i.e. the signal-to-noise ratio. Since, for these data, the allometric model explained much of the variability that could be explained by the collected covariates, the investigated covariates mainly contributed with noise. Any covariate investigation starting from the allometric base model was best performed with a strict criterion, i.e. a low p-value. Further, the results from the standard base model showed that dataset size (influencing the signal-to-noise ratio) had an influence on the optimal p-value. In the small datasets with 40-60 subjects the signal could not be separated from the noise and the strictest p-value performed best. However, in the larger dataset with 120-180 subjects...
the strictest p-value performed worst.

Figure 19. The mean absolute prediction error (mae) for SCM with three different p-values, investigated in 16 scenarios. Each scenario was a combination of four different analysis dataset sizes (x-axis), two different starting models (allometric or standard) and two different sets of investigated covariate relations (only linear or additionally piece-wise linear). The optimal p-value resulting in the best predictive performance (i.e. low mae), was dependent on the number of investigated covariate relations, but more importantly on the signal-to-noise ratio. Since the allometric model (top panel row) accounted for most of the variability that can be explained by the covariates, a conservative (low) p-value performed best for these scenarios. However, with the standard starting model (bottom panel row) the optimal p-value was dependent on the dataset size. For the very small datasets (40 subjects), the strictest p-value performed the best whereas for larger datasets (120-180 subjects) this p-value performed the worst. For comparison to the lasso later in this work, SCM was represented by the model that used the optimal p-value for each scenario, i.e. different p-values for different scenarios.

In Figure 20, SCM is compared with the lasso and the base model. In this graph, SCM is represented by the p-value that performed best in each scenario. Still, SCM performed worse than the base model in the small datasets. In all 16 scenarios, the lasso performed better than the base model and SCM with any of the three p-values. However, the differences between the lasso and SCM were very small for the large datasets when the optimal p-value was used.
The mean absolute prediction errors (mae) for the base model, the lasso, and the optimal SCM model (SCMoptimal) were compared in the 16 scenarios. “SCMoptimal” denotes SCM with the p-value that performed best in each scenario. Even so, for small datasets with 40-60 subjects SCM performed worse than the base model, i.e. no covariate investigation at all. The lasso, on the other hand, performed best in all 16 scenarios: better than the base model and SCM with any of the three p-values. However, the differences were small when the allometric model was used (top panel row). This was because the investigated covariates did not explain much variability that was not already accounted for by the base model. Further, also because the optimal P-value was used for SCM. Also, for large datasets with the standard base model as starting point for the covariate analysis (bottom panel row) the difference was small between SCM and the lasso.

Figure 21 shows that lasso cross-validation provided a nearly unbiased estimate of the prediction error. SCM, which used all analysis data to develop the model, provided only an apparent prediction error which was overly optimistic, i.e. biased low.
Figure 21. Box plot displaying the bias and precision of the estimated predictive performance of the final covariate model produced by optimal SCM and the lasso in 16 scenarios. For each of the 100 validation datasets in a scenario, the mean absolute prediction error (actual mae) was calculated and compared with the value estimated on the analysis dataset (estimated mae). For the lasso, the ratio of the estimated and actual mae was centred around one, indicating no bias. The estimate from SCM, however, was overly optimistic, i.e. biased low. This was because SCM used the complete analysis dataset to select the covariate model and estimate model parameters. Thus, only an apparent prediction error within the same dataset could be obtained. The spread of the ratio was similar for the two methods, indicated by the similar box length representing the inter-quartile range. To keep the scale in an appropriate range, outliers (outside the whiskers) are not displayed in the graph.

4.4. Run-time Comparison of the Lasso and SCM (Paper II)

For the scenarios with linear models and the standard base model, the computer run-time for SCM was compared with that for the lasso. The two methods were run, one at a time, on a single processor. For each analysis dataset, the ratio of the run-time for SCM to that of the lasso was calculated. For these data, the lasso was often 2-3 times faster for small datasets. However, for the larger datasets, the lasso run-time was only marginally faster (results not shown).
4.5. Design Optimization (Paper III)

When analysing the realized optimal design the PopED estimate of parameter uncertainty seems to be very accurate (unbiased) for most parameters, as seen in Figure 22. As seen in Figure 23, the ratio of REE and CV followed the standard normal distribution (indicated by the grey broken line) for six of the eight parameters displayed. The REE used in this figure was calculated for the model that was used to optimize the design in each replicate and consequently a covariate parameter could only be evaluated if present in that model (which occurred in 23-59% of the replicates for the four true covariates). The two covariates with the lowest inclusion rates in Figure 13 also had the highest selection bias in Figure 16 (WT and ATAR). They generally exhibited larger relative estimation errors than those estimated by PopED. This could be expected since the standard error of these parameters must be biased low\(^{15}\) while the magnitude of the estimated coefficients was biased high, both as a result of model selection.

![Relative estimation error (REE) when estimating the true model using the second study in the Default simulation set-up and the Optimal Sparse Design set-up. Due to the sparseness of the design in the latter, some structural and variance parameters were estimated with slightly lower precision: SIGMA (i.e. \(\sigma^2\)), IIV-ka, IIV-V, ka, and V. For the covariate parameters on the other hand, the sparseness was compensated for by the stratification. Therefore the parameter precision was unchanged or even improved e.g. as for ATAR, the covariate which had the most uneven distribution without stratification. For both the Default and Optimal Sparse Designs there was a downward bias in IIV-CL. Further, in the Optimal Sparse Designs, IIV-ka was estimated with an upward bias at the expense of SIGMA which was estimated with a downward bias.](image)
Figure 23. Quantile-quantile plot of the distribution of the z-score (the ratio of REE and PopED-estimated coefficient of variation) for the 1 000 estimates in the Optimal Sparse Design set-up. This ratio followed the standard normal distribution for most estimates (grey broken line). The covariate parameters for ATAR and WT showed a wider distribution of REEs than expected from the PopED estimates. This was probably due to the data-driven model selection causing both selection bias and seemingly too precise estimates of selected covariate effects. The parameters CL and IIV-CL were excluded from this figure since calculation of REE was not straight-forward for these: the references for these two parameters changed with the covariate model that was estimated in each replicate.

4.6. Diabetes Model (Paper IV)

From paper IV, the VPC on the FPG data is presented in Figure 24, stratified on both DGR and dose. The model described both the median and the 95% confidence interval well. Any major discrepancies between the model-predicted and observed medians were probably due to too few observations; e.g. the 0.1 mg dose group in DGR 2, where the treatment effect appeared to be higher than that in the individuals in the same DGR treated with 0.5 or 1 mg. No clear trend was seen in the 95%-PI for dose groups or studies where the coverage was not appropriate. The VPC on the FI data is presented in Figure 25, stratified on both DGR and dose. The model described both the median and the 95% confidence intervals very well.
Figure 24. Visual predictive check of FPG on the new model developed on all three studies. The blue lines represent model predicted median and 95%-prediction interval (PI). The black circles represent observations within the PI and the red numbers display the outlying ID numbers. The black line is a linear interpolation of all observations. Observations in DGR 5 are jittered in the horizontal direction for visibility since this group includes almost half of the total number of observations. The model describes the observed median and 95%-confidence interval well.
The model prediction for each of the five DGR is given in Figure 26 for a hypothetical long term study with 0.5 mg tesaglitazar which is an intermediate dose size. This figure not only displays the directly observed FPG and FI but also shows the unobserved variables BCM and insulin sensitivity. Since the model did not include any disease progression, the BCM reached a new steady state after about six months of the
same treatment (slightly dose-dependent). The BCM adapted in the opposite direction in DGR 1, due to the high improvement in $S$ that reduced the FPG more than the treatment reduced $OFFSET$. The response in $S$ was fast, with a half-life of about 11 days, i.e. steady state was reached within approximately five weeks. The response in FI was fast but with a small rebound in DGR 2-5 due to beta-cell adaptation. FPG exhibited a rather fast initial improvement but reached steady state only after about six months as a consequence of both improved $S$ and increased BCM.

Figure 26. The changes in BCM, S, FI and FPG simulated by the new model using a fictitious study design including all five DGR. The median response to the typical exposure from 0.5-mg daily dosing of tesaglitazar is displayed for each DGR. DGR 1 has a normal BCM whereas the diabetic patients have only 40-60 percent of this at the start of the run-in. Treatment increases the BCM in the diseased but not in DGR 1, where the increased $S$ even forces the adaptation in the other direction. Shortly before the treatment is commenced the underlying disease is seen, due to the long duration of the run-in period. If the patients were to be left untreated in this manner, DGR 3 would have the lowest $S$ whereas DGR 4 would have the lowest BCM. DGR 2 and 5 are almost identical in the disease.
5. Discussion

5.1. SCM

All the approaches for identifying a covariate model based on an investigation outside NONMEM are associated with several problems on top of those discussed for SCM. SCM is the dominating procedure for identifying the covariate model within NONMEM. Therefore, it is important to understand its properties and investigate alternative methods where these properties are not appealing.

As the correlation between two competing covariates increases, substitution with the false covariate will become more frequent and the relative power will decrease. Strong covariates and large datasets not only increase the power but also increase the resistance to this substitution. Surprisingly, however, the results shown in Figure 14 show that, even at moderate to high correlations, competition bias constitutes a very small proportion of the selection bias. However, this is with the exception of covariates masking each other’s effects so that both end up in the model, as seen in Figure 16. When selecting only the best of several covariates, without requirement for statistical significance, most of the selection bias still remains, but now only consists of the competition bias (results not shown). This is one of the reasons for the attention that this phenomenon has received.

When two potential covariates are extremely correlated ($r \geq 0.95$), it is not uncommon that only one of them is tested for selection on a model parameter. If no prior preferences are available, one of the two covariates has to be discarded at random. However, given that statistical significance is required, it seems plausible that selection bias will never be severely inflated by two competing covariates. Thus, for this case, investigating both and letting the data decide which one to retain would not be inappropriate.

Comparing SCM to the lasso highlighted some interesting properties of the SCM-stopping criterion - the p-value. Whereas inclusion of parameters would become stricter for SBC with increasing dataset size, the AIC and p-value are both neutral in this respect. This means that they require a fixed increase in the log-likelihood, independent of the dataset size. However, if predictive modelling is the primary purpose and model parsimony is not an issue, the strictness of the inclusion criterion should be adjusted in the opposite direction to that in SBC. For example, the p-value required for selection should be less strict (higher) if the signal-to-noise ratio is higher due to a large dataset or because the investigated covariates are highly plausible according to prior knowledge. In the same spirit, testing many covariates instead of few should only require a stricter p-value if this means that less plausible covariates are tested among these (cf. Figure 19).
5.2. The Lasso

Within this thesis, the lasso method was successfully implemented for covariate identification in NONMEM. In 16 different scenarios, this implementation was superior to SCM with respect to predictive performance and computer run-time. Further, the lasso correctly validated the final covariate model by providing a nearly unbiased estimate of predictive performance.

The differences between the two methods were small in large datasets and also in small datasets when the investigated covariates did not explain much of the variability between covariates, i.e. when starting with the allometric model, given that the optimal p-value was used for SCM (cf. Figure 20). However, a small average difference can make a large difference for small subgroups of extreme individuals in a couple of the covariate analyses. The lasso would be superior to SCM even in large datasets when investigating covariates with small subgroups, e.g. rare categories with fewer than approximately 25 subjects or continuous covariates where most subjects are homogeneous with respect to the covariate. In such situations, the selection bias may be high as indicated from e.g. ATAR in Figure 16. The same can be expected when investigating covariates on structural model parameters for which the data are uninformative.

Although 16 different scenarios were investigated using 100 replicates, the comparison in this work is only a case study in the sense that the data originate from a single drug-development programme. In a scenario where all the investigated covariate relations are either highly influential or not influential at all, SCM with an optimal p-value is expected to perform slightly better than the lasso. However, highly influential covariates are often known in advance and can be included in the base model (pre-specified or included using a very liberal p-value, e.g. p<0.50) before the lasso is run. It is not advisable to investigate covariate relations that are thought to be unimportant in small datasets if the objective is to obtain a predictive model. This applies regardless of which method is used for covariate analysis.

In all 16 scenarios, the lasso estimate of prediction error was nearly unbiased because the lasso uses cross-validation in the right way. Most estimates were within ±15% of the true value even for small datasets (cf. Figure 21). However, the intra-individual error was high for these data, almost as high as the IIV in CL. This restricted the possible spread of estimated and actual prediction errors. Thus, even if cross-validation is better than data splitting, such good precision can not be expected on small datasets in general.

The SCM estimate of prediction error was heavily biased, with almost 75% of the estimates below the true value. This was rather independent of dataset size, probably because more covariate relations were included from the larger datasets. However, the precision improved with dataset size (cf. Figure 21). Using cross-validation instead of a p-value in SCM would be possible. However, performed the right way, five-fold cross-validation increases the computer run-time for SCM by almost six-fold. It is unlikely that SCM with cross-validation would perform better than the optimal SCM used in this comparison. However, the optimal p-value is unknown in a real covariate analysis, making SCM perform even worse in a real comparison. For SCM, data splitting is a quick and unbiased alternative although not suitable for small datasets. This is because at least 20% of the data must be withheld during model selection and since the prediction error estimated from such a small data subset is highly variable (cf. Figure 6). Other methods that are far more computer intensive than cross-valida-
tion also exist.\textsuperscript{90}

In this comparison, the lasso was faster than SCM. However, in a situation where only a small set of sensible predictors are investigated, it can be expected that the lasso requires longer run-times than SCM if run on a single processor. More importantly, in the lasso, all model estimations are independent and can be run in parallel. SCM, on the other hand, has to run the model selection in steps and may only execute NONMEM in parallel within each step. In this manner, using a large cluster or grid, the lasso would always be appreciably faster than SCM.

Even if the purpose of the modelling is to create a simple schedule for individualised dosage of new patients, it is suggested that the lasso is very useful. The model produced by the lasso can be used together with the empirical distribution of the covariates to investigate subpopulations at risk of sub-therapeutic or toxic treatment. Based on these results, a simple dosage schedule can be designed using one or a few covariates. The empirical covariate distribution that is used should contain a large number of subjects from the relevant patient population, but does not necessarily have to be derived from the current drug-development programme.

Given the interaction of the covariate model with the structural and stochastic models,\textsuperscript{49} if other parts of the model are re-evaluated after the covariate analysis, the covariate model may also need re-evaluation. Using the lasso, a fast alternative is to only re-fit the final lasso model conditioned on the lasso cross-validated t-value being the same. However, it may be safer to re-evaluate the covariate model completely, especially if the covariate model changes when using the fast alternative. Even so, using the final lasso model for re-evaluating the structural or stochastic parts of the model to some extent takes this sub-model interaction into account. This is still another advantage of the lasso method.

5.3. Approaches to Knowledge Propagation

In paper III, the effects of different approaches to knowledge propagation were explored. This was done by simulating two studies under different conditions. Clearly, this is a simplified view of reality; a drug-development process includes numerous studies in serial/parallel and with larger differences in design than simulated here. However, for the purpose of illustrating the merits of different knowledge-propagation approaches, the simulation system used in this paper is complex enough to provide useful input.

The full Bayesian methods reside at the heart of formalizing quantitative knowledge propagation.\textsuperscript{124, 125} Nevertheless, in the current thesis work it was decided not to investigate these methods since they are still of a less widespread use in drug development. Further, model selection in the Bayesian setting is often unpractical\textsuperscript{125, 126} and requires specification of the prior probabilities for each model that is investigated.\textsuperscript{127} This is problematic because of the correlations between covariates.

Of the five knowledge-propagation methods investigated in this work, data pooling was found to be superior in terms of finding the correct model and predicting external data (cf. Figure 13 and Figure 18). Whenever possible, this approach should be used for combining information to generate knowledge. Data pooling can be expected to perform similar to the NONMEM-prior functionality as implemented
The Naïve-Independent analysis of the second study performed worst in this simulation study. Therefore, its use is recommended only if previous studies do not contain enough information or possibly if study design or analyst time does not allow handling of possibly confounded study differences.

Knowledge on the model structure (e.g. which covariates are important) can be beneficially propagated by with a pre-specified model rather than the data-driven NI approach. The two pre-specified approaches are notably quicker in that only a single model has to be estimated. As evident from Figure 16, a further benefit with re-estimation on the second dataset is that the estimates are not biased by model selection, since this is performed on the first dataset. The use of population modelling has recently been advocated for confirmatory analyses. Although freely data-driven model selection can not be accepted in this case, a treatment-blinded limited exploratory analysis has been suggested in the literature. Naturally, a pre-specified covariate model can be used even on the drug-specific parameters when performing a model-based confirmatory analysis. However, with regard to the predictive performance (Figure 18), applying the pre-specified model to the pooled dataset will result in better performances than the unbiased alternative, unless the model selection was highly data-driven (i.e. when making too many decisions based on too little information). For this purpose, it can also be advantageous to reduce the pre-specified model by removing coefficients where the sign is opposite to that expected or where the parameter cannot be estimated with the necessary precision, e.g. where CV > 50%.

Merging the models (MM) obtained from analysing the studies separately performed almost as well as DP in this simulation study. However, the approach taken in this work, using weights proportional to the number of subjects, will not work as well if the study designs of the first and second studies had been highly different so that the information content per subject would have been considerably different between the studies. There are several ways to handle this problem as a more appropriate meta-analysis. The MM approach is different from a meta-analysis in that the information on which covariates to select is not fully used, but also in that the uncertain covariate effects (symbolized by the white parts of the MM bars in Figure 13) are shrunk compared to the full information in the data. This reduces the selection bias (Figure 16) and variability of the estimate.

5.4. Adaptive Learning

PopED was successfully employed to optimize sparse sampling times and to stratify subjects. The ability of PopED to provide sample times that were robust to parameter uncertainty may also provide a certain robustness against model misspecification. Where PopED failed to provide unbiased estimates of parameter precision (Figure 23) this was likely the result of model selection on the first study resulting in biased estimates of the parameters and their uncertainty.

With both the default and the Optimal Sparse designs, $\omega_{\alpha}$ was estimated with a slight downward bias (Figure 22). This was expected, since the covariate parameters were estimated on TVCL with imprecision, soaking up some of the actually random variability. The Sparse Design further added bias in $\sigma$ and $\omega_{\alpha}$. This could
have been avoided by excluding $\theta_{10}$ and $\omega_{10}$ in the design optimization and estimating the second study conditioned on these parameters being the same as estimated from the first study,\textsuperscript{133} or by using prior information from the same.\textsuperscript{39} However, this would have violated the way in which knowledge was propagated. Furthermore, since $\theta_{CL}$ was still estimated well (cf. Figure 22), it was not necessary to do so.

### 5.5. The Glucose-Insulin Model

This FPG-FI model described all observations well, although they originated from a wide range of heterogeneous subjects, ranging from insulin-resistant subjects to patients with long term diabetes. The foundation of the model had been previously suggested by Topp et al. This model originates from several different sources and this is the first time it has been fitted as a whole to a population of diabetic patients. A few parameters that were different in this heterogeneous subpopulation were re-estimated. In addition, using a population model approach, within-subgroup variability was included as an IIV on the parameters. The dynamic effects of discontinuing or initiating an anti-diabetic treatment were also included in the model, which necessitated the inclusion of a relation between insulin sensitivity and insulin elimination.\textsuperscript{95, 96} The fact that the model fits all subgroups very well with merely a few parameters estimated on the actual data is a validation of the model structure in itself.

The complexity of the new PK-PD model and the limitations of the available data necessitated the assumption that many parameter values were the same for the patients as for the healthy individuals described by Topp et al. Because of this, the BCM predicted from this model should be interpreted as the relative mass, reflecting the actual mass only if these assumptions are correct. The model identified a strong relation between insulin sensitivity and insulin elimination, which is well in line with what has been reported in the literature.\textsuperscript{95, 134} Further, the relative difference between insulin-resistant subjects (DGR 1) and patients with diabetes (DGR 2-5) is also in line with autopsy data reported in the literature.\textsuperscript{135-137} Nonetheless, including data from short-term provocation studies of glucose-insulin homeostasis and observations of BCM\textsuperscript{138} would be of benefit to the model.

The requirement for heterogeneous data is not a weakness of the model but instead offers potential to incorporate information from a wide range of experiments into one quantitative framework. The interest in this kind of model in drug development programmes is growing with the increasing use of quantitative pharmacology.\textsuperscript{2-7, 54} The benefits of using more mechanistic models include potential for both increased knowledge by incorporating more heterogeneous data and improved predictions when the model is extrapolated to previously unexplored areas. For example, if including disease progression in the parameters over time and fitting the model to studies including one-year’s treatment and observations of BCM, it is likely that the predictions of 10-year outcomes would be better than those predicted by its forerunners.\textsuperscript{98} This model allows separation of the symptomatic and disease modifying effects on beta-cell adaptation. The symptomatic effect arises when treatment is initiated or stopped. The disease modifying effect alters the underlying gradual disease progression.\textsuperscript{98} This approach will gain in importance when non-invasive quantification of BCM becomes readily available.\textsuperscript{138}
6. Conclusions

This thesis has investigated different methods and approaches to covariate modelling. Selection bias and high variability of the estimates are indeed severe problems associated with stepwise covariate modelling and any other method using maximum likelihood estimates of the selected covariate coefficients, e.g. GAM, WAM, GA, and graphical analysis. The selection bias is especially obvious when investigating weak covariate relations in small datasets or subgroups, since the power of identifying these is low. As a result, the final covariate model may reduce the predictive performance, even though it appears to explain a significant amount of the variability. This means that covariate relations may falsely appear to be clinically relevant.

Since the reduced predictive performance is mainly caused by low statistical power, it would be especially advantageous to improve on this aspect. This can be achieved by

1. using better criteria for selection, e.g. cross-validation instead of a p-value. Alternatively, the p-value can be adapted to prior beliefs of the quality and information of hypothesised covariate relations, e.g. a very strict p-value can be used for covariate relations with less potential or if analysing small datasets.

2. increasing the statistical power by increasing the information content in the data by propagating knowledge, e.g. by pooling data from several studies. Further, the studies can be more informatively designed, e.g. using a population-optimal design.

Another approach to increasing the predictive performance of the model is to reduce the problems of selection bias and high variability of estimates directly by

3. correcting for bias or, even better, reducing the variability between the estimates by applying shrinkage estimation, e.g. by using the lasso.

4. using a pre-specified model structure, since this is the only way of completely avoiding selection bias. This approach may even be used in a confirmatory model-based analysis.

The implementation of the lasso within NONMEM is superior to SCM in obtaining a predictive covariate model when analysing a small dataset or when investigating covariates with small subgroups in a large dataset. If run in parallel, the lasso is appreciably faster than SCM. The lasso validate the covariate model using cross-validation techniques. Further, it does not require the user to specify the p-value for selection.

If possible, the best approach to increasing the information content and to propagating knowledge is to pool data from several sources; this helps to establish the correct model components and to obtain a model with good predictive performance. With sufficient prior information from the literature on which covariates are important,
estimating a pre-specified model without selection bias can be a fast, reliable process also useful for model-based confirmatory analysis. Merging the estimates from different studies or estimating the pre-specified model on the pooled dataset often results in a more predictive model. Further, using an ED-optimal design to optimize sample times and to stratify subjects from different subgroups was found to be a successful strategy, allowing sparse sampling and handling of prior parameter uncertainty.

A mechanism-based model was developed for the interaction between tesaglitazar exposure, fasting plasma glucose, fasting insulin, insulin sensitivity and beta-cell mass. The model described the observations of fasting insulin and plasma glucose very well, even though the data originated from a wide range of heterogeneous subjects, ranging from insulin-resistant subjects to patients receiving long-term treatment for diabetes. Changes in beta-cell mass for the different groups of patients were also predicted in line with reports in the literature. The model allows the incorporation of data from short-term provocation experiments and from actual observations of beta-cell mass. Thus, the here developed model has the potential to well predict the outcomes of new studies which often is the ultimate goal with modelling.
7. Acknowledgements

The work presented in this thesis was performed at the Department of Pharmaceutical Biosciences, Division of Pharmacokinetics and Drug Therapy, Faculty of Pharmacy, Uppsala University, Sweden.

I would like to express my sincere gratitude to all who have contributed to this thesis and especially to:


Professor Margareta Hammarlund-Udenaes för att du som chef månar om att alla ska må bra och för att du även engagerar dig i undervisningen.

Medförfattare: Joakim Nyberg och Ola Caster som gjorde sina magisterarbeten för mig och samtidigt bidrog starkt till paper II i denna avhandling, både vad gäller praktiskt genomförande och teoretiska inslag. Dr Andrew Hooker, som var den optimale medförfattaren till paper III och som dessutom kan engelska (trots att han är amerikan). Bengt Hamrén och Dr Maria Svensson för ett mycket trevligt distanssamarbete, med intressanta diskussioner och intensivt mailande. I det här sammanhanget vill jag även nämnna min syster Carolina, som helt oförskräckt plöjt igenom hela avhandlingen och förvandlat mitt ursprungliga skatbo till något mycket bättre. Och
Siv, som också läst och med några effektiva handvändningar plötsligt gjort det hela förståeligt och enkelt. Den enda gången hon inte är trovärdig är när hon säger att hon inte förstår någonting, eller när hon påstår att hon bara ser ut som en ängel, men inte är det på riktigt.

För teknisk support skulle jag vilja tacka Pontus Pihlgren och Dr Lars Lindbom, för att ni hållit ihop våra linuxdatorer i ett kluster och utvecklat PsN. Kjell Berggren och Magnus Jansson för att de lappat ihop mina ständigt kraschande Dell-laptops, och för den grafiska sättningen av avhandlingen.

Alla andra som gör vår avdelning till en sådan fantastiskt kul och lärorik arbetsplats. Tack för alla skojiga resor, likstela lunchklubbar, sena kvart-i-fem-fi kan, eldiga julfester, badmintonmatcher och turer i skogen! Tack även för alla nedkluddade formler och utbytta av idéer.

Jag vill även nämligen min gymnasielärare i kemi och bioteknologi, Krister Åkervall. Mycket tack vare dig började jag på Apotekarprogrammet. Då jag var sommarväkande på Karolinska blev Dr Olof Beck min första närkontakt med forskningen. Under denna period bestämde jag mig för att doktorera.

Tack alla gamla vänner som till största delen inte alls har bidragit till avhandlingen men som jag uppskattar desto mer: Olle, Mats, Annika, Tobias, Mikael, Kakan, Sofi, Matilda, Andreas, Monica, Helena, Karin, Anna E, L och B, Hanna M, Sven och Anette.

Sweet ass, alla paddelvänner vars sällskap varit direkt destruktivt för forskningen, men som bidragit med så mycket annat. Framförallt tänker jag på MrPresident, Pastorn, Måts, PimpDaddy, Lena, John, Pelle, Oljan, Larsken och Tobias.

8. References


Covariate Model Building in NONMEM


121. FDA. Drug Interaction Studies - Study Design, Data Analysis, and Implications for Dosing and Labeling (http://www.fda.gov/cber/gdins/interactstud.htm); 2006.


Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy* 59

Editor: The Dean of the Faculty of Pharmacy

A doctoral dissertation from the Faculty of Pharmacy, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-7923