Model-Based Optimization of Clinical Trial Designs

CAMILLE VONG
Dissertation presented at Uppsala University to be publicly examined in B41, BMC, Husargatan 3, Uppsala, Friday, 21 November 2014 at 13:15 for the degree of Doctor of Philosophy (Faculty of Pharmacy). The examination will be conducted in English. Faculty examiner: PhD Oliver Sander (Novartis AG, Basel, Switzerland).

Abstract

General attrition rates in drug development pipeline have been recognized as a necessity to shift gears towards new methodologies that allow earlier and correct decisions, and the optimal use of all information accrued throughout the process. The quantitative science of pharmacometrics using pharmacokinetic-pharmacodynamic models was identified as one of the strategies core to this renaissance. Coupled with Optimal Design (OD), they constitute together an attractive toolkit to usher more rapidly and successfully new agents to marketing approval.

The general aim of this thesis was to investigate how the use of novel pharmacometric methodologies can improve the design and analysis of clinical trials within drug development. The implementation of a Monte-Carlo Mapped power method permitted to rapidly generate multiple hypotheses and to adequately compute the corresponding sample size within 1% of the time usually necessary in more traditional model-based power assessment. Allowing statistical inference across all data available and the integration of mechanistic interpretation of the models, the performance of this new methodology in proof-of-concept and dose-finding trials highlighted the possibility to reduce drastically the number of healthy volunteers and patients exposed to experimental drugs. This thesis furthermore addressed the benefits of OD in planning trials with bio analytical limits and toxicity constraints, through the development of novel optimality criteria that foremost pinpoint information and safety aspects. The use of these methodologies showed better estimation properties and robustness for the ensuing data analysis and reduced the number of patients exposed to severe toxicity by 7-fold. Finally, predictive tools for maximum tolerated dose selection in Phase I oncology trials were explored for a combination therapy characterized by main dose-limiting hematological toxicity. In this example, Bayesian and model-based approaches provided the incentive to a paradigm change away from the traditional rule-based “3+3” design algorithm.

Throughout this thesis several examples have shown the possibility of streamlining clinical trials with more model-based design and analysis supports. Ultimately, efficient use of the data can elevate the probability of a successful trial and increase paramount ethical conduct.

Keywords: nonlinear mixed-effects models, pharmacometrics, likelihood ratio test, NONMEM, power, sample size, study design, proof-of-concept, dose-finding, population optimal design, LOQ, BQL data, neutropenia, docetaxel, myelosuppression, thrombocytopenia, MTD, Bayesian methods, 3+3 algorithm, dose escalation study

Camille Vong, Department of Pharmaceutical Biosciences, Box 591, Uppsala University, SE-75124 Uppsala, Sweden.

© Camille Vong 2014

ISSN 1651-6192
ISBN 978-91-554-9063-8
urn:nbn:se:uu:diva-233445 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-233445)
You can never cross the ocean until you have the courage to lose sight of the shore.

Christopher Columbus

A Papa et Maman
The background artwork ("Binary Galaxy") by Charlie Brej is Copyright Red Hat, Inc, and is used under the terms of the Creative Commons Attribution-Share Alike 3.0 Unported License:
https://creativecommons.org/licenses/by-sa/3.0/
The original work can be found here:
http://fedoraproject.org/wiki/File:Artwork_F10Themes_Binary_galaxy.svg
The global artwork was inspired by gesecolor.com
List of Papers

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


IV Vong C, Nyberg J, Karlsson MO, Friberg LE, Hooker AC. Optimal Design Applied to Hematological Toxicity-Induced Anticancer Treatment. *In manuscript.*

V Vong C, Fouliard S, Chalret du Rieu Q, Kloos I, Friberg LE, Chenel M. *In Silico* Comparison of Maximum Tolerated Dose Determination in a Phase I Dose-Finding Framework: Application to Hematological Toxicity for a Histone Deacetylase Inhibitor Abexinostat, Co-Administered with Free or Liposomal Doxorubicin in Solid Tumors. *In manuscript.*

Reprints were made with permission from the respective publishers.

*The authors contributed equally to this work.*
## Contents

1. Introduction ......................................................................................... 13  
   1.1. Anatomy of clinical trials .......................................................... 14  
   1.2. Design and analysis of clinical trials .......................................... 16  
      1.2.1. The methodology ............................................................ 17  
      1.2.2. Basic study designs .......................................................... 18  
      1.2.3. Traditional analysis methods and drawbacks ................... 19  
   1.3. Model-based drug development ................................................. 21  
      1.3.1. Nonlinear mixed-effects models ........................................ 23  
      1.3.2. Maximum likelihood estimation ...................................... 24  
   1.4. Model-based trial data analysis .................................................. 26  
      1.4.1. Hypothesis tests and statistical power.............................. 27  
      1.4.2. Handling censored and missing data .................................. 31  
   1.5. Model-based trial design ............................................................ 34  
      1.5.1. Stochastic/Monte-Carlo simulations ................................ 35  
      1.5.2. Cramér-Rao inequality ...................................................... 35  
      1.5.3. Population Fisher information matrix ............................... 36  
      1.5.4. Design optimization ........................................................ 37  
   1.6. Oncology dose-escalation trial design ........................................ 39  
      1.6.1. Chemotherapy treatment .................................................. 39  
      1.6.2. Hematological-induced anticancer treatment .................... 41  
      1.6.3. Optimization of cancer treatment .................................... 42  
      1.6.4. Phase I study designs ....................................................... 44  

2. Aims ..................................................................................................... 49  

3. Methods ............................................................................................... 51  
   3.1. Analyzed data ............................................................................ 51  
      3.1.1. Study drugs ................................................................. 51  
      3.1.2. Study designs and clinical data ......................................... 53  
   3.2. Model implementation ............................................................... 54  
      3.2.1. PKPD benchmark models ................................................. 55  
      3.2.2. Thrombocytopenia model building strategy ..................... 58  
      3.2.3. Data analysis and model evaluation .................................. 59  
   3.3. Power estimation ......................................................................... 61  
      3.3.1. Monte-Carlo based power ................................................ 61  
      3.3.2. Monte-Carlo Mapped power ............................................. 61  
      3.3.3. Proof-of-Concept and Dose Finding sample size .............. 62
3.4. Optimal design methodology .......................................................... 64
   3.4.1. Information-based criteria for LOQ ....................................... 64
   3.4.2. Utility-based criteria for dosing regimen optimization .......... 66
   3.4.3. Information-based criteria for sampling optimization ........... 67
3.5. Dose-escalation algorithms .......................................................... 69
   3.5.1. Pharmacometric-based maximum tolerated dose ............... 69
   3.5.2. Bayesian approaches ......................................................... 70
3.6. Methodology assessment .............................................................. 72
   3.6.1. Evaluation of Monte-Carlo Mapped power ........................ 72
   3.6.2. Comparison of optimized LOQ sampling schedules .......... 73
   3.6.3. Comparison of optimized dosing regimens ....................... 74
   3.6.4. Comparison of optimized sampling schedules ................. 75
   3.6.5. Maximum tolerated dose determination framework .......... 76
4. Results ................................................................................................. 77
   4.1. Statistical hypothesis testing .................................................... 77
      4.1.1. Performance of the MCMP method .................................. 77
      4.1.2. Demonstration of application benefits ............................. 79
   4.2. Anticipation of prospective constraints .................................... 81
      4.2.1. Application to limit of quantification constraints ............ 82
      4.2.2. Application to dosing and sampling optimizations .......... 85
   4.3. Dose-escalation oncology trials .............................................. 88
      4.3.1. PKPD model .................................................................. 88
      4.3.2. Comparison between dose-escalation methods .............. 90
5. Discussion ........................................................................................... 95
   5.1. Statistical hypothesis testing .................................................... 95
   5.2. Anticipation of prospective constraints .................................... 98
   5.3. Dose-escalation oncology trials .............................................. 101
6. Conclusions ....................................................................................... 105
Acknowledgments .................................................................................. 107
References ............................................................................................ 113
**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ac</td>
<td>Acetylation</td>
</tr>
<tr>
<td>(S)AE</td>
<td>(Serious) Adverse Event</td>
</tr>
<tr>
<td>AIDS</td>
<td>Acquired Immunodeficiency Syndrome</td>
</tr>
<tr>
<td>ANC</td>
<td>Absolute Neutrophil Count</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>BLRM</td>
<td>Bayesian Logistic Regression Model</td>
</tr>
<tr>
<td>BQL</td>
<td>Below Quantification Limit</td>
</tr>
<tr>
<td>BSA</td>
<td>Body Surface Area</td>
</tr>
<tr>
<td>CL</td>
<td>Clearance</td>
</tr>
<tr>
<td>CRB</td>
<td>Cramér-Rao Bound</td>
</tr>
<tr>
<td>CRM</td>
<td>Continuous Reassessment Method</td>
</tr>
<tr>
<td>CTCAE</td>
<td>Common Terminology Criteria for Adverse Events</td>
</tr>
<tr>
<td>CTS</td>
<td>Clinical Trial Simulation</td>
</tr>
<tr>
<td>CT-scan</td>
<td>Computed Tomography</td>
</tr>
<tr>
<td>CUI</td>
<td>Clinical Utility Index</td>
</tr>
<tr>
<td>CV</td>
<td>Coefficient of Variation</td>
</tr>
<tr>
<td>DF</td>
<td>Dose-Finding</td>
</tr>
<tr>
<td>DLT</td>
<td>Dose Limiting Toxicity</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>EBE</td>
<td>Empirical Bayes Estimate</td>
</tr>
<tr>
<td>EI</td>
<td>Equivalent Interval</td>
</tr>
<tr>
<td>EORTC</td>
<td>European Organization for Research and Treatment of Cancer</td>
</tr>
<tr>
<td>EPI</td>
<td>Expected Prediction Intervals</td>
</tr>
<tr>
<td>EWOC</td>
<td>Escalation With Overdose Control</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>FN</td>
<td>Febrile Neutropenia</td>
</tr>
<tr>
<td>FO(I)</td>
<td>First-Order (with Interaction)</td>
</tr>
<tr>
<td>FOCE(I)</td>
<td>First-Order Conditional Estimation (with Interaction)</td>
</tr>
<tr>
<td>FPG</td>
<td>Fasting Plasma Glucose</td>
</tr>
<tr>
<td>G-CSF</td>
<td>Granulocyte-Colony-Stimulating Factor</td>
</tr>
<tr>
<td>GEE</td>
<td>Generalized Estimating Equation</td>
</tr>
<tr>
<td>H₀</td>
<td>Null Hypothesis</td>
</tr>
<tr>
<td>H₁</td>
<td>Alternative Hypothesis</td>
</tr>
<tr>
<td>HAT</td>
<td>Histone Acetyltransferase</td>
</tr>
</tbody>
</table>
HbA1c  Glycosylated Hemoglobin  
HDACi  Histone Deacetylase inhibitor  
HIV  Human Immunodeficiency Virus  
IDSA  Infectious Disease Society America  
IIIV  Inter-Individual Variability  
IOV  Inter-Occasion Variability  
IPP  Individual PK Parameters  
IV  Intravenous  
K  Lysine  
LL  Log-Likelihood  
LLOQ  Lower Limit of Quantification  
LOCF  Last Observation Carried Forward  
LOD  Limit of Detection  
MAR  Missing at Random  
MBDD  Model-Based Drug Development  
MCAR  Missing Completely at Random  
MDT  Mean Dose Time  
ML  Maximum Likelihood  
MMR  Moderated Multiple Regression  
MNAR  Missing Not at Random  
MRI  Magnetic Resonance Imaging  
MTD  Maximum Tolerated Dose  
mTPI  modified Toxicity Probability Intervals  
MTT  Mean Transit Time  
NIHSS  National Institutes of Health Stroke Scale  
NLMEM  Nonlinear Mixed-Effects Models  
NME  New Molecule Entity  
OD  Optimal Design  
ODE  Ordinary Differential Equation  
OFV  Objective Function Value  
OS  Overall Survival  
PBPK  Physiological-Based Pharmacokinetic  
PD  Pharmacodynamic  
PET-scan  Positron Emission Tomography  
PK  Pharmacokinetic  
PLD  Pegylated Liposomal Doxorubicin  
POC  Proof-of-Concept  
PPC  Posterior Predictive Check  
PsN  Perl-speaks-NONMEM  
RBC  Red Blood Cell  
RBCT  Randomized Biomaker-Controlled Trial  
RCCT  Randomized Concentration-Controlled Trial  
RDCT  Randomized Dose-Controlled Trial  
REE  Relative Estimation Error  
RP2D  Recommended Phase 2 Dose
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUV</td>
<td>Residual Unexplained Variability</td>
</tr>
<tr>
<td>(R)SE</td>
<td>(Relative) Standard Error</td>
</tr>
<tr>
<td>SSE</td>
<td>Stochastic Simulation and Estimation</td>
</tr>
<tr>
<td>TDM</td>
<td>Therapeutic Drug Monitoring</td>
</tr>
<tr>
<td>TITE-CRM</td>
<td>Time-To-Event-CRM</td>
</tr>
<tr>
<td>ULOQ</td>
<td>Upper Limit of Quantification</td>
</tr>
<tr>
<td>UPM</td>
<td>Unit Probability Mass</td>
</tr>
<tr>
<td>(pc)VPC</td>
<td>(prediction-corrected) Visual Predictive Check</td>
</tr>
<tr>
<td>V</td>
<td>Volume</td>
</tr>
</tbody>
</table>
1. Introduction

Drug development is a risky venture. Releasing a new drug on the market takes on average 10 to 15 years and costs approximately one billion dollars \(^1\,^2\). Despite the long-heralded trend of higher development costs in the future, which makes it challenging for biopharmaceutical companies to recoup their investment, the number of newly approved novel drugs (i.e. new molecular entities; NMEs) keeps decreasing. For instance, the number of drugs in 2010 approved for marketing has declined to an all-time low with only 15 NMEs to receive marketing approval\(^3\). Furthermore, about 85% of the early phase therapies fail and for those that manage to reach phase III, the latest stage before regulatory approval, only half of them will be registered\(^4\). Naturally, such soaring costs, long timelines and high attrition rates may clearly not uphold the sustainability of the system.

The lack of growth in the drug development pipeline is inadvertently a setback for societal demand for more cost effective accessible treatments and for the global disease burden. It affects patient care since less new and affordable drug treatments are available for therapy, and has already skewed research focus on compounds with a large potential instead of the development of cures for high-prevalent third-world diseases\(^5\,^6\). Although repurposing strategy efforts may help reducing the time and money of afresh investments by proposing new applications of existing drugs, the gap remained unabated.

The main reasons for the staggeringly high cost and the scarcity of new innovative medicines are multifactorial: increasing technology and clinical costs, safety issues, lack of efficacy, company’s directives, difficulty in patient recruitment and retention, higher ethical and regulatory requirements, generic competition, reimbursement challenges, etc. From a scientific perspective, many of the methodologies used to predict and evaluate product safety and efficacy are not optimal. As a direct consequence, a substantial number of potential drugs are being lost to a traditional clinical trial paradigm established in the 1960s, which utilizes study designs and data analyses in an uninformative and ineffective manner\(^7\,\,^{10}\). In fact, in an attempt to consistently achieve increasing Food and Drug Administration (FDA) statistical standards for demonstrating safety and efficacy, clinical trial sample sizes have increased at an average rate of 7.47 per cent per annum, from the 1970s to 2001\(^3\).
Oncology clinical development, which has moved quickly at the heart of the pricing debate, exacerbates furthermore this hurdle\textsuperscript{11–13}. The complex nature of cancer treatments not only requires costly medical procedures in the protocol (biopsies and MRIs) at regular intervals to assess endpoints like tumor shrinkage but also extends the total trial duration. The latter is mainly explained by the FDA requirement to have improved overall survival (OS) as the primary endpoint\textsuperscript{14}. Furthermore, inadequate study designs\textsuperscript{15}, poor predictability of current animal models in humans\textsuperscript{16,17}, very limited validated surrogate endpoints\textsuperscript{14}, disease splintered into subtypes with different etiologies and lack of well-controlled phases due to additional approved salvage therapies, render oncology trials nontrivial and challenging.

It is clear that more focused and informed decision-making based on novel methodologies that maximize information gain throughout the drug development phases, provide clear roadmap to early stop ineffective drugs and speed up development of successful drugs, reduce logistic, time and human resources, and revert safety concerns will increase the probability of clinical trial success rate and lower the overall costs of drug development. Additionally, the prospect of embarking on such paradigm change will be accompanied by newly acquired knowledge that will require various competences. Therefore, it will not be possible anymore to simply rely on unilateral knowledge of a single institute, predominant in the so-called “Big Pharma” model\textsuperscript{5}; the pathway from concept and target validation to patient benefit will rather be a collective effort between different stakeholders.

1.1. Anatomy of clinical trials

Clinical trial is defined as a \textit{planned experiment} that involves human beings and that is designated to compare the effect and value of intervention(s) against a control. Perhaps the essential characteristic of clinical trials is that one uses prospectively the results of a \textit{limited} sample to infer the most appropriate treatment for a \textit{population} of future patients given a medical condition.

Given this definition, a clinical trial must enroll several conditions: (i) It must follow the principles of scientific experimentation, (ii) patients are followed forward in time from a well-defined time point (an identical calendar date is not prerequisite) (iii) one or more intervention(s) are applied to induce change in baseline aspects of participants in the intervention group (unlike observational studies), and (iv) a control group with sufficient similarities to the intervention group must be used to solely assess the action of the intervention. Note that the best available standard of care treatment often confers the most ethical option for a control group. Nonetheless, a \textit{placebo} denoting no active intervention at all is acceptable.
More importantly, the underlying motive of a clinical trial should be the aptitude of “improvement” over the present state of therapy. This is principally translated into the primary question that a clinical trial must address, generally framed as a hypothesis test. The question may be whether an intervention has the postulated effect or not, whether there is no difference in outcome between two interventions, whether the variation of an intervention has a better tolerability profile than the original one, etc. There may also be subsidiary questions that are directly or indirectly related to the outcome variable. In some instances, one might want to see within the trial in progress incidence of cause-specific of death instead of death (response variable is different) or to quantify the difference in surrogate biomarker change in different subtypes of disease progression (response variable is similar). In any cases, if the secondary questions are formed ad-hoc to the protocol, the likelihood of supporting such findings is generally low due to small number of patients in each subgroup. Additionally, it will be less credible since multiplicity of statistical tests inflates the number of false positives. In fact, stating well-defined questions at the start of the protocol should be the foundation of any properly planned and executed trial.

Both primary and secondary questions should be medically relevant to the welfare of the patient. The type of intervention, e.g. a drug, a procedure, a lifestyle modification, should be ideally optimized to promote benefits over the detriments of toxicity in therapy. Unfortunately, adverse or side reactions are major stumbling blocks to the continuity of a treatment afoot. Unlike treatment effect, which adverse event might occur and the severity of it is not always predictable and definitely unacceptable to be full-scale explored for ethical reasons. Therefore, clinical judgment and moral responsibility have bestowed a conservative approach toward the utmost safety of the patient in clinical trials.

The key element for valid inferences is the reciprocation between the intervention and the outcome variable, the primary variable. When scientific questions are conceived, the proper choice of a “measurable” response variable occurring during the course of the trial is of a primordial importance. A primary response variable should ideally be able to be assessed noninvasively as possible in all participants, in the same manner (in terms of precision and accuracy) and insusceptible to interpretation. Furthermore, a valid response variable should be causal to the intervention, hence highly correlated with changes in the intervention’s intensity, amplitude, duration and variability. Lastly, for credibility warrantee, the response variable should preferably be selected with sufficient acceptance in the scientific and medical communities.

Finally, defining the study population is an integral part of the ability to detect the hypothesized results of the intervention, to generalize the findings to a broader population and to assess the study’s merit and appropriateness. The study population is a subset of the population at large with a condition
that strictly meets specific inclusion and exclusion criteria precisely defined in the study protocol. Inclusion criteria may range from demographic factors, such as age and gender, matching pathology and precise disease stage, corresponding genotype and blank medical history. Exclusion criteria may include specific criteria in laboratory values (Absolute Neutrophil Count (ANC), creatinine etc.), risky conditions (pregnancy, congenital heart defect), historical interventions (bypass surgery) etc. Once all these criteria are satisfied, the enrollment process will further reduce the study population to the study sample.

Ultimately, all clinical trials are a compromise between theory and practice. Nonetheless, the wisdom of these decisions determines the validity and the ethical conduct of a trial, which together make clinical trial a powerful experimental technique. One last aspect that has not yet been covered relates to the design and analysis of a clinical trial. The former and the latter majorly impacts on the efficiency of the trial and constitute the main proponents of this thesis.

Figure 1  Potential components of the scientific methodology as applied to clinical trials.

1.2. Design and analysis of clinical trials
Statistical valid clinical trials date back from the eighteenth century\textsuperscript{18,19} but advancement in the field did not further ensue until Louis (1834) first estab-
lished clinical trials on a scientific footing. He introduced the notion of “nu-
merical method” in assessing therapies by defining the need of exact obser-
vation of natural disease progression in untreated patient versus patient out-
come on a course of treatment\textsuperscript{20}. Decades later, Fibiger (1898) in a trial for
diphtheria illustrates the first alternate assignment of patients by distinguishing
two groups: the untreated as a basis of control and the treated one\textsuperscript{21}. However, the concept of randomization and blindness were not applied until
Fisher (1926) introduced the technique to agriculture research\textsuperscript{22} and Amber-
son (1931) to a pulmonary tuberculosis trial\textsuperscript{23}, respectively. It is nonetheless
only in the past few decades that the contemporary form of clinical trial has emerged from the work of Hill (1962)\textsuperscript{24}.

1.2.1. The methodology

The archetype of a sound scientific study should proceed through the general
sequence of events illustrated in Figure 1. From conception to dissemination
of the results, the principles of the method should fulfill the scientific, ethi-
cal, organizational requirements by focusing onto two main aims:

a. The genuine difference in response due to the intervention
b. The avoidance of bias

The genuine impact of the intervention over the absence of intervention must
be tested with statistical methods to assess how strong the evidence is. Addition-
ally, the trial must recruit enough patients to obtain statistically satisfac-
tory precision of the response estimate in each patient group. Both aspects
are directly translated into the fields of hypothesis test and sample size calcu-
lation.

Another concern of any clinical trials should be also to obtain a truthful
answer to the primary question, regardless to “all causes other than sampling
variability”\textsuperscript{25}. This requires that conclusions are drawn from evidences that
are unbiased and without confounding factors. Possible sources of bias could
be selection and allocation bias, investigator bias, ecological fallacy, nonsys-
tematic methodologies, subjective clinical judgment in patient evaluation,
distorted view of therapy or placebo, accidental unblindings, mishandling of
missing data mechanisms, etc. An unbiased study guarantees that statistical
tests will have valid significance levels.

Two types of trials embedding the previously stated scientific principles
are routinely used in clinical investigation. The confirmatory trial should be
sufficient to provide firm evidence in support of efficacy and safety of a new
intervention. Therefore, key hypotheses of interest are always pre-defined
and a precise effect size must be related to a clinical significance. The expl-
oratory trial by contrast is more flexible to design and may entail change of
hypotheses linked to data accrual. Such trials cannot be the basis of the
formal proof of efficacy and safety.
1.2.2. Basic study designs

Like nowhere else than in clinical trials is ignorance highly valued. Presently, randomization and blinding are the two prominent methods to ensure unbiased assessment in clinical investigation of treatment studies (unlike observational studies). Randomized Dose-Controlled Trial (RDCT), currently the “gold standard” of trial design, is a comparative study between a control and an intervention group, in which patient allocation is determined by using a random number-generated algorithm that will assign randomly to each patient the group the patient will enroll in. Randomization could either be fixed (pre-specified allocation probabilities), such as the simple, blocked and stratified, or adaptive (change of allocation probabilities as trial progresses), such as the baseline and response adaptive randomization. A balanced design, which is if the sample sizes for the treatment combinations are all equal, is often the most preferred design, since it provides the largest statistical power and is less sensitive to small departures from the equal variance in any statistical test. Another desirable feature of randomization is the possibility to randomize on different metrics of interest. Figure 2 displays the different possibilities of randomization in a conventional mechanistic pathway of drug response. The idea of selecting a randomization point closer to the clinical endpoint in the sequence of causal events could potentially reduce within-patient variability in the response variable, then the advantage of such approach should be reflected by a higher statistical power.26

![Figure 2. Schematic of potential randomization points in a typical mechanistic pathway of drug response. RDCT corresponds to a dose-, RCCT a concentration-, and RBCT a biomarker-randomization.](image)

Blinding is the act of keeping the participant, the investigator, or both masked to the identity of the assigned intervention. A double-blinded trial
should ideally be favored if the design is possible. Otherwise, a single-blinded trial, in which only investigators are aware of which intervention the patient is taking, might also reduce potential bias in the study. A further extension is the triple-blinded trial, in which the blinding is carried out for the participant, the investigator and the committee monitoring the response variable. Nonetheless, this latter form is flexible to return to a double-blinded trial when directions of observed trends yield ethical concerns.

Apart from bias concerns, treatment assignment is core to the configuration of a study design. For instance, most of trials’ interest lies in “between-patient” comparison in response variable after treatment; therefore, parallel study best fulfills this criterion whereby each patient receives only one treatment. Additionally, this type of study might be more suitable if there are any concerns about carryover effects or if a disease or disorder progression is likely to take place. On the other hand, if one is concerned with intra-subject variability, which is defined as the variability in response occurring within the same patient, one might prefer to utilize a cross-over study, whereby each patient receives several treatments, with wash-out periods in between. The cross-over design allows therefore having the patient as his own control. If two or more interventions are to be evaluated versus a control within a same trial, a factorial design could be utilized. Such an experiment allows the investigator to study the effect of each factor separately, as well as the effect of interactions between factors on the response variable. Finally, it is important to design a study that has a demonstrative purpose. Studies such as superiority trials are designed to demonstrate that one treatment is superior to another one. Non-inferiority trials are intended to prove that a new treatment is at least not appreciably worse than another, and the objective of equivalence trials is to test whether the new treatment is as good as an established one. Dose-response trials, mostly used in drug development strategies, correspond to the investigation of the dose-response curve by which multiple inferences will be made on the efficacy and on the estimation of optimal dosing strategies.

The design of a trial is closely related to the analysis approach intended to be used in the study protocol. In fact, both the design and the analysis strategy must be informed and documented precisely prior to the conduct of the clinical trial. Additionally, any amendment of clinical protocol must be made available without invalidating the legitimacy of the scientific practice.

1.2.3. Traditional analysis methods and drawbacks
The analysis of clinical trials starts by defining which dataset at hand the analysis will be carried on. In practice, protocol violations, patient non-adherence or dropout, incorrect randomizations, etc. always yield to imperfect data and hence may hinder the original planned efficiency of a study design. Therefore, three analysis strategies were proposed to handle these
shortcomings in the data. Intention-to-treat, as the most widely employed method, relies on the basic principle that participants in the trial should be analyzed on the initial treatment assignment and not on the treatment eventually received in the groups. Alternatively, if one considers primarily the treatment regimen that patients actually receive instead of the treatment he/she was randomized to, an as-treated analysis will be better suitable. Lastly, per-protocol or on-treatment analysis is restricted only to the use of data from participants who adhered perfectly to the instructions as stipulated in the protocol, thereby so-called “ideal patients”.

Once the dataset has been properly defined, several analysis methods are applied depending on the primary objective of the trial. These methods might entail multiple testing and adjustment, sample size calculation, analysis of multiple endpoints, handling missing data, prognostic factors, survival analysis, etc. In 2010, the journal Clinical Pharmacology & Therapeutics published a summary of statistical procedures commonly used in the analysis of clinical trials. In this article, although clinical trials are generally longitudinal studies with repeated measurements sampled, merely a few of these procedures allow for longitudinal data analysis while the objectives of most traditional analyses tend to be limited to observations at the end of the trial. Additionally, these tests are often applied individually on each treatment arm in comparison to placebo which make interpolations between treatment arms difficult and reduce the ability to propagate knowledge about dose/exposure–response to future studies. As a consequence, traditional analyses tend to be commonly described as a pairwise comparison of two dose groups, i.e. the lowest and the highest dose group, and at a specific time of the study, i.e. the end of the study, “throwing away” most of the available data and information.

In a drug development perspective, the analysis and interpretation are commonly limited to frequentist statistics based on individual experiments and seldom utilize prior knowledge gained from previous stages, other drug candidates, competitors, and experimental systems. Given a performance criterion in measuring the success of a study, e.g. p-value, the development pace focuses on moving as soon as possible the candidate to the next milestone. Bridging consecutive experiments is rarely made to best characterize the compound and confirm the mechanistic interpretations of the model parameters. As a natural consequence, any factors or changes delaying the development timeline, although it is knowledge-generator, are barely considered; hence adaptive strategies are mostly not practiced.

Another consequence of such directives is the nature of the decision making itself, often based on the perceived requirement for statistical significance (see section 1.4.1) in few attributes separately, rather than quantitative risk-benefit assessments based on probabilities of achieving a global target. Therefore, models used for data analysis do not necessitate expressing the same complexity as required in system biology or in disease models. In most
cases, they are actually largely empirical, mostly restricted by the data quality and quantity and their building process relies primarily on describing the observed trend in the data. Thus, this general statistical modeling generates models independent on the subject area and on assumptions but less apt to extrapolate to a different setting. In fact, these models are mostly developed in late stage clinical development, as a confirming tool, rather than serving as an instrument to guide the process.

Obviously, these few examples were just cherry-picked, but such approaches clearly provided the ground to consider alternative analysis strategies and examine how different the traditional versus the new methods are performing. For instance, power calculation difference and effectiveness of dose-escalation oncology trials between traditional methodologies and new methodologies were covered in Paper II and Paper V, respectively.

1.3. Model-based drug development

More recently, a promising field so-called “Model-based drug development” (MBDD) as defined in the FDA “Critical Path Initiative” document has emerged considerably over the past two decades as a substantial tool to improve drug development and decision-making. Grounded on Sheiner’s “learning” (hypothesis generating) and “confirming” paradigm, MBDD exemplifies the continual process of drug development by integrating quantitative analyses of all accrual data collected across trials, doses, compounds etc. in order to reduce the uncertainty in the knowledge of the new compound of interest. Depending on the particular approach employed, MBDD can provide support at different critical levels in pharmaceutical research as illustrated in Figure 3: (i) a better understanding of the underlying mechanisms of a disease or drug action, (ii) a natural bridging platform for incorporating prior knowledge and pooling data across studies, (iii) more efficient data analysis by utilizing longitudinal data over time and multiple response variables, (iv) a forecasting framework for individualized treatments and Therapeutic drug monitoring (TDM), (v) a possibility to extrapolate to other target populations, and (vi) an improvement of study design by using optimal design (OD) theory and clinical trial simulations (CTS). In fact, Jonsson and Sheiner have shown that the use of model-based statistical tests can improve the efficiency of clinical trials.

At the center of this relatively new paradigm is the concept of Quantitative pharmacology, a mindset that emphasizes the quantitative integration of relationships between drug characteristics, disease development, and individual variabilities across studies and development phases. The quantification usually means an iterative construction of mathematical and statistical models which serve to leverage the focus of the drug development. The use
of such quantitative approaches is described in the scientific discipline of pharmacometrics.

![Figure 3](image.png)

Figure 3. Quantitative Model-Based Drug Development (MBDD) as a cornerstone of the drug development process. Listed are potential applications embedded in the two consecutive learn–confirm cycles (arrows) driving the process of drug development.

Pharmacometrics is the scientific discipline that uses modeling and simulation as primary tools to facilitate the “learning-confirming” process. The act of modeling means here to describe mathematically the aspects of a data-generating system, which the scale can range from a clinical trial, the human or animal body to a cell unit. The complexity varies from an empirical description suitable to its intended use to a multi-compartmental realistic depiction. Therefore, a given model is usually topic-specific, inexplicitly rich in assumptions and may take in various forms such as dynamical systems, statistical models, and differential equations.

In the context of MBDD, several components are usually interconnected. A Pharmacokinetic–Pharmacodynamic modeling is dedicated to the description of the links between the change in drug concentration over time (PK) to the relationship between the concentration and the intensity of the observed response (PD), both beneficial and adverse². This nominal PK–PD model can be altered or extended in several ways, such as using an “exposure”, e.g. area under the plasma concentration-time curve or average steady-state concentration, subcomponents of the concentration-response interface such as biophase distribution, biosensor process and transduction, surrogate biomarker, trajectory of the disease and placebo effect. Ultimately, the evol-
tion in complexity of the model is driven to a large extent by the aspiration to better describe the entire system, such as endorsed by the physiological-based pharmacokinetic (PBPK) field\textsuperscript{39}. Resultant physiological or semi-mechanistic PKPD models are then used in a simulation setting to generate different clinical trial design strategies for future trials to be analyzed subsequently. This typical circular knowledge generation and propagation allows a constant refinement of the scientific knowledgebase for the drug candidate under consideration.

1.3.1. Nonlinear mixed-effects models
Most pharmacometric research is based on nonlinear mixed-effects models (NLMEMs), which allow recognition of multi-level random variations inherent to the biologic data. In population modeling, parameters are assumed to take a typical individual value in the population (fixed effects), with variances (random effects) describing the differences between individuals (between-subject variability) and occasions (intra-subject variability), and the variability due to measurement and model misspecification errors (residual error).

For continuous data, the general NLMEM can be formalized as:

\[ y_{ij} = f(t_{ij}, g(\theta, \eta_i, x_i, a_i, \kappa_i) + h(t_{ij}, g(\theta, \eta_i, x_i, a_i, \kappa_i), \epsilon_{ij}) \] (1.1)

With \( f(.) \) the prediction function mapping the observation \( y_{ij} \) for individual \( i \) at time \( t_{ij} \) and \( h(.) \) the residual error function describing the deviations between predictions and observations with the random variable \( \epsilon_{ij} \). Individual parameters are further described by \( g(.) \) the vector function of population value \( \theta \), the between-subject variability \( \eta_i \) and intra-subject variability \( \kappa_i \), discrete design variables \( x_i \) (e.g. dose) and individual specific covariates \( a_i \). All random effects \( \eta_i, \kappa_i \) and \( \epsilon_{ij} \) are assumed to be normally distributed with mean zero and covariance matrices \( \Omega \), \( \Gamma \) and \( \Sigma \), respectively. These matrices describe the correlations between the individual parameters (\( \Omega, \Gamma \)) and between the residual error parameters (\( \Sigma \)). Additionally, the parameters in the general function \( g(.) \) can have any transformed parameter distribution, e.g. normal additive or proportional, log-normal, etc. The residual error can also vary, e.g. homoscedastic (additive), heteroscedastic (proportional), combination of both, serially correlated etc.

On the other hand, the NLMEM for discrete data is probabilistic and describes the probability of observing the individual response \( y_{ij} \) given model parameters, such as:

\[ p \left( y_{ij} \mid t_{ij}, g(\theta, \eta_i, x_i, a_i, \kappa_i) \right) \] (1.2)
where $p(\cdot)$ is the probability (likelihood) function, and the other function, variables and parameters are defined as before.

Finally, the population parameters can be combined for simplicity purposes into a vector $\Theta$ incorporating all population parameters ($\theta$, $\Omega$, $\Gamma$ and $\Sigma$).

### 1.3.2. Maximum likelihood estimation

Given the data collected in an experiment and the specification of a model, the next step is inevitably the estimation of the model parameters. Maximum Likelihood (ML) methods are commonly used in NLMEM and seek to find the set of parameters for which prediction is most likely matching the data, i.e. the highest probability for these data to occur under the model (see Figure 4). Given the parameters, the function defining how likely the data are, is the Log-Likelihood function (LL). For NLMEM, the individual $L_i$ contribution to the LL for given population parameters $\Theta$, is described as:

$$L_i(y_i, \theta) = \int_{-\infty}^{+\infty} P(y_i | \eta_i, \kappa_i, \theta, \Sigma) \cdot P(\eta_i, \kappa_i | \Omega, \Gamma) \, d\eta_i \, d\kappa_i$$  \hspace{1cm} (1.3)$$

where $P(y_i | \eta_i, \kappa_i, \theta, \Sigma)$ is the individual data density and $P(\eta_i, \kappa_i | \Omega, \Gamma)$ the population parameter density of the individual random effects, and their product referred to as the joint likelihood density.

The population LL function or marginal likelihood for given population parameters $\Theta$ is the product of all individual $L_i$ over all subjects $N$ in the data, given as:

$$L(y, \theta) = \prod_{i=1}^{N} L_i(y_i, \theta)$$  \hspace{1cm} (1.4)$$

In numerous regression and statistical softwares, the population LL also called Objective Function Value (OFV) is calculated instead as a sum of the logarithms of all $N$ individual $L_i$ in order to reduce computational issues. It is further transformed to a negative value to minimize instead of maximizing the OFV which could then be rewritten as:

$$L(y, \theta) = -2 \cdot \sum_{i=1}^{N} \log L_i(y_i, \theta)$$  \hspace{1cm} (1.5)$$

The computation of the marginal likelihood over all random effects is non-trivial and has no closed-form solution for NLMEM. This absence of analytical solution is due to the random effects entering in the model in a nonlinear manner. Thus, numerical approximation methods must be used to approximate the marginal likelihood. Generally, these approximation methods can be classified into two categories: Gradient-based algorithms, which are deterministic and linearize the model with respect to $\eta_i$ and $\kappa_i$, and the Expec-
tation-Maximization algorithms, which are sampling-based approximation of the LL integral (Monte-Carlo integration).

Figure 4. Schematic for the Maximum Likelihood $\mathcal{L}(y, \Theta)$ estimation method. The top-left panel represents $f(.)$ the prediction model at different tested typical parameter values $\theta_n$ and the bottom-left panel the corresponding Likelihood values. The right panel represents the approximation methods of the marginal likelihood.

Numerous softwares incorporate these methods. For instance, NONMEM\textsuperscript{40}, Monolix\textsuperscript{41}, R (nlme package)\textsuperscript{42}, SAS (NLMIXED procedure)\textsuperscript{43}, etc. NONMEM, which was initially developed by Beal and Sheiner at the University of California in San Francisco, and now maintained by Icon Development Solutions is by far the primary modeling software for most pharmacometrists and contains most of the previously stated algorithms. Three gradient-based approximation methods used in this thesis are briefly described below.

**Gradient-based algorithms**

Gradient-based algorithms are iterative by nature. They require local differentiability of the function by which the gradient at the currently evaluated point will define the search directions for the next iteration. The differentiation method used here is a second-order Taylor series expansion around the random effects of the conditional distribution of the observed data, which provides a quadratic local approximation of the function at the current design point. This second-order differential contains the Hessian matrix of the logarithm of the joint density with respect to $\eta$ which is often complex and computationally expensive. Depending on the trade-off of accuracy and computation time, the three presented algorithms differ in the computation of this Hessian matrix.
The Laplace approximation

The Laplace algorithm provides the most precise approximation of the gradient-based algorithms, and is the only suitable method to fit categorical data among these 3 methods. The reason behind this is that the Laplace method linearizes the model around the mode of the joint density of the random effects, involving both the first and second terms of the Taylor expansion of the integrand. Because the Hessian matrix is used, this method is expected to be closer to the true underlying marginal likelihood, simply because it uses a higher order approximation. It has naturally also the longest runtime.

The first-order conditional estimation (FOCE)

The FOCE algorithm is the most frequently used algorithm for continuous data models. Such as in the Laplace method, the FOCE algorithm linearizes the model around the conditional estimates of $\eta$ and $\kappa$. However, the Hessian is approximated by its expected value under the assumption of the Gaussian distribution of the individual data, which in this case corresponds to a function of the model gradient with respect to $\eta$ and $\kappa$. This is the primary reason why the FOCE algorithm is considered as a first-order method. Interactions between $\eta_i$ and $\varepsilon_{ij}$ can be accounted by the addition of extra terms in the alternative FOCE with interaction (FOCEI) method.

The first-order approximation (FO)

The FO method approximates the Hessian matrix similarly as in the FOCE method. Nevertheless, the model is linearized around the median of the individual parameters, i.e. random effects set to zero. For $\eta$-dependent residual error functions, a variant of the FO algorithm exists as well, so-called First-order with interaction (FOI).

In comparison to other approximations, the FO method is faster but has been proven to produce biased estimates. Additionally, the efficiency of the method decreases with increasing inter- and intra-individual variability and increasing deviations from the normality assumption of the data. Therefore, this approximation is not suitable for probabilistic models, hence for discrete data.

1.4. Model-based trial data analysis

Since longitudinal repeated measurement study is the paramount method to assess drug treatment over time, a model-based data analysis provides an elegant and effective way to utilize fully the information contained in the data. It allows the use of sparse and rich data within a same analysis, delineates the population parameters from the subject-specific variations such as found in covariates, permits flexibility in trial designs by merging analysis of
balanced and unbalanced designs, integrates multiple mechanisms of action and multiple endpoints, allows inter- and extrapolation for non-tested dose groups and treatment duration, incorporates missing data mechanisms, analyses simultaneously various types of data, etc.

The following sections introduce the trial analysis components that were covered and further extended in this thesis with the assistance of a pharmacometric approach.

1.4.1. Hypothesis tests and statistical power

Most often, confirmatory evidence is generated through testing an alternative hypothesis (H₁) against a null hypothesis (H₀) using a "test of significance"⁴⁶. This hypothesis test provides a sound inference whether results obtained from the trial contain enough information to cast doubt on conventional belief that has been used to establish the null hypothesis. A statistically significant result must thereby have been demonstrated to have occurred not by chance alone, given pre-determined threshold probabilities of two conceptual types of error: type I and type II errors. Figure 5 illustrates the relationship between the different components of a hypothesis test.

![Hypothesis Testing Components](image)

*Figure 5. Statistical hypothesis testing components and distributions of an observation when H₀ is true or false for a one-sided test. “Null” represents the mean outcome variable given no treatment, “Alternative” the mean with the postulated effect size, α the type I error rate, β the type II error rate, and P the computed statistical power.*

Given a probabilistic framework, a hypothesis test in a dose-response trial can be formalized as the decision between two mutually exclusive hypotheses such as:

\[
\begin{align*}
H_0 & : \Psi(\Theta) = 0 \\
H_1 & : \Psi(\Theta) \neq 0
\end{align*}
\]  

(1.6)
Where $\Psi(.)$ is a function mapping model parameters $\Theta$. A proper way to
determine whether to reject the null hypothesis is to compute a test statistic
$t_{\text{stat}}(\hat{\Theta})$ which is a scalar value summarizing all the observations by a single
number. If the conventional decision rule is based on $p$-value, the test statistic
is converted to a conditional probability defining the strength of evidence
in the data against $H_0$ given the null hypothesis was true. The test statistic is
next compared to the decision boundary given by the equation $t_{\text{stat}}(\hat{\Theta}) =
t_{\text{critic}}$ according to a pre-specified significance level $\alpha$ (the probability of
making a type I error, i.e. the false identification of a drug effect), such as a
significant left-tailed test, i.e. rejecting the null hypothesis, is described as:

$$p_{\text{value}} = \int_{-t_{\text{stat}}(\hat{\Theta})}^{t_{\text{critic}}(\hat{\Theta})} p(t_{\text{stat}}|H_0) \; dt_{\text{stat}} \leq \alpha$$

with $\alpha = \int_{t_{\text{critic}}(\hat{\Theta})}^{t_{\text{stat}}(\hat{\Theta})} p(t_{\text{stat}}|H_0) \; dt_{\text{stat}}$ (1.7)

With $p(t_{\text{stat}}|H_0)$ being the pdf of $t_{\text{stat}}(\hat{\Theta})$ and $\alpha$ typically set at 0.1, 1 or
5% in behavioral research. As such, if the $p$-value is small, the probability
that the observed data can randomly arise from the null hypothesis distribu-
tion is unlikely, hence the rejection of the null hypothesis. By contrast, a
high $p$-value suggests that the observed data is consistent with the assump-
tion that the null hypothesis is true; therefore, the null hypothesis could be
retained.

In practice, a $p$-value can match any form of inequalities expressed in the
hypothesis test. Thus, a corresponding test statistic can be formulated such as
if deviations of the estimated parameter are considered possible in either
direction from a postulated value, a two-tailed test will be used (the region of
rejection $\alpha/2$ is on both sides of the sampling distribution); in contrast, if
only deviations in one direction are considered possible, a one-tailed test is
used, e.g. right-tailed or left-tailed test. Additionally, a test statistic follows
usually a distribution determined by the function used to define that test sta-
tistic and the hypothesized distribution of the observational data. Conse-
quently, different null hypothesis tests exist such as the $z$-test for normal
distribution, $t$-test for Student's $t$-distribution and $f$-test for $f$-distribution.

In this thesis work, the two-tailed $t$-test was selected as a referential
method of comparison in Proof-of-Concept (POC) and Dose-Finding (DF)
studies. POC studies are designed to give preliminary evidence of efficacy
and safety, with the aim to inform a decision about proceeding into full de-
velopment of the drug. On the other hand, DF studies are performed to ex-
plor e a dose–response relation. In practice, the POC decision is often based
on whether a required effect size can be detected in comparison to placebo or
a comparator treatment given a specific study design.

When designing a study, a statistical hypothesis is framed primarily in de-
termining the number of study participants (sample size) required to ensure a
minimal probability of detecting a drug effect if it exists. The power of a
study π corresponds to this probability and is defined as the probability of rejecting the null hypothesis if the alternative hypothesis is true, given the type II error β (failing to assert an existing drug effect). It can then be computed using:

$$\pi = 1 - \beta = \int_{-\infty}^{t_{\text{critic}}} p(t_{\text{stat}}|H_1) \, dt_{\text{stat}}$$ (1.8)

With $p(t_{\text{stat}}|H_1)$ the pdf of $t_{\text{stat}}(\bar{\theta})$ and β usually set by the regulatory agency and the sponsor of a trial (typically 20%). The primary focus of proper sample size calculation lies in having sufficient statistical power given all possible factors from the prospective data that might influence the outcome, such as displayed in Figure 6.

![Diagram](image)

Figure 6. Possible set of factors influencing the statistical power of a study. Positive and negative correlations to the statistical power are depicted in this diagram.

The underlying reason of properly powering a trial is in fact to minimize variations (standard error) of the unknown effect estimate by increasing the sample size and thus makes the standardized effect size larger. Thus, sample sizes are judged by the quality of the resultant estimates that one can predefine with acceptable confidence interval for an estimate or a target value for the power of a statistical test to be applied. For instance, for an independent two-sided $t$-test (used in Paper II), the test statistic is expressed as:

$$t_{\text{stat}} = \frac{\bar{\theta}_1 - \bar{\theta}_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \quad \text{with}$$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_0-1)s_0^2}{n_1 + n_0 - 2}}$$ (1.9)
where $\hat{\mu}_0$ and $\hat{\mu}_1$ are the estimated means of the null and alternative group, $n_0$ and $n_1$ their respective sample size, $s_p$ the pooled standard deviation of $s_0$ and $s_1$ their respective standard deviation. The required sample sizes given $t_{\text{critic}}$ may be derived using simple algebra in this trivial example. However, more technically challenging sample size methods can be used depending on the nature of the data (continuous, dichotomous or time-to-event etc.), the choice of analysis (single observations or longitudinal data), the design (number of comparison groups), etc. For instance, in repeated measures experiments, several sample size computation methods exist based on ANOVA or ANCOVA, the Generalized Estimating Equation (GEE)$^{47-52}$, Confidence Intervals, and Moderated Multiple Regression (MMR). However, in NLMEs, two main hypothesis tests are primary used for sample size computation: the likelihood ratio test (LRT) and the Wald test.

**Likelihood ratio test**

The LRT is a statistical hypothesis test based on the value of a ratio calculated between the likelihood functions of two competing but nested models that are defined according to a hypothesis test. Consequently, the reduced model ($H_0$) and the full model ($H_1$) are compared in the likelihood domain and can be written as:

$$t_{\text{LRTn}}(\hat{\theta}) = \frac{\mathcal{L}(y_n, \hat{\theta}_n^{H_1})}{\mathcal{L}(y_n, \hat{\theta}_n^{H_0})}$$  \hspace{1cm} (1.10)

with $\mathcal{L}(.)$ representing the likelihood evaluated for the observed data $y_n$ under both hypotheses and $n$ the sample size. Again, the LRT can be computed as twice the difference in their log-likelihoods. Given the null hypothesis is true, the LRT statistic follows a chi-squared distribution with degrees of freedom equal to the difference in dimensionality $df_1 - df_0$ representing the number of free parameters of the alternative and the null model, respectively. If the alternative is true, it follows a non-central chi-squared distribution with similar degrees of freedom and the non-centrality parameter $\lambda$.

Power calculation for the LRT has been described based on multiple simulations and re-estimations of the full and the reduced model$^{53-55}$. Sample size calculations in NLMEs remain however time-consuming and computer-intensive (many replicated datasets for one tested sample size) and embed drawbacks such as the assumption of both the correct degrees of freedom for the hypothesis variable in consideration (e.g. the degree of freedom for a random effect) and the distribution shape around the null hypothesis (i.e. if the chi-squared distribution is achieved). Additionally, a need for correction for type I error inflation has been demonstrated by Wahlby et al.$^{45,56}$ to be necessary for small sample sizes, demanding therefore more computational
effort. A faster and reliant alternative method was the subject of Paper I in this thesis work.

**Wald test**

Unlike the LRT, Wald test depends only on the estimates of the alternative model. The general Wald test statistic of a nonlinear hypothesis becomes therefore a quadratic form in the vector $\Psi(\hat{\theta})$ and the inverse of a matrix that estimates its covariance matrix, such as:

$$t_{WNL} = n \cdot \Psi^T(\hat{\theta}) \left( R(\hat{\theta})I(\hat{\theta})^{-1} R^T(\hat{\theta}) \right)^{-1}\Psi(\hat{\theta})$$

under $H_0: \Theta = 0$, $t_{WL} = n \cdot \frac{\hat{\theta}}{SE(\hat{\theta})}$ (1.11)

with $I(\hat{\theta})$ be the Fisher Information Matrix (FIM), $R(\hat{\theta})$ the Jacobian matrix of $\Psi(\hat{\theta})$, $SE(\hat{\theta})$ the corresponding standard error and $n$ the sample size. Similar to the LRT, the Wald test is asymptotically distributed as chi-squared with $df_i - df_0$ degrees of freedom when the null hypothesis is true.

With the availability of a closed-form expression for the non-centrality parameter, it is straightforward to compute the sample size for a given power and the critical value for a given value of Type I error $\alpha$.

1.4.2. Handling censored and missing data

Incomplete data is almost always an integral part of a data analysis. Respondents may decline to answer some items in a questionnaire, data collection may have been poorly conducted or a patient may drop-out due to toxicity of the drug under investigation. Sometimes the portion of such missing data can become sizeable and failing to account for it might cause substantial invalidity about the model and the power of the trial.

The most appropriate way to handle missing data will depend upon how data points are missing and how much information it is known about their value. Missing observations are characterized by the fact that the value is completely missing. A specific terminology has been adopted to describe the processes generating the missing data that are typically classified into 3 categories: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). By contrast, censored observations are characterized by the fact that factual value of the observation is missing, but not the interval in which the value lies beyond or within. It is thereby assigned to the MNAR classification, since the missingness is related to the missing value itself. Four types of censoring exist: Left censoring for data points below a certain threshold value, right censoring for data
points above a certain value, *interval censoring* for data points between two values, and *random censoring* when the censoring times are independent of their failure time.

**Figure 7.** Illustration of the concept of detection limit and quantitation limit by showing the theoretical normal distributions associated with blank, detection limit (LOD), quantification limit level (LLOQ and ULOQ) samples. In parallel, distributions of observed concentration samples with respect to time are displayed for the dynamic range of the bio-analytic method. Hashed pattern represents the values being censored.

For a bio-analytical method the lower limit of quantification (LLOQ) as well as the upper limit of quantitation (ULOQ) constitute both a case of censoring in the assessment of the pharmacokinetic (PK) of a drug. The reason behind this is primarily intrinsic to the detection of the raw analytical signal for measurements below the LLOQ as illustrated in Figure 7.

During a conventional method validation procedure, a variety of technical operating parameters, such as accuracy, precision and dynamic linear range is evaluated. Two key parameters are the Limit of Detection (LOD) defined as the lowest concentration of an analyte detectable and clearly distinguishable from background noise or "blank" (typically 3 times the signal-to-noise, based on the standard deviation of the blank), and the Limit of Quantification (LOQ) being the lowest concentration of an analyte that can be determined with a stated degree of reliability in terms of precision (repeatability) and accuracy. Typically, the LOQ is defined as 10-fold the standard devia-
tion of the blank and should ensure an inter-precision of 20% at least for any samples at the LOQ. This alternatively signifies that any measurements below the LOQ might not have a value with enough statistical certainty to be trustfully used. It is therefore not unexpected that US Regulatory guidelines advise that “concentrations in unknown samples should not be extrapolated below the LLOQ or above the ULOQ of the standard curve”, despite a general agreement among PK scientists, that reporting all measurements would be a better solution\textsuperscript{60,61}.

The effect of such directives, in practice, is that most laboratories do not report the quantitative value but merely the qualitative information, i.e. a plasma concentration is known to be below the LLOQ but it is unknown by how much. Additionally, since it is rather assumed that the concentration could not be below 0, Below Quantification Limit (BQL) observations are generally treated as an interval censoring between 0 and a specific LLOQ.

In traditional statistical analyses, these values are often omitted or substituted with a constant value, such as half the LLOQ or zero. These methods result in a uniform distribution for those substituted values, instead of their true respective distributions. As a result, this can produce questionable descriptive statistics and more importantly, introduce bias in parameter estimates\textsuperscript{62,63}. An alternative method uses the characteristics of the distribution of the values above the LLOQ to estimate the values below the LLOQ using maximum likelihood estimation methods\textsuperscript{64}. This joint modeling of the probability of the observation being censored and the response variable of interest have demonstrated to reduce parameter bias. Nonetheless, censored and missing observations still constitute a complicating factor in the analysis of repeated-measures longitudinal study that is typically the case in clinical research.

Table 1. Methods for handling BQL data in a model-based analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Discard all BQL observations</td>
</tr>
<tr>
<td>M2</td>
<td>Discard BQL observations and apply the method of maximum conditional likelihood to the remaining observations</td>
</tr>
<tr>
<td>M3</td>
<td>Maximize the likelihood for all the data and treat BQL observation as censored</td>
</tr>
<tr>
<td>M4</td>
<td>Like M3 but the likelihoods for data above and below the LOQ are conditioned on measurements being greater than zero</td>
</tr>
<tr>
<td>M5</td>
<td>Replace all BQL observations with LLOQ/2</td>
</tr>
<tr>
<td>M6</td>
<td>Replace the first BQL observation with LLOQ/2 and discard consecutive succeeding BQL observations</td>
</tr>
<tr>
<td>M7</td>
<td>Replace all BQL observations with zero</td>
</tr>
</tbody>
</table>

In the context of NLMEMs, Stuart Beal published in 2001 an overview of possible approaches to handle BQL data in a model-based analysis\textsuperscript{65} as displayed in Table 1. Generally, these techniques are referred to as the “M-
methods” and are either classified as discard-, substitution- or likelihood-based methods. The primary methods involving ML estimation are method M2, M3 and M4. With the method M2, the likelihood for the data assumes that all non-BQL values are by default right-censored at the LLOQ, therefore conditional likelihood for the data being above LOQ are maximized with respect to the model parameters. The methods M3 and M4 are both based on simultaneous modeling of continuous and categorical data where the BQL observations are treated as categorical data. The likelihood in these cases is divided into 2: a formal likelihood for observations above LLOQ is maximized with respect to the model parameters and the likelihood for a BQL observation is taken to be the likelihood that it is indeed below the LLOQ. The only difference between M3 and M4 stands in the way that the M3 method assumes no lower bound for any measurement values (−∞ is permitted per se), while the M4 method assumes for observations above and below the LLOQ an interval censoring between ULOQ and 0 and LLOQ and 0, respectively.

Several publications investigated different methods of handling BQL data and their effect on model selection and parameter estimation in NLMEMs. Generally, these studies have shown that dropping BQL data leads to bias and imprecision in parameter estimates, and likelihood-based methods, notably the M3 method, avoid such impediments even with 30% of the collected data below the LLOQ. Nonetheless, although these methodologies can be used as a last resort in data analysis, it could be beneficial to design studies to avoid unnecessarily large amounts of data outside of the range of quantification, while balancing the potentially high information content of data where the potential of LOQ measurements is high. This could possibly wean pharmacometric analyses off from often being referred to as “rescue” analyses of ill-planed trials. Such novel perspective of clinical trial designs including LOQ information was explored in Paper III of this thesis using a model-based optimal design approach.

1.5. Model-based trial design

A design which generates data incapable of providing meaningful answers cannot be easily salvaged retrospectively. R.A. Fisher himself described the importance of planning of an experiment, such as “To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of”68. Thus, the importance of getting the study design correct cannot be overstated.

Control over the design of clinical studies is one of the arenas that pharmacometricians can intervene. The opportunity to leverage the type of data to be collected, the timing of its collection, the number of patients to be in-
cluded in, etc. is possible with Monte-Carlo (MC) methods and optimal design of experiments being the central components.

1.5.1. Stochastic/Monte-Carlo simulations

Due to the inherent multiple levels of random effects in NLMEMs, a single clinical trial represents merely a sample of the infinite number of possible realizations. Trials need thereby to be replicated over multiple conditions in order to better characterize the central tendency of the outcome and to better quantify the variation among the population. MC simulation methods are a broad class of computational algorithms that rely on repeated random sampling to investigate the distribution of an unknown probabilistic entity. Within pharmacometric area, MC simulations allow to assess the overall robustness of a model, a statistical test or a methodology, to evaluate the efficiency of a trial design though CTS, to compute the uncertainty in an outcome of interest, to visualize the possible spectrum of a variable, to investigate the impact of the model assumptions, to retrieve the solution of complex integrals, etc.

In the context of trial designs, the most frequently use of the MC methods is the CTS. CTS is the generation of multiple datasets of individuals simulated from the random effects of a given model and a given planned design, in order to study a large variety of outcome metrics. Often, the simulation stage is followed by an evaluation of the model for each of the samples in order to evaluate the overall efficiency of a design to provide reasonable expected parameter precisions. This procedure is commonly referred to as Stochastic Simulation and Estimation (SSE) and may aid in trial design selection. Nonetheless, such cyclic procedures are to be repeated for each proposed design, which inevitably render MC methods computational expensive and with tedious runtimes rising with the number of objectives being studied. Additionally, MC simulations are generally heuristic, and do not always ensure the global solution of the optimization problem. It is therefore conceivable that alternative methodologies may be required regarding to these multi-objectives problems.

1.5.2. Cramér-Rao inequality

Optimal design theory is based on the premise that the information contained in the resulting experimental data is conditioned on the design variables of an experimental setup. Alike the MC methods, OD focuses therefore on varying design parameters to provide optimal trial results for the subsequent data analysis. However, instead of multiple SSEs, OD is based on a formal mathematical inequality established by Cramér and Rao, which summarizes both the relationship between the information content of a trial and its
theoretical upper limit given a trial design. This asymptotic behavior formed the Cramér-Rao bound (CRB) and is a fundamental part of the OD theory.

The CRB establishes the inverse of the Fisher Information Matrix (FIM) as the asymptotic lower bound of the variance-covariance matrix of any unbiased estimator\(^{69,70}\), formalized as:

\[
I(\xi, \hat{\theta})^{-1} \leq \text{COV}(\hat{\theta})
\]  

(1.12)

where \(I(\cdot)^{-1}\) is the inverse of the FIM, \(\xi\) the vector of design variables and COV the variance-covariance matrix for \(\theta\) given observations \(y\).

The FIM is not data-dependent since it can be defined mathematically with respect to the parameters and its inverse be interpreted as the expected variance-covariance matrix which yields to the expected parameter standard errors (SE). Thus, the CRB links the data-dependent variance-covariance matrix to a quantity which can indicate the amount of information about the unobservable parameters given a design, and be calculated prior to the collection of the data. The CRB provides therefore an opportunity to optimize a future design by selecting the experimental design variables such that the FIM is maximal, hence the parameter SEs minimal. The difference between MC simulations and OD theory in the search of the most efficient design is displayed in Figure 8.

1.5.3. Population Fisher information matrix

The FIM is mathematically defined as the expectation over the second moment of the score function (i.e. the partial derivative of the log-likelihood with respect to the parameters)\(^{71}\). However, since the first moment of the score function for the maximum likelihood estimates is zero (i.e. maximum of a function at its first derivative equaled to zero), the FIM is also the variance of the score such as illustrated in Figure 9. Alternatively, if LL is twice differentiable with respect to \(\theta\), and under certain regularity conditions, the FIM may also be written as:

\[
I(\xi, \hat{\theta}) = -E\left( \frac{\partial^2}{\partial \theta^2} \log L(y, \hat{\theta}) \right)
\]  

(1.13)

For a general NLME, the “population” FIM cannot be calculated in an analytic way and approximation methods have to be used. In 1997, Mentré et al. proposed a closed-form solution of the FIM for continuous data by using the FO approximation to linearize the model around the typical response, i.e. setting the individual random effect parameters to zero\(^{72}\). Since the original FIM approximation was only for homoscedastic residual error model with a known residual variance parameter, the work was further extended to
include heteroscedastic residual error models with other distributions\textsuperscript{73–75}, other variability components and the covariances between the fixed and random effects\textsuperscript{76,77} to form the commonly called full FIM (as opposed to the reduced FIM with off-diagonal elements set to zero). In this work, unless stated differently, the FIM was calculated using the FO approximation as described in Nyberg et al.\textsuperscript{78} and is by default set to a full FIM.

A convenient property of the FIM, which was utilized at multiple occasions in \textit{Paper III and IV}, is its additivity with respect to experimental design units, e.g., the FIM for an experiment with two groups is simply the sum of the group FIMs.

1.5.4. Design optimization

Since the FIM is the negative of the expectation of the second derivative of the LL function with respect to $\Theta$, information can be interpreted as a measurement of the "curvature" of the support curve near the maximum likelihood estimate of $\Theta$. A flat support curve would have a low negative expected second derivative, and thus low information; while a sharp one would have a high negative expected second derivative and thus high information. Within OD, the primary goal of a design optimization is therefore to maximize the FIM by finding the optimal set of design variables, e.g. optimal sampling times or their corresponding sampling windows, subject assignment to different elementary designs, discrete doses, covariates, etc.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure8.png}
\caption{Model-based trial design methodology comparison between MC simulations and Optimal Design of experiments. In MC simulations, multiple designs are simulated to generate the design with minimal variance of parameter $\theta$. In contrast, optimal design theory minimizes the inverse of the Fisher Information Matrix (FIM), which will asymptotically decrease the actual parameter covariances given the data.}
\end{figure}
Generally, a scalar function is used to measure the magnitude of the FIM and a multitude of different functions with different qualities exists\textsuperscript{68}. The determinant of the FIM is the most common function used in practice and designs which are optimal with respect to the determinant are referred to as D-optimal designs. Another alternative of the D-optimality is the Ds-optimality criterion, in which only a subset of parameters of interest is targeted in the optimization, hence estimated with the highest possible precision. However, both of these two methods reveal limitations due to the assumptions made on a model structure and on the known parameter estimates to calculate the FIM. One option to account for parameter uncertainty may be the use of a global design approach, such as ED-optimality\textsuperscript{79}. Global optimal design techniques assume that the parameters come from distributions and hence allows for incorporation of uncertainty in parameter estimates. Generally, this family of E-criteria is considered more robust than local optimal design methods. When the prior information of the parameters is of high quality, a Bayesian OD is a suitable option, especially if one considers to optimize on the individual level, given a population prior\textsuperscript{80,81}. In theory, any other user-defined criterion could be utilized for optimization\textsuperscript{68}, especially if yielding to a specific objective different from the parameter precision. The latter was explored in an oncology example for dosing regimen optimization in Paper IV.

![Figure 9. Schematic of the mathematical relationship between the log-likelihood function $\mathcal{L}(\cdot)$, the Maximum Likelihood Estimates and the Fisher Information Matrix FIM for two designs, a very informative (high FIM, i.e. low variance) and an uninformative (low FIM, i.e. high variance) one.](image-url)
Once the scalar objective function is defined according to one of the criteria previously stated, a generic optimizer is used to generate the best setup of experiment. Again, the FIM is used here to compare different experimental designs and to find the best experimental design among a set of candidates. However, before this level is reached, a number of characteristics may be considered such as the sequence of the optimization when optimizing over multiple design variables, the stiffness of the differential equations resulting in alternate ODE solvers, the search algorithm strategies which will provide a faster and systematic search in the defined design space, etc. This whole set of considerations makes the optimization of trial designs a nontrivial exercise, even occasionally reluctant to the pharmacometrician community itself. Nonetheless, with increasing applications of OD, this approach has since demonstrated usefulness within different areas of drug development and has been developed into several established softwares.

1.6. Oncology dose-escalation trial design

Oncology is a therapeutic area where intensive drug development is underway. However, due to the large heterogeneity and complexity of the disease, the task is rather challenging and daunting. Additionally, most drugs in this field cause severe side-effects and the recommended dose is often based on the tolerability instead of the efficacy. Therefore, the research area has been struggling with a large proportion of drugs failing in late phases of development due to unacceptable toxicity or lack of efficacy. As a fact, only 5% of drugs tested in human proceed to marketing approval.

1.6.1. Chemotherapy treatment

Chemotherapy is a category of cancer treatment that uses chemotherapeutic agents to effectively target fast-dividing malignant cells. As these drugs act by killing cells, they are termed cytotoxic. They prevent mitosis by various mechanisms including the inhibition of certain cellular components involved in the cell division machinery, the damage of DNA or the induction of the natural programmed cell death known as apoptosis. In many ways, there is a growing variety of chemotherapies to potentially target each of the possible disease pathways, such as summarized in Table 2.

Chemotherapy may be used either with a curative intent, or may aim to extend lifetime or relieve symptoms, depending on the type of malignancy, the disease stage and the health of the patient. Contingent on the objective, a treating physician may tolerate more side effects if the treatment is curative whereas a trade-off between benefits of survival or tumor-burden alleviation and the side-effects must be considered if the treatment is palliative.
It is also often used in conjunction with other cancer treatments to enhance the likelihood of success of a radiation therapy or a surgery, before or after as a neo-adjuvant or adjuvant therapy. This complementarity has been demonstrated to reduce relapse and death in osteosarcoma\textsuperscript{86}, breast\textsuperscript{87} and colon cancer\textsuperscript{88}. Naturally, chemotherapy may be used as well with several other chemotherapeutic agents at once; in this case it is referred to as combination chemotherapy. Typical modes of administration are intravenous (IV), intra-arterial or intraperitoneal injection, or topical or oral forms.

Table 2. Different types of chemotherapy drugs

<table>
<thead>
<tr>
<th>Category</th>
<th>Mechanism of action</th>
<th>Example drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alkylating antineoplastic agent (including platinum)</td>
<td>damage DNA &amp; not phase-specific</td>
<td>Cyclophosphamide, busulfan, dacarbazine, cisplatin, streptozocin, mechlorethamine</td>
</tr>
<tr>
<td>Anti-metabolite</td>
<td>impede DNA and RNA synthesis by substitution, block the enzymes required for DNA synthesis during only the S phase</td>
<td>5-fluorouracil, Capecitabine, Gemcitabine, Methotrexate, fluorouracil</td>
</tr>
<tr>
<td>Anti-microtubule agent</td>
<td>block cell division by preventing microtubule function affect blood vessel growth cycle-dependent (M-, G2-M, S-phase)</td>
<td>Paclitaxel, Docetaxel, Ixabepilone, Vinblastine, Etoprebase, Estramustine</td>
</tr>
<tr>
<td>Topoisomerase inhibitor</td>
<td>interfere with topoisomerases, which usually help separating the strands of DNA</td>
<td>Topotecan, Irinotecan, Etoposide, Teniposide, Novobiocin, Merbarone</td>
</tr>
<tr>
<td>Cytotoxic antibiotics (Anthracyclines)</td>
<td>DNA intercalation, generation of highly reactive free radicals and topoisomerase inhibition not phase-specific</td>
<td>Daunorubicin, Doxorubicin, Epirubicin, Bleomycin, Mitoxantrone</td>
</tr>
</tbody>
</table>

Classical cytotoxic anticancer agents act by non-specifically killing rapidly dividing cells. Therefore, the sensitivity of tumors with high growth rates, e.g. acute myelogenous leukemia and Hodgkin's disease, is higher to chemotherapy, whereas malignancies with slower growth rates will tend to respond much more modestly\textsuperscript{89}. If the tumor is heterogenic, varying sensitivities to chemotherapy agents might then occur. This also means that chemotherapy harms healthy somatic cells: cells in the bone marrow, digestive tract, hair follicles, etc. This results in the most common side-effects of chemotherapy: myelosuppression (decreased production of blood cells, hence also immunosuppression), mucositis (inflammation of the lining of the digestive tract), and alopecia (hair loss) among others.
1.6.2. Hematological-induced anticancer treatment

Adverse Events (AE) may occur as a collateral effect in chemotherapies. In clinical trial, AEs are clearly demarcated between adverse events and serious adverse events (SAE). A SAE is any undesirable experience associated with the use of a medical product in a patient which should be reported to the FDA and in the drug labelling when the patient outcome is either death, life-threatening risk, hospitalization, disability or permanent damage, congenital anomaly or birth defect prior to conception or during pregnancy. Unlike SAE, AEs are liable to the drug levels at the target organs, so they may be decreased by the variations of drug levels in the organism with respect to time. Additionally, AEs may also be caused by drug interaction: one agent may interact agonistically or antagonistically with the other one. Again, the same logic follows with the drug removal or with newly re-evaluated strategies for the drug’s administration. In oncology, the severity of side effects has a standardized classification of five different grades according to the Common Terminology Criteria for Adverse Events (CTCAE).

Treatment with chemotherapies results in a large range of toxicities that can occur any time after administration, within days, weeks to years, or chronically. Myelosuppression or bone marrow suppression is the leading Dose Limiting Toxicity (DLT) in most cytotoxic treatments and is defined as the temporarily suppression of the hematopoietic system with decreased production of cells responsible for providing immunity (leukocytes), carrying oxygen (erythrocytes), and responsible for normal blood clotting (thrombocytes). Anemia and thrombocytopenia are mostly palliated with blood transfusion; however neutropenia defined as a low Absolute Neutrophil Counts (ANC) in blood below $0.5 \times 10^{9}/L$ (Grade 4) can be associated with higher risks for life-threatening infections. Since cytotoxic drugs weaken the immune system by lowering the first-line defense cells against infections in the body, the patient is susceptible to develop infections, notably Febrile Neutropenia (FN) which is described as a fever coexisting with neutropenia grade 4 with body temperature $> 38.5 \, ^{\circ}C$ according to the Infectious Disease Society America (IDSA) and the European Organization for Research and Treatment of Cancer (EORTC). This life-threatening condition, once occurring, may require hospitalization and treatment with antibiotics; otherwise, the risk of developing FN can be reduced by prophylactic treatment with synthetic G-CSF (Granulocyte-Colony-Stimulating Factor, e.g., lenograstim, filgrastim). To allow recovery of the bone marrow in patients that develop grade 4 neutropenia, the next dose is often reduced systematically by 25% or 50% or delayed until the return of an ANC baseline value. However in most cases, the treatment reduction is rarely tailored to the patient because of a lack of individual drug exposure and toxicity rela-
tionship. Furthermore, this same reduction in dose intensity (mg/m\(^2\) per week) may not be quantitatively evaluated as a possible hindrance of the long-term individual clinical outcome.

Several studies have reported a correlation between the maintenance of high dose intensity and improved overall survival\(^{98,99}\), while other clinical studies of breast cancer and non-small cell lung cancer\(^{100,101}\) have demonstrated that optimal dosing schedules which maintain most of neutropenia in grade 2 and 3 have a better potential to increase survival rate, since an appropriate toxicity profile prevents from toxicity-related deaths. It is therefore of a particular value to have models with the ability to explain and predict the dynamics of neutrophil counts based on both variabilities of the population and the drug itself. Even more, the use of these models can be served in the optimization of dose size and dosing regimen, such as explored in Paper IV.

**Figure 10.** Schematic illustration of the hematopoiesis.

### 1.6.3. Optimization of cancer treatment

Dosing strategies for chemotherapeutic drugs are complex. A fundamental issue particularly is the difficulty of finding the best dosing schedule within a very narrow therapeutic window, in which too low dosage may compromise the effectiveness of the treatment against the tumor, whereas excessive doses may become intolerable to the patient due to the side-effects. In oncology through, it is commonly held that a higher dose is more likely to be suc-
cessful in tumor treatment than a lower dose. As a direct consequence of this paradigm, severe side-effects are often observed and clinical doses are rather adjusted empirically determined by the risk of severe toxicity than persistently optimal. Besides, efficacy of treatment is usually difficult to assess in tumors since they are performed sparsely due to invasive and/or expensive techniques, such as biopsies, CT-scan, MRI and PET-scan etc.

The choice of dosing a patient population is also problematic due to both between–subject variability in the response profiles and the heterogenetic profile of tumors. A standard method of computing chemotherapy dosage is based on calculated body surface area (BSA), which is based on the height and weight of the individual. The underlying assumption behind such metric is that drug exposure derived from a dose determined by BSA is identical in individuals with equal proportion of height and weight. However, many physiological and drug-related factors such as drug absorption, metabolism, clearance, organ function, drug-to-drug interactions, disease state, covariates (age, gender), genetics, etc. may influence the drug uptake, hence its actual concentration in the patient's bloodstream. As a result, a high variability in the systemic drug concentration among patients dosed by BSA are often observed, leading to over- or under-dosing patients and not achieving optimal treatment effects.

In a trial design perspective, these complications mentioned above often lead to conduct trials that will usher inaccurate doses and dose schedules though drug development phases. More importantly, they are poorly designed to fit individual needs and hence are unethical and underpowered. Naturally, such designs often generate unnecessary expenditures and ultimately potential drug development failures.

Many chemotherapies ‘effectiveness shows schedule dependency. In other words, the response or toxicity is liable to the dosing intensity. Mechanisms which possibly cause this dependence may include circadian rhythms, cell-cycle specificity, drug resistance development over time, time-dependent cellular repair mechanisms, saturate systems, co-factor depletion\textsuperscript{102,103}, etc. Another reason may be partly explained by the fact that for anticancer drugs to be effective, the tumor needs to be killed and maintained undetectable during each treatment cycle. However, if not completely eradicated, tumor cells induce the proliferation rate in the remaining tumor to regrow between treatment cycles. Thus, this same tumor will become more sensitive to a subsequent treatment cycle since they are more rapidly dividing. This expectation may however be conflicted between excessive acute or cumulative exposure-dependent toxicities if dosing intensity is too high, or a tumor regrowth during time delay between two administrations due to toxicity. The choice of administration schedule is therefore important, since changing the first dose (\textit{a priori}) and carrying forward the information obtained from this first cycle to influence subsequent doses (\textit{a posteriori}) may improve rapidly the patient’s outcome by treating him individually and
at the right target exposure. In fact, drastic changes in dosing regimen may result in greater effect than increasing the dose on the outcome.

The use of PKPD models would allow more efficient characterization of schedule dependency by establishing optimal target exposure for treatment effectiveness while monitoring toxicity levels described by the PD models. Dosing can therefore be personalized to achieve target exposure and optimal results for each patient. Such protocol, referred to as TDM under certain conditions, has already been applied in clinical trials to implement individualized dosing such as for Carboplatin\textsuperscript{104} or 5-FU\textsuperscript{105}.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{timeline.png}
\caption{Timeline of principal dose-escalation methodologies for phase I oncology trials. In 2014, over 90\% of dose-escalation trials are still conducted with the “3+3” design, introduced by Storer et al. in 1989.}
\end{figure}

1.6.4. Phase I study designs

Phase I clinical trials are a critical step in the development of anticancer drugs. The main goal of these studies is to estimate the maximum tolerated dose (MTD) and to establish the recommended Phase 2 dose (RP2D) and/or schedule of new drugs or drug combinations for subsequent phase trials. Because the early clinical development of a novel agent may condition its ultimate fate, a careful attention is brought notably to the determination of the MTD. This is extensively explained by one inherent characteristic of oncology trials: the RP2D defined as the dose equaling the MTD in the USA or as the dose below the MTD in Europe and Japan, is mostly the only dose (or one additional dose at the vicinity of the MTD) carried forward to further drug development phases. Two main reasons have sustained this standard
conduct: a global paradigm which expects treatment efficacy to be detected around 80-120% of the MTD and an ethical commitment to allocate most patients to adjacent levels of the MTD to provide them prospective therapeutic benefits. As a result, a restricted dose range is merely explored around the MTD, rendering its accuracy and precision in estimation as primordial. In 2013, a retrospective study conducted by Jardim et al. confirmed the impact on drug approval underpinning such approach: 73% of cytotoxic drugs were approved by the FDA at a final dose established within 20% of their RP2D, with 82% of these doses being exactly their RP2D.

Although the determination of the MTD is of major importance in oncology trials, 95% of dose escalation trials merely use a common rule-based algorithm, namely traditional “3+3” design despite the new methodologies developed in the past decades as illustrated in Figure 12. This approach consists in adaptively allocating small cohorts of patients based on fixed and easily implementable rules that rely exclusively on observed toxicity data from the most recent subset of recruited patients. It mainly proceeds with cohorts of 3 patients treated at increasing dose levels either pre-specified in the protocol or following a modified Fibonacci sequence (i.e. the dose increment is decreased while the dose level increases). The procedure displayed in Figure 13, is as followed: The first cohort of 3 patients is treated at the starting dose. If none of them experiences a DLT, then the next cohort of 3 patients will be treated at the higher dose level. By contrast, if one DLT is observed in the 3 patients, 3 additional patients will be treated at the same dose level. The dose escalation continues until at least 2/3 or 2/6 of the patients experience a DLT (i.e. ≥ 33% of patients) at a dose level that will be defined as the MTD.

As previously seen, the “3+3” design does not required any advanced statistics in the monitoring of the trials and a majority of cytotoxic agents has benchmarked this methodology over the last three decades. Therefore, the operating characteristics of the “3+3” design are appealing for its historical presence and the simplicity of its implementation. Nonetheless, it has been extensively demonstrated that the MTD determination is mainly biased downward (i.e. under-predict the real MTD) and imprecise, and the dose escalation scheme tends to be unethical by treating a high percentage of patients at doses outside of the therapeutic range. Given such impediments, the likelihood of most oncology trials to success nowadays is largely hindered by their study design.
To overcome these limitations, several alternative methodologies (Figure 13) have been proposed using statistical model- and interval-based approaches, such as the Continual Reassessment Method (CRM) and its further ad hoc adaptations\textsuperscript{109–111}, the Escalation With Overdose Control (EWOC)\textsuperscript{112–114}, the Bayesian Logistic Regression Model (BLRM)\textsuperscript{115} or the modified Toxicity Probability Intervals (mTPI)\textsuperscript{116,117}. These innovative approaches guide dose escalation based on a dose-toxicity model that is updated with incoming trial data, from which the dose recommendation for the next cohort is derived. Generally, these new methods have been demonstrated to be superior over the rudimentary “3+3” design in simulation and theoretical work\textsuperscript{118,119}. Nonetheless, the uptake of these methods is fairly low since many assumptions must be made on the model and the distribution to describe the uncertainty around the dose-toxicity curve, which further exacerbates the “black box” perception of model-based inferences, already present in the scientific community.

In this context, an interesting shift toward an explicit and quantitative monitoring of the full time-profile of major toxicities has been proposed using NLME models\textsuperscript{120}. Unlike previously used models describing the drug effect as a binary response, i.e. toxicity or efficacy, the toxicity profile of a drug can be characterized by linking the dose–exposure–and toxicity relationships in both mechanistic and quantitative basis and such models have notably demonstrated their superiority during the last decade in clinical trials decision-making\textsuperscript{28–31}. For hematological toxicity, a previously published semi-mechanistic model developed for myelosuppression has been investigated successfully in various drug therapies and show to be a good predictive tool in terms of neutropenia outcome\textsuperscript{121–124}. This same model was ex-
tended and used for a combination therapy in Paper V to derive a model-based MTD based on predicted individual DLT rates for thrombocytopenic AE.

**Phase I dose-escalation strategies for combination treatment**

The cellular heterogeneity of tumor often leads to the emergence of drug resistance. In this case, combination therapy may be an efficient strategy to target different cell susceptibilities or reduce the likelihood of this resistance.

Although individual PK and PD are usually known for single agent and hence an estimate of the individual MTD, when it comes to combination therapy, unexpected acute and delayed toxicity may occur, which makes combination dose-escalation studies nontrivial. In the case of non-overlapping toxicities, it is often presumed that the combined MTD will be the MTD of each agent. However, if the premise is such that toxicities from each agent are overlapping, three possible hypotheses can be generated: (i) the side-effects can be either additive if there is no interaction at the PK or at the PD level (ii) the side-effects of one agent are increased due to the PK interaction between the two agents, i.e. increase drug exposure, and (iii) the side-effects of both agents are increased due to a PD interaction, i.e. synergistic toxicity. There is currently no “gold standard” methodology on how to treat these 3 scenarios. However, it is clear that the determination of one MTD single value in single-agent treatment must be replaced by the natural exploration of an infinite number of MTDs derived from the number of possible drug combinations. This range of MTD is often referred to as the RP2D contour, or the envelope of tolerability.

Four strategies are usually employed in dose-escalation studies: 

a. Alternate dose escalation  
b. Simultaneous dose escalation  
c. Single-agent dose escalation (agent A is fixed to its MTD or a dose nearby the MTD with more prior information, while agent B is escalated)  
d. Compromised dose escalation with only one of the two agents achieving full dose escalation. (adaptive design)

The choice between these four strategies is often at the discretion of the investigators and is dependable on the prior information obtained on the degree of drug interaction. Nonetheless, when it comes to dose-escalation decision-making, the most currently used methodology for combination treatment is the “3+3+3” design. Clinical data used in Paper V were originated from such trial design.
2. Aims

The general aim of this thesis was to propose and evaluate methods for handling critical aspects of clinical trial designs in the context of nonlinear mixed-effects models and optimal design of experiments. The goal was to offer an improvement over traditional methodologies in decision-making, censored data and adverse events management and in dose selection.

The specific aims were divided into three objectives:

I. **Statistical hypothesis testing in drug development:** The first goal was to simplify the power calculation of model-based clinical trial simulations by the development of a more time-efficient new power calculation method based on the difference of individual objective function values. The second step was to demonstrate the potential benefits of pharmacometric model-based inference over the classical pairwise comparison in interim analyses.

II. **Anticipation of prospective analysis and clinical constraints:** The second objective focused on the pre-incorporation of possible analysis issues and patient’s safety constrictions at the trial planning stage. Two particular interests were here the evaluation and development of novel optimization criteria for sampling schedules to handle below limit of quantification data, and for dosing regimens that reduce severe toxicity events for hematological toxicity-induced agents. Both new functionalities were implemented using utility-based functions and were assimilated in an optimal design program.

III. **Application of a model-based approach in dose-escalation oncology trials:** The third aim was to explore traditional and Bayesian methodologies in Phase I dose-escalation trials and to propose a model-based approach for the assessment of the true maximum tolerated dose. A quantitative dose-exposure-response model for the toxicity of a combination therapy was developed for this purpose.
3. Methods

3.1. Analyzed data

A large part of this work was based on simulated data using a variety of known pharmacometric models. This has allowed for the investigation of new methodology performance (*Paper I, II, III, IV, V*) and operating characteristics (*Paper V*), model selection and evaluation (*Paper V*), precision of parameter/metric comparison (*Paper III & IV*), trial design comparison (*Paper III, IV & V*) and the impact of analysis strategies (*paper II*). In only *Paper V*, real clinical data was used to develop a thrombocytopenia model for a combination therapy in solid tumors. The following sections describe the study drugs and the source and characteristics of the clinical datasets used in this paper for modeling.

3.1.1. Study drugs

The drugs used in *Paper V* is a combination treatment of abexinostat (S78454, PCI-24781) an orally pan-Histone Deacetylase inhibitor (HDACi) currently in Phase I and II clinical trials for treatment of solid tumors and lymphoma, combined with doxorubicin, a topoisomerase II inhibitor and conventional cytotoxic agent from the anthracyclines family. Two different formulations were tested: a free doxorubicin and a Pegylated Liposomal Doxorubicin (PLD) form.

**Doxorubicin**

Doxorubicin also termed Adriamycin PFS®, Adriamycin RDF®, or Rubex®, is mostly employed for chemotherapy to treat leukemias, breast cancer, soft tissue sarcomas, and aggressive lymphomas, e.g. Hodgkin's lymphomas. The major toxicity associated with this agent is a cumulative exposure-dependent cardio-toxicity due to the interference with the pumping action of the heart and the production of free oxygen radicals damaging the heart cell membranes. Due to this cardiomyopathy often leading to a congestive heart failure, the current dosing regimen is restricted to a maximum lifetime cumulative dose, which in this case is 550 mg/m² of doxorubicin. Commonly used doxorubicin combination regimens are AC (Adriamycin, cyclophosphamide), TAC (Taxotere, AC), ABVD (Adriamycin, bleomycin,
vinblastine, dacarbazine), CHOP (cyclophosphamide, hydroxydaunorubicin, vincristine, prednisone) and FAC (5-fluorouracil, Adriamycin, cyclophosphamide).

**Pegylated liposomal doxorubicin**

PLD, also named Doxil® or Caelyx®, is an encapsulated form of doxorubicin, which protective coating contains surface-grafted segments of the hydrophilic polymer methoxy polyethylene glycol that protect the liposomes from detection by the mononuclear phagocyte system. Additionally, this formulation also reduces interactions between the lipid bilayer membrane and the plasma components. For both reasons, PLD has a higher increase in blood circulation time and is less cardiotoxic than free doxorubicin. The average diameter of PLD is approximately 100 nm, which allows them to better accumulate in defective and much more convoluted tumor vessels. PLD is mainly indicated in metastatic breast cancer, advanced epithelial ovarian cancer and AIDS-related Kaposi sarcoma as a monotherapy. PLD safety profile is associated with a reversible dose and schedule dependent muco-cutaneous toxicity (stomatitis and hand-foot syndrome), which can be reduced by a less intense dosing regimen,

The postulated mechanism of action of the abexinostat-doxorubicin combination is sequential such as displayed in Figure 13. Hyper acetylation of histones and tubulins is first carried out by abexinostat, leading to loosened chromatins. This in return induces three possible pathways: (i) an increase of gene expression involved in tumor suppression, cell cycle and apoptosis, (ii) an increase of doxorubicin binding to their DNA substrate, acting as a topoisomerase II inhibitor and (iii) intercalations from doxorubicin between base pairs in the DNA helix, resulting in the inhibition of chain elongation and ultimately leading to DNA double-strand breaks, hence cell death.

Regarding the safety profile, the combination treatment was conducted in dose escalation studies in which escalation of abexinostat to the recommended dose for phase II trials was performed while keeping doxorubicin at a fixed clinically standardized dose. This dose escalation strategy allowed characterizing thrombocytopenia as the major DLT of the combination. This is consistent with the results obtained in phase I dose escalation trials of abexinostat as a single agent. Additionally, recent publications have suggested that HDACi-induced thrombocytopenia was related to a decreased ability of platelet release from megakaryocytes transiently delayed in maturation.
Figure 13. Illustration of the sequential mechanisms of action achieved by the combination between abexinostat and doxorubicin. Histone acetylation of lysine (K) residues on the histone tails are regulated by acetylation and deacetylation mechanisms catalyzed by histone acetyltransferases (HATs) and histone deacetylases (HDACs), respectively. Abexinostat inhibits HDACs by which results in the net increase in histone acetylation levels (Ac) and a relaxed chromatin conformation that is capable of activating transcription, inducing cell death mediation. Alternatively, doxorubicin creates DNA double-strand breaks by intercalation.

3.1.2. Study designs and clinical data

Study designs
Data were collected from two multicenter, non-randomized, non-comparative, open-label, dose escalation phase I studies, conducted to assess the tolerability, the pharmacokinetics and early pharmacological effects on tumor responses of abexinostat. Abexinostat was studied in combination with doxorubicin in patients with advanced refractory solid tumors (study 1) and in combination with PLD in patients with primary partially platinum-sensitive and platinum resistant malignancies in advanced ovarian cancer (study 2).

In study 1, a total of 24 patients were first enrolled in the dose escalation phase to determine the Recommended Dose 1, and 12 more patients were further included at this same dose in the cohort expansion phase. During the treatment period, patients received capsules of abexinostat orally twice a day.
(b.i.d) 4 h apart on days 1 to 4 for 3 weeks out of a 4-week cycle for a total duration of 6 cycles. 25 mg/m² of doxorubicin were infused over 15 minutes on day 3 for the first 3 weeks of the 4 week cycle and at 2 h after the second dose intake of abexinostat.

In study 2, a total of 17 patients received abexinostat orally in a similar schedule as study 1. 40 mg/m² of PLD was administered as a 1h-IV infusion to the patients 2 h after the second dose intake of abexinostat on day 3 of each 4-week cycle, for a total duration of 6 cycles.

Both dose escalation studies followed a conventional “3+3” design without dose de-escalation (see section 1.6.4 in Introduction). The explored dose range for abexinostat varied between 30 to 75 mg/m² b.i.d by increment of 15 mg/m² (4 dose levels). DLTs were assessed only during the first cycle according to the CTCAE version 3.0 December 12, 2003.

**PK and PD data**

In study 1, abexinostat and doxorubicin concentrations were collected during week 1 (day 1-4) of cycle 1 and day 1-2 of cycle 2. A total of 5 PK samples were taken for abexinostat at 0.5, 1, 2, 3, 4 h after each dose intake on day 1 and 3, and one pre-dose PK sample on day 2 and 4. The pharmacokinetics of doxorubicin were assessed on day 3 at pre-infusion, 0, 5, 10, 30 min and 1, 2, 4, 7, 24 and 48h after the end of the infusion. Platelet count was performed pre-dose, then on day 1, 3, 8, 10, 15 and 17 in all cycles.

In study 2, abexinostat concentrations were collected during week 1 (day 1-4) of cycle 1 and PLD concentrations during the whole first cycle. A total of 1 pre-dose and 5 PK samples were taken for abexinostat at 0.5, 1, 2, 3, 4 h after the first dose intake and 4 PK samples at 0.5, 1, 2 and 4 h after the second dose intake on day 1 and 3, and one pre-dose PK sample on day 2 and 4. The pharmacokinetics of PLD were assessed on day 3 at pre-infusion, 0, 0.5, 1, 2, 3, 8, 24 and 48h after the end of the infusion, and trough values on day 8, 15 and 22. PD platelet count measurements were performed pre-dose and on days 1, 3, 5, 8, 12, 15, 19 and 22 in cycle 1, and on days 1, 3, 8, 15 and 22 in cycle 2 and 3.

**3.2. Model implementation**

The following sections present the different pharmacometric models used in this thesis and the followed model building procedure to analyze the clinical data in *Paper V.*
3.2.1. PKPD benchmark models

Generic PKPD models were used to evaluate novel methodologies in the thesis, ranging from simple to literature mechanistic models from different disease areas.

**Paper I**

PK/PD datasets were simulated from different models and study designs. Two basic pharmacokinetic models were used as proof-of-concept to validate the novel power methodology. The first example involves a one-compartment IV bolus model with first order elimination, with typical clearance (CL) and volume (V) values being 10L/h and 100L, respectively. The second example involves a one-compartment model with first order elimination and zero-order absorption where steady state conditions were assumed, with typical CL and dosing rate values being 10L/h and 1mg/h, respectively.

The implemented new method was also tested on four distinctly different PK/PD models of varying complexity: a linear disease model with a drug effect on the slope and lognormal distributed IIVs on baseline, slope and effect parameters, a nonlinear mixed-effects model in Type 2 Diabetes Mellitus describing the mechanistic relationship between tesaglitazar exposure, fasting plasma glucose (FPG), glycosylated hemoglobin (HbA1c) and aging red blood cell (RBC) with drug effect added on the rate of elimination ($K_{\text{out}}$) of FPG, a nonlinear mixed-effects model describing the decrease of viral load in HIV-infected patients after initiation of antiretroviral treatment and a nonlinear mixed-effects base model with no original covariate inclusion, describing the relationship between the plasma concentration of digoxin, the estimated concentration at the effect site and the reduction in heart rate during atrial fibrillation with a drug effect linearly added on the heart rate baseline value.

**Paper II**

Two complex models were utilized in POC and DF trials to derive the respective sample size needed. The second model is identical as the FPG-HbA1c-RBC model previously described in **Paper I**.

*Stroke model*

A nonlinear mixed-effects model has previously been developed for stroke disease progression after an acute ischemic stroke, assessed by the 42 point NIH stroke scale (NIHSS). The model consists of three sub models for conditional probabilities reflecting the likelihood of disease improvement or deterioration, reaching complete recovery (i.e. NIHSS=0 as in no neurological disability) and dropout of the study, in combination with two linear sub models for of the relative magnitude of improvement or deterioration (visualized in Figure 14). The model also includes covariates such as age and
baseline NIHSS score. In this simulation study the drug effect was added linearly on the magnitude of improvement (i.e. relative score change given an improvement in disease state).

Figure 14. A schematic illustration of the concept of the National Institutes of Health Stroke Scale model, in which a zero score represent complete recovery and a score of 42 represent maximum severity. S1, S2, S3, and S4 are observed scores at observations 1, 2, 3, and 4, respectively. Gray circles indicate potential scores after each type of transition (which, in reality, could be any value between the score minimum and the last observation in the event of a score decline, between the last observation and one unit below the score maximum in the event of a score improvement, or the score maximum). Bold lines indicate actual score progression, whereas gray lines represent events that were possible, but did not take place, at every transition.

Diabetes model

A nonlinear mixed-effects model in Type 2 Diabetes Mellitus has previously been developed by Hamrén et al.\textsuperscript{136} to describe the mechanistic relationship between tesaglitazar exposure, FPG, HbA1c and aging RBC. The model as shown in Figure 15 consists of three sub models including an indirect response model on the effect of drug exposure on FPG over time, a transit compartment model to describe the RBC lifespan with a zero-order release of RBC into blood circulation. The model also includes at any stage of the RBC maturation a function describing the glycosylation of RBC into HbA1c related to the FPG level. The structure of the model allows the possibility to evaluate different mechanisms for the drug effect. However, this simulation study has only investigated one plausible mechanism, which is a drug effect increasing the rate of elimination of FPG.
Figure 15. Schematic representation of the mechanism-based model for the FPG–HbA1c relationship, in which plasma concentration \( C_p \) vs. FPG effect is described by a sigmoidal \( E_{\text{maxFPG}} \) function including \( E_{\text{maxFPG}} \), the maximum effect on \( K_{\text{outFPG}} \) the first-order degradation rate constant of FPG and \( \text{EC}_{50FPG} \), the plasma concentration achieving half-maximal effect on \( E_{\text{maxFPG}} \); red blood cell (RBC) maturation is described by \( K_{\text{inRBC}} \), a zero-order rate constant of RBC release in blood circulation and \( K_{\text{t}} \), a first-order transit rate constant between each maturation stage; FPG mechanism is described by zero-order production rate constant of FPG and a glycosylation rate \( K_{\text{glucose}} \) from RBCs to HbA1c. FPG, fasting plasma glucose; HbA1c, glycosylated hemoglobin. Reprinted with permission from Macmillan Publishers, ref. 21 copyright 2008

**Paper III**

Two trivial models were used for design evaluation and design optimization using the 7 methods developed for handling BQL data in OD. A one-compartment model with fixed effect parameters of CL and V of 0.693 and 1 was used for design evaluation such as described in Byon et al.\(^67\). For design optimization, a two-compartment model was used, the biphasic elimination was described by the CL and the inter-compartmental distribution set to 10L/h and 100L/h, respectively, and the central and peripheral compartment volumes of distribution set to 100L and 80L, respectively.

**Paper IV**

The semi-mechanistic model by Quartino et al.\(^123\) was used as an example model to apply dose size and dosing regimen optimization methodologies.
The anti-mitotic chemotherapy docetaxel (Taxotere®) was selected to represent a typical anticancer drug with substantial dose-limiting neutropenia.

The model was described by one compartment representing proliferative cells such as the stem and proliferating precursor cells, linked to three transit compartments describing the non-mitotic maturation of the neutrophils in the bone marrow, before migrating into the blood compartment and emerging as the circulating observed neutrophils. A feedback function acting as the moderation of G-CSF on the proliferation rate when circulating neutrophil count is below the baseline value was also incorporated in. As the model encompasses mechanistic over physiological features, several system-specific parameters were introduced such as the mean transit time (MTT), the baseline levels of neutrophils before drug administration (ANC₀), a feedback parameter GAMMA (γ) describing the amplitude of proliferation simulation dependent on the current circulating neutrophil count, and the half-life of circulating neutrophils (K₉). The drug effect relating to plasma concentration levels was generally found to affect the proliferation pool by inhibiting the proliferation rate or inducing cell loss. In this study, the drug effect was modeled as a linear function (SLOPE). On a population level, this model accounts for IIV which is incorporated on the system parameters MTT, ANC₀ and drug-specific parameter SLOPE by assuming they are log-normal distributed, but not for any interoccasion variability (IOV). Residual error was described using an additive error model.

To describe the administration and disposition of docetaxel, concentration-time profiles were predicted using the PK model developed by Bruno et al. A schematic representation of the full PK/PD model is presented in Figure 16.

3.2.2. Thrombocytopenia model building strategy

The general modeling workflow in Paper V consists in using both abexinostat and doxorubicin PK and PD data such as described in section 3.1.2 to develop a semi-physiological model describing the pharmacokinetics of abexinostat, doxorubicin and PLD as well as the resulting pharmacodynamics of the platelet count time course. The sequential Individual PK Parameters (IPP) modeling approach proposed by Zhang et al. was performed using a previous semi-physiological PKPD models established with phase I trial data of abexinostat as single therapy and literature models for the other therapeutic agents. The PK for abexinostat described as a three-compartment distribution model with first-order absorption and first-order elimination was combined with a PK model of doxorubicin, a PK model of PLD and a PK model describing the release of the PLD formulation into the active doxorubicin form to perform an external validation on the PK data. The absence of a PK interaction between abexinostat and doxorubicin or PLD was also assessed by external validation. Those models were
then used to derive the Empirical Bayes Estimates (EBE) necessary to compute the corresponding concentration-time profiles.

The next step is a joint modeling of platelet count data from both studies. A semi-physiological model, structurally similar to the neutropenia PKPD model described in Paper IV is adapted to the thrombopoiesis. This modified model consists of one compartment representing the proliferating cells in the bone marrow, three transit compartments mimicking the maturation of the non-proliferating CFU-GEMM precursors into CFU-Meg and megakaryocytes and one compartment representing the platelets in the blood circulation. The regulation of the thrombopoiesis by endogenous growth factors was described by a feedback function between circulating platelet count and the proliferation rate.

Different drug effect models (linear, Imax, sigmoid Imax, power) were tested for the effects of abexinostat and doxorubicin on the progenitor cell proliferation. An interaction term was also investigated to determine if the two agents could be acting additively or synergistically.

A full variance-covariance matrix was investigated for correlations between random effects. Several residual error models were investigated and log-transformation of the data was tested to verify the normality assumption in the residual errors.

### 3.2.3. Data analysis and model evaluation

The data analysis was carried out in Paper V by nonlinear mixed-effects modeling using the software NONMEM 7.2 using the subroutine ADVAN13 and the FOCEI estimation method. Throughout the model building process, selection between models was based on the fit of the data, precision of parameter estimates, goodness-of-fit plots and visual predictive checks (VPCs).

Discrimination between two nested models was based on the drop in the OFV computed in the LRT. Generally, a decrease in OFV of 3.84 for one additional parameter was considered significant for a 5% type I error rate. Non-nested models were compared using the Bayesian Information Criterion (BIC). The model with the lowest BIC value was selected. Precision in parameter estimates was evaluated using the relative standard error outputted from the SCOV option in NONMEM. Graphical evaluation was performed using standard goodness-of-fit plots generated by the Xpose 4.4 R package (http://xpose.sourceforge.net/).

Two simulation-based diagnostics were employed using PsN toolkit 4.2.0 (http://psn.sourceforge.net/). A prediction-corrected Visual Predicted Check (pcVPC) was used to evaluate the predictive performance of the model while a Posterior Predictive Check (PPC) was used to assess the prediction abilities of the nadir value at cycle 1. For pcVPC, one thousand datasets were simulated using the final model and the original dataset struc-
ture. The median, 10th and 90th percentiles were computed for the observed data and for the simulated data sets. For each bin across the independent variable, both observed and simulated dependent variables were normalized for the typical population predictions. The non-parametric 95% confidence intervals for the median, 10th and 90th percentiles were computed based on the simulated datasets and superimposed on the observed data for visual inspection. For PPC, a mean nadir value at cycle 1 was derived from observed data for each dose level. One thousand datasets were simulated using the final model and the original dataset structure. For each dose level, the distribution of all nadir values from all replicated datasets was displayed as the reference distribution of the corresponding observed mean nadir value.

Figure 16. PKPD model for docetaxel with linear drug effect \( E_{\text{drug}} \). The PK model is a three-compartment model with a first-order of elimination. The PD model consists of one compartment representing the proliferating cell pool, three transit compartments with maturing cells and one compartment of circulating observed neutrophils. MTT is the mean transit time though the chain, \( k_{tr} \) proliferation rate constant, \( k_{circ} \) elimination rate constant for circulating neutrophils, \( E_{\text{drug}} \) drug effect and Feedback represents the feedback loop from circulating neutrophils.
3.3. Power estimation

The following sections focused on the expansion of the methodology around the statistical power. Both novel algorithm for a faster power curve calculation (Paper I) and its application in critical stages of phase II (Paper II) are the specific subjects.

3.3.1. Monte-Carlo based power

**Power assessment**

Traditional model-based power assessment is usually conducted using the SSE method for a selection of sample sizes and models (see section 3.2.1). In this methodology, each corresponding power assessment was derived separately from a given sample size. For a defined design, i.e. sampling times, sample size, dosing regimen, etc., 1,000 replicates were simulated from the full model (with drug or covariate effect, $H_1$) and both full and reduced (without drug or covariate effect, $H_0$) models were fitted to the simulated data. For each replicate, the difference in the OFV was computed and submitted to the hypothesis chi-squared $\chi^2$ test. The number of replicates where the difference in OFV indicated a significant subgroup effect was counted. The ratio of this number over the total number of replicates provides the estimated power of the study for the tested sample size $N$. The process was carried out iteratively for a range of sample sizes to cover different areas of a full power curve.

**Type I error calibration**

To correct for the difference between the actual and nominal type I error due to the deviation of the LRT from its properties at small sample sizes, a systematic type I error calibration was applied to the critical \( \Delta OFV \) obtained from the SSE: 10,000 replicates of the same design used in the SSEs were simulated from the reduced model and both the full and reduced models were fitted to the simulated data. The \( \Delta OFV \) for each replicate was calculated and ranked to determine the nominal cut-off OFV from the fifth percentile. This new empirically determined OFV cut-off is used to reassess the power for the present sample size: the percentage of \( \Delta OFV \) greater than the new cut-off OFV is taken as the power for the current sample size. The type I error corrected SSE determined power is further referred to as the calibrated power.

3.3.2. Monte-Carlo Mapped power

In Paper I, a new power calculation method—the Monte Carlo Mapped Power (MCMP) approach was developed to allow a complete mapping of the power curve without the main impediments of run-time intensity and
need for Type I error correction, as mentioned for the SSE methods in the Monte-Carlo based power. The approach adopted here attempted to extend the same simulation and hypothesis testing settings as proposed previously but to only one single step of simulation and estimation from a large simulated data set.

In NONMEM version 7.1.2, the overall objective function value (OFV) of a model for a given dataset, which is approximately proportional to minus twice the natural logarithm of the likelihood of the data, can be easily outputted as individual objective function values, such as:

\[ \text{OFV} = \sum_{i=1}^{n} \text{iOFV}_i \]  \hspace{1cm} (3.1)

where \( \text{iOFV}_i \) denotes the \( i \)th individual contribution to the overall OFV. The MCMP method as outlined in Figure 17 tests the hypothesis of a possible drug/covariate effect using the substitution of the overall OFV value by the summation of \( \text{iOFV}_i \) values in the LRT. Given a defined study design of \( n \) individuals per study group, a large simulated dataset is first computed from a model containing the tested drug/covariate effect. The generated data are then estimated with a single full and a single reduced model, providing a large pool of values for the full model denoted as \( \text{iOFV}_{\text{FULL}} \) and for the reduced model, denoted as \( \text{iOFV}_{\text{REDUCED}} \). In the LRT, the difference in the overall objective function value (\( \Delta \text{OFV} \)) between two nested models can be replaced, such as:

\[ \Delta \text{OFV} = \sum_{i=1}^{n} \Delta \text{OFV}_i = \sum_{i=1}^{n} ( \text{iOFV}_{\text{FULL}} - \text{iOFV}_{\text{REDUCED}} ) \]  \hspace{1cm} (3.2)

This sum is tested against the theoretical value obtained from the \( \chi^2 \) distribution (i.e. cut-off of 3.84 in \( \Delta \text{OFV} \) for nominal significance level of 0.05 with \( df=1 \)). To map the whole power versus sample size relationship up to a predefined sample size, this procedure is repeated 10,000 times for each sample size of the power curve by randomly sampling the sum of all \( \Delta \text{OFV}_i \) (e.g. in increments of one subject per study group). At each current design (i.e. each sample size), the power is assessed as the percentage \( \sum \text{iOFV}_i \) of out of 10,000 times that is greater than the significance level criterion defined by the LRT.

3.3.3. Proof-of-Concept and Dose Finding sample size

POC studies in Phase IIa are designed to give preliminary evidence of efficacy and safety, with the aim to inform a decision about proceeding into full development of the drug. In practice, the study size is usually determined by the primary objective of the trial and the number of subjects should be large enough to be able to detect the defined drug effect but at the same time ex-
pose a minimum number of subjects to an experimental drug. It is also common that DF studies are performed in Phase IIb to explore a dose–response relationship. In *Paper II*, the use of clinical trial simulations using MCMP was explored to reduce sample sizes in both types of studies for the two examples discussed in section 3.2.1.

Figure 17. Comparison between the Stochastic Simulation and Estimation (SSE)-based power versus Monte-Carlo Mapped power (MCMP) algorithms. ** Asterisk indicates that the cut-off value in SSEs needs calibration for Type I error, notably for small sample sizes.

In the two example applications, two scenarios were simulated and the results were analyzed using MCMP and a conventional statistical analysis using a two-sided *t*-test (*p* < 0.05) (see section 1.4.1).

a. A POC study in which a placebo arm was compared with the highest active dose group

b. A dose-ranging scenario with placebo being compared to all three active treatment arms

The scenarios simulated and analyzed are summarized in Table 3. The conventional study sizes were based on *t*-test comparing placebo and the highest dose group, and the size of the dose-ranging study was calculated under the assumption of four equal sized groups, i.e., the conventional dose-ranging study was twice the size of the conventional POC study.
Table 3. Simulation and analysis settings for Proof-of-Concept and Dose-Finding studies in two disease area examples: Stroke and Diabetes Type 2.

<table>
<thead>
<tr>
<th></th>
<th>Stroke model</th>
<th>Diabetes model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dose level</td>
<td>2.5, 3.8, and 5mg</td>
<td>0.1, 0.5, and 1 mg</td>
</tr>
<tr>
<td>Big dataset MCMP</td>
<td>2500</td>
<td>500</td>
</tr>
<tr>
<td>sample size/arm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-test characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔEffect</td>
<td>1.75 (equivalent to 55% recovery)</td>
<td>-0.63% for ΔHbA1c</td>
</tr>
<tr>
<td>SD placebo vs active</td>
<td>6.23 and 5.98</td>
<td>1.02 and 1.02</td>
</tr>
<tr>
<td>Missing data method</td>
<td>LOCF</td>
<td>N/A</td>
</tr>
</tbody>
</table>

LOCF: Last Observation Carried Forward; SD Standard Deviation; MCMP: Monte-Carlo Mapped power; HbA1c: Glycosylated Hemoglobin; ΔEffect: effect size

3.4. Optimal design methodology

OD was used in two papers of this thesis, *Paper III and IV*. The conventional application of OD consists in maximizing the information content of an upcoming study by optimally selecting study design parameters. Several novel approaches were explored in this thesis to extend this general application to handle BQL data in the planning of a trial with assay constraints and to perform dose optimizations when toxicity is the primary interest (instead of parameters precision). All models were implemented in MATLAB version R2010b (The MathWorks Inc., Natick, MA, 2010) and all optimization works were performed in the optimal experimental design software PopED version 2.13⁷⁷,⁷⁸.

3.4.1. Information-based criteria for LOQ

The premise of *Paper III* was the absence, to our knowledge, of systematic investigations considering LOQs in OD. General approaches currently ignore LOQs or use trivial imputation methods. It is therefore not surprising that experts in the pharmaceutical industry have recently ranked the development and comparison of methods handling LOQs as one of the top priorities in the field of OD⁷⁷.

This section describes seven different methods, numbered D1-D7, to calculate the FIM when LOQ data may be present. Most methods are applicable to data above and below the LOQ; however when this is not the case, a remark will be explicitly made in the description.

The general procedure for all methods, except for method D1, is to compute the FIM by weighting each experimental sampling time’s expected information content, given the probability that this sampling time will be below the LLOQ or above the ULOQ. Therefore, if a sampling time in a design is predicted to have a 100% probability of being an observation, e.g.,
above the LLOQ, the classical expression of the likelihood for normally distributed data is used to compute the FIM and the sampling time is fully weighted, conditioned by the model’s predicted response. However, if a sampling time is predicted to have an X% probability of, for instance, being a BQL data point, different approaches can be used to scale down the weight of this sampling time. These different weighting mechanisms are proposed in Figure 18.

These 7 different methods, standard and newly developed, were implemented, such as:
- **Standard D1**: Ignore LOQ.
- **Standard D2**: Non-informative Fisher information matrix (FIM) for median response below LOQ (FO) i.e. for each design point, set the contribution to the FIM to zero if it gives a median response below LOQ.
- **New D3**: Non-informative FOCE linearized FIM for individual response below LOQ i.e. for each design point, set the individual contribution to the FIM to zero if it gives an individual response below LOQ.
- **Standard D4**: Addition of a homoscedastic variance to account for the LOQ. Five different levels of additive residual error were tested in this work: 1/4, 1/3, 1/2, 1 and 2 times the LLOQ.
- **New D5**: Simulation & Rescaling i.e. for each design point, scale the FIM with the probability of BLQ predicted from simulation.

**Figure 18.** Schematic depicting the scaling principle applied to sampling times which are below the Limit of Quantification (dashed line) in methods D1-D7. "+ε" represents an additional additive error.
New D6: Integration & Rescaling i.e. for each design point, scale the FIM with the probability of BLQ calculated from the FO approximated joint density.

New D7: Calculation of FIM for each design point by integrating over simulated data with a joint likelihood for data above (normal likelihood) and below LOQ (M3 method) using the Laplace approximation.

3.4.2. Utility-based criteria for dosing regimen optimization

The aim of Paper IV was to illustrate how a population optimal design can be used to optimize phase I/II oncology trials in order to limit neutropenia. Different metrics of interest such as predictions of absolute nadir neutrophil count and clinical constraints were used in the optimization of dose size and dosing schedules, based on different objectives. The template model (described in section 3.2.1) was used for the optimization of the anti-mitotic chemotherapy docetaxel, generally administered as a standard dose of 60-100 mg/m² given as a one-hour IV infusion every three weeks. When treated with this posology, 75-85% and 6-12% of patients experience grade 4 neutropenia and febrile neutropenia, respectively\textsuperscript{154,155}.

A user-defined penalty function was implemented for each criterion of interest. Any incident of grade 4 neutropenia was defined as a DLT, however, in further design optimization, more indications were added to refine this definition, such as a duration of grade 4 \( \geq 7 \) days. For a prediction model defined as \( f(.) \) for individual \( i \) at time \( t_{ij} \) and subject-specific parameters described as the function \( g(.) \) of population \( \theta \), individual specific covariate \( a_i \) and individual specific random effect \( \eta_i \) and \( \kappa_i \) assumed to be normally distributed with a mean of 0 and a variance of \( \Omega \) and \( \Gamma \), respectively, four metrics of interest were selected for dose optimization.

To maximize the typical nadir value:

\[
\arg \max_{a_i} \left( \text{nadir } f \left( t_{ij}, g(\theta, \eta_i, x_i, a_i, \kappa_i) \right) \right)_{\eta_i=0} \quad (3.3)
\]

To maximize the profile to be within grade 2 and 3 by using the Area Under the Curve (AUC) subtraction

\[
\arg \max_{a_i} \left( \text{AUC}_{G2-G3} f \left( t_{ij}, g(\theta, \eta_i, x_i, a_i, \kappa_i) \right) \right)_{\eta_i=0} \quad (3.4)
\]
To minimize the 5<sup>th</sup> percentile of time spending in Grade 4 ($T_{G4}$) more than 7 days by assuring $\text{ANC}_i \geq \text{ANC}_0$ at end of cycle (day 21) and for individual specific random effect $\tilde{\eta}_i$

$$\arg \min_{a_i} \mathbb{E}_{\eta} \left( T_{G4} \left( f \left( t_{ij}, g(\theta, \eta_i, x_i, a_i, \kappa_i) \right) \bigg| \eta_i = \tilde{\eta}_i \right) + \text{Baseline}_{t_{ij}=21} f \left( t_{ij}, g(\theta, \eta_i, x_i, a_i, \kappa_i) \right) \bigg| \eta_i = \tilde{\eta}_i \right) > 7 \text{ days}$$ (3.5)

To maximize the dose sizes for a typical nadir $\geq 0.5 \times 10^9 / \text{L}$

$$\arg \max_{a_i} \left( \text{nadir} \ f \left( t_{ij}, g(\theta, \eta_i, x_i, a_i, \kappa_i) \right) \bigg| \eta_i = 0 \right) \geq 0.5 \times 10^9 / \text{L}$$ (3.6)

Both dose sizes and dose times were optimized using the Random Search, Stochastic Gradient and Linear Search methods for a discrete variables optimization. Constraints on dose size or dosing interval, e.g. identical doses or intervals for instance, were added to the optimization.

For criteria D, three doses were optimized without any upper limit constraint on the dose size and with a fix spread of doses gradually increasing, e.g. 1 dose every 2 days, 1 dose every 3 days etc.

### 3.4.3. Information-based criteria for sampling optimization

In the same Paper IV, the goal was also to obtain a precise and unbiased nadir value, either based on model prediction or on a sampled observation, using different optimization strategies for sampling times. Four optimization criteria were tested and detailed below.

#### D-optimality for the uncertainty of the typical population parameters (D-OPT)

$$\arg \max_{t_{ij}} \left| FIM_{pop} \left( t_{ij}, g(\theta, \eta_i, x_i, a_i, \kappa_i) \right) \bigg| \eta_i = 0 \right|$$ (3.7)

The application of D-optimality results in a sampling schedule that provides minimum variance for all equally weighted population parameter estimates in the model.
MAP-optimality for the uncertainty of the expected individual parameters (MAP-OPT)

In Maximum a Posteriori optimality (MAP-OPT), the interest is shifted to the precision of the expected individual parameters instead of the population parameters. The criterion thus corresponds on average to the minimization of the uncertainty of the Empirical Bayes Estimates (EBEs) while population parameters are used as prior information. The $FIM_{MAP}$ is expressed such as:

$$FIM_{MAP} = \frac{1}{N} \sum_{i=1}^{N} FIM_{pop} \left(t_{ij}, g(\theta_i, \eta_i, x_i, a_i, \kappa_i)\right) + \Omega^{-1} \quad (3.8)$$

with $\Omega = diag(\omega_1^2, ..., \omega_R^2)$ providing the prior population information, $\theta_i$ being both the population parameters and the transformed individual parameter estimates to fixed effects and $N$ the total number of individual parameters to be sampled from the population distribution, selected to be 100 samples in this example. For more information on the methodology, refer to Hennig et al.\textsuperscript{81}. The objective function was then to maximize the logarithm of the determinant of the $FIM_{MAP}$ such as:

$$arg \ g \ max \ \log |FIM_{MAP}| \quad (3.9)$$

C-optimality for the uncertainty of the expected nadir

In C-optimality, the criterion determines a design such that a linear combination of the unknown parameters, specified in the vector $c$, has a minimal variance.

In the context of neutropenia, the metric to minimize was the variance of the predicted nadir $var(nadir)$, which is not explicitly defined as a parameter in the model. To circumvent this issue, the relative impact of each parameter in the model to the nadir value was expressed by a vector $c$ defined as the gradient of the response function at time $t_{ij} = t_{nadir}$ with respect to the population parameters, such as:

$$c = \left\{ \frac{\partial f(\cdot)}{\partial \theta_1}, \frac{\partial f(\cdot)}{\partial \theta_2}, ..., \frac{\partial f(\cdot)}{\partial \theta_n} \right\}_{t_{ij}=t_{nadir}} \quad (3.10)$$

and $t_{nadir}$ expressed as either (i) the predicted time to nadir from the model later referred to as C-OPT FIX and (ii) the observed time to nadir for each set of sampling times currently optimized later referred to as C-OPT OBS. The C-optimality criterion is then defined such as:

$$arg \ min_{t_{ij}} \left( c^T FIM \left(t_{ij}, g(\theta_i, \eta_i, x_i, a_i, \kappa_i)\right) \right)_{\eta_i=0}^{-1} c \quad (3.11)$$
Sample Reuse Simulation (SRS) method

The SRS method\textsuperscript{156} is a simulation-based optimal design procedure for obtaining optimal sampling schedules for nonparametric estimates of interest in the model. The sampling schedule was optimized to allow direct sampling at the nadir value. The mean-squared error is used as the design criterion and consists in four steps:

a. For $N$ simulated responses from both inter-individual and residual error variability, $N$ true time to nadir are calculated.

b. Using the currently evaluated sampling times, $N$ simulated responses from either inter-individual (SRS ETA NO EPS) or both inter-individual and residual error variability (SRS ETA EPS) are used to compute $N$ estimated time to nadir.

c. Mean-squared error of each pair is approximated by a summation over the $N$ estimates, such as for $N$ number of simulations:

$$\arg \min_{t_{ij}} \left( \frac{1}{N} \sum_{i=1}^{N} (t_{ij}^{\text{true nadir}} - t_{ij}^{\text{est nadir}})^2 \right) \quad (3.12)$$

d. Find $t_{opt}$ that minimizes equation 3.12 using a Newton-type minimization procedure.

3.5. Dose-escalation algorithms

In Paper V, a selection of the most frequently applied dose-escalation methods including the “3+3” design, CRM, EWOC, BLRM and mTPI were used to determine the RP2D of the drug combination described in section 3.1.1. These methods described in the following sections, were used in a comparison framework in order to evaluate the operating characteristics of each of them (section 3.6.5).

3.5.1. Pharmacometric-based maximum tolerated dose

One way to determine the RP2D is to use simulations from the PKPD model developed using the combination therapy data. The “reference” RP2D was computed from 4 steps: (i) simulating 2000 individual PD profiles per dose level using the original sampling design protocol, (ii) computing number of DLTs per dose based on how many individual observed nadir values were found above (DLT=0) or below (DLT=1) thrombocytopenia grade 4 threshold value, (iii) deriving the MTD as the dose level leading to at least 33% of DLTs and (iv) determining the “reference” RP2D as the dose level below the MTD. The evaluation of DLT risk was assessed using only the nadir value of the response curve during the first cycle, corresponding to how DLT was determined according to the protocol.
3.5.2. Bayesian approaches

The following sections present the Bayesian approaches used in Paper V to derive the RP2Ds (not exceeding a probability of toxicity superior to 33%) and further compared to the PKPD model-derived “reference” and the clinical RP2Ds.

**The Continuous Reassessment Method (CRM)**

The model-based CRM method proposed by Quigley et al. in 1990 defines the MTD as a quantile of the dose-toxicity probabilistic model (DTPM) curve, which is iteratively updated with patient data available until every new accrual time. With every occurrence of toxicity, the DTPM’s parameter estimates are adjusted according to Bayes’ theorem and the dose with a posterior expected toxicity probability closest to a predefined MTD rate is administered to the next patient. The resulting dose escalation scheme is continued until a maximum sample size is reached. The RP2D is selected to be the dose with the posterior expected toxicity probability closest to (but without exceeding) the target MTD rate at the end of the trial.

Due to occasional aggressive dose “jumping” recommendations in the original CRM, several modifications (renamed as the “modified” CRM) have been implemented here to address those issues. First, the first cohort was always treated at the lowest pre-specified dose level; secondly, an increment step never exceeds one dose level at a time, and thirdly more than one patient was treated at each dose level.

The operating characteristics of the modified CRM were evaluated using the R package bcrm 0.4.4. and MCMC methods in OpenBUGS via the BRugs package.

**The Escalation with Overdose Control (EWOC)**

The risk of over-dosing patients is overlooked in the CRM. To overcome this limitation, an EWOC suggested by Babb et al. was implemented with an additional parameter called the feasibility bound $\alpha$. In the EWOC method, the posterior distribution of the MTD is normally calculated and the next patient is given the dose closest to the $\alpha$th quantile of the MTD distribution. This method includes thereby the ethical constraint of minimizing the chance of treating patients at unacceptably high doses.

The operating characteristics of EWOC were evaluated using the same model and prior settings as in CRM. The feasibility bound $\alpha$ was set at 0.25, i.e. the recommended value in the original method.

**The Bayesian Logistic Regression Model (BLRM)**

The BLRM method proposed by Neuenschwander et al. uses a fully-Bayesian approach which allows dose-escalation decisions to be based on the whole posterior distribution. This method proposes to classify the proba-
bility of DLT into 4 regions of the posterior distribution: Under-dosing, Targeted toxicity, Excessive toxicity and Unacceptable toxicity. For the selection of the next candidate dose, the posterior probabilities in each region are examined and a dose is recommended based on a pre-specified acceptance of risks and benefits. A common recommendation rule is to maximize the probability of toxicity within the Targeted toxicity category, while keeping the probability of excessive and unacceptable toxicity under 25%. In the implementation, the Bayes risk associated to each of this region is defined by a corresponding loss function and the dose which minimizes the overall Bayes risk is thereby selected as the next dose.

Using the same settings as in CRM and EWOC, the toxicity category intervals were specified as Under-dosing (0, 0.17], Targeted toxicity (0.17, 0.33], Excessive toxicity (0.33, 0.6] and Unacceptable toxicity (0.6, 1]. The corresponding loss function was defined such as:

\[
R(\theta, \text{dose}) = \begin{cases} 
  l_1=1 & \text{if under-dosing} \\
  l_2=0 & \text{if targeted toxicity} \\
  l_3=1 & \text{if excessive toxicity} \\
  l_4=2 & \text{if unacceptable toxicity}
\end{cases}
\]  

(3.13)

The Modified Toxicity Probability Intervals (mTPI)

Similar to the BLRM method, the mTPI framework as defined by Ji et al. relies on subsequent allocation of patients based on toxicity probability intervals computed at the current dose. An Equivalent Interval (EI) like the Target Interval in BLRM, partitions the probability interval (0, 1) into 3 sub-intervals in which toxicity probabilities are deemed to indicate under-dosing, target-dosing, and over-dosing strategies, respectively. Given toxicity observations of the currently evaluated dose, a beta prior binomial model is used to compute the posterior density function, i.e. the posterior probability for each of the subintervals. The core decision rules for the next stage dose level are based on the Unit Probability Mass (UPM), defined as the ratio of the probability of the interval and the length of the interval. The dose-escalation proceeds according to these two steps: (i) compute the UPM for the three subintervals and (ii) Escalate (E), Stay (S) or De-Escalate (D) the dose for the next cohort if the largest UPM is found in the under-dosing, the target-dosing or the over-dosing intervals, respectively. The resulting scheme is applied for each dose level selected in the dose escalation until the maximum sample size is reached.

The mTPI method was conducted using the web-based statistical tool available at http://compgenome.org/NGDF/. An EI defined as [0.17-pT] and a beta (1, 1) prior (equivalent to the uniform distribution) were used as suggested by the authors.
3.6. Methodology assessment

The MC-based simulation techniques were used several times in this thesis to evaluate novel methodologies in terms of bias and imprecision assessment of the parameter estimates and of methods robustness. Operating characteristics of these methods were also evaluated for general recommendation to the end-users. Simulations and estimations were conducted in NONMEM version 7.1.2 in conjunction with PsN 3.5.1.

3.6.1. Evaluation of Monte-Carlo Mapped power

In addition to the comparison with the SEE-derived power values, the MCMP method described in section 3.3.2 in Paper I was investigated for optimal implementation.

MCMP Dataset Size

The relation between the MCMP dataset size and imprecision in the estimated sample size needed to reach 90% power ($N_{90\%}$) was explored. Several MCMP dataset sizes ($n=250, 500, 1,000, 2,000, 4,000, 8,000$ and $10,000$) were investigated in the MCMP simulation step by simulating 1,000 replicates. Each replicate was then estimated under the full and reduced models. For each MCMP dataset size, 1,000 MCMP curves were obtained and used to compute the relative standard error, the mean and the standard deviation in $N_{90\%}$.

Number of SSEs for Equivalent Relative Standard Error in MCMP Power Prediction

For each dataset size described in the previous section, a Relative Standard Error (RSE) in power predicted by MCMP is calculated based on the 95% confidence interval derived from the same 1,000 MCMP curves simulated from the previous infusion model. The number of SSE ($n_{SSE}$) replicates for equivalent RSE was computed from the following relationship for the power of interest of $\hat{p} \pm \text{SE}$:

$$\hat{p} \pm Z_{1-\frac{a}{2}} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n_{SSE}}}$$

(3.14)

Impact of $\eta$-Shrinkage on MCMP Power Prediction

To evaluate the impact of $\eta$-shrinkage, the MCMP method was run with a sparser number of samples per subject (i.e. 2 versus 4 samples per subject) and a residual error increased up to 30%. The total sample size resulting in 90% power from the MCMP method was selected for power assessment using a calibrated SSE. Power predictions from both methods were then compared.
3.6.2. Comparison of optimized LOQ sampling schedules

The evaluation and comparison of methods D1-D7 in Paper III was performed in two separate steps, each step was performed for several different LLOQ levels:

A. Comparison between methods for observed versus expected parameter precision and runtime in a base design using the one-compartment model.

B. Performance of the optimized designs using the fastest methods based on results of step A in the two-compartment model. The optimized sampling schedules and comparison of methods were assessed for bias, imprecision, robustness and predictability.

The explored LLOQs for A were 0.0625, 0.0884, 0.125, 0.1768 and 0.25 and 0.05, 0.10, 0.15 and 0.20 for B.

**Expected precision**

Expected parameter precision on the standard deviation scale was calculated from the inverse of the FIM, obtained with each method (D1-D7) and for all 5 LLOQ levels. The expected precisions were compared to the corresponding empirical parameter precisions obtained in an SSE study. In addition to the comparison of each parameter’s precision, the matrix determinant was used as a summary metric. For each method and each level of LLOQ, the determinant of the FIM was compared to the determinant of the empirical FIM. The latter was calculated by calculating the inverse variance-covariance matrix of the parameter estimates from the simulated and re-estimated datasets.

**Runtimes**

For each method, the runtime for one base design evaluation of the one-compartment model was measured and expressed as runtime factor relative to method D1. Default settings for method D3, D5 and D7 were 500, 10,000 and 2,000 simulations per FIM calculation. For runtime comparisons, a dedicated cluster node was used.

**Bias and imprecision**

The bias and imprecision of the parameter estimates were assessed by calculating the Relative Estimation Error (REE) for each parameter $p$ in each dataset ($i=1...N$) at each LLOQ, $l$, and for each method $m$ according to:

\[
REE_{p,i,l,m} = \left(\frac{\hat{p}_{i,l,m} - p_{\text{true}}}{p_{\text{true}}}\right) \times 100
\]

(3.15)
where $\hat{p}_{i,l,m}$ represents the estimated value for parameter $p$, $P_{true}$ is the true value of parameter $p$ used for simulating.

Robustness
When using the M3 method for re-estimation, the number of successful runs (i.e. minimization successful) per method and for each LLOQ was recorded as a percentage and compared to the base design.

Expected Prediction Intervals (EPI)
The effect of the parameter uncertainties under the different optimized designs was visualized using the EPI. Prediction intervals for fixed and random effects were obtained for each method at a selected LLOQ of 0.15 (corresponding to 57% data censoring under the base design):
A. 200 datasets were simulated from the optimized design obtained using each optimization method.
B. Parameters were then re-estimated for each dataset with the M3 method. All sets of parameters were utilized in the next 2 steps regardless of their termination status.
C. Each set of fixed parameters was used to simulate out a typical population response curve with a very rich design ($>2000$ sampling times) and to compute a 90% EPI for the fixed effects.
D. Each set of fixed parameters was further used with its respective set of estimated random effects and residual error to simulate out $50$ individual curves with a very rich design ($>2000$ sampling times) and to compute a 90% EPI for the random effects.

The resulting EPIs from all methods were then compared to an EPI derived from a saturated non-optimized design ($70$ sampling times), referred to as the True Model, using the same four-step methodology.

3.6.3. Comparison of optimized dosing regimens
In Paper IV, the optimal dose sizes and administration times obtained from maximizing one of the criteria described in section 3.4.2 were used to simulate one large trial dataset. Subsequently, each dataset was used to derive the proportion of patients experiencing each toxicity grade. The comparison was then made with respect to the baseline dosing regimen study design, i.e. a one-hour infusion of $100$ mg/m$^2$ on the first day of the treatment cycle.

Additionally, maximizing the typical nadir value was used as the design criterion to optimize three doses (with no upper limit on the dose size) given different fixed dosing intervals and a sensitivity analysis was carried out to assess the impact that changes in PD parameter may have on the optimal dosing regimens. Designs were then optimized over a range of $0.1$ to $10^{-74}$.
fold (step decrement/increment by 10 %) the original parameter values of ANC₀, MTT, SLOPE, γ and K₉CIRC. A summary metric representing the dosing intensity was computed as the Mean Dose Time (MDT), such as described by:

\[ MDT = \frac{1}{Total \, Dose} \sum_{j=1}^{14} Dose_j \cdot Day_j \] (3.16)

where Total Dose refers to the maximal dose tolerated in the cycle, Doseᵢ refers to the \(i^{th}\) dose on the cycle and Dayᵢ refers to the \(i^{th}\) day of the cycle. If the optimal schedule gives most doses early in the cycle, the MDT value will be low and if they are given during the last week of treatment, the MDT value will be high. Along the nadir value, the MDT was used to monitor the evolution of optimal treatment with respect to each parameter change.

Figure 19. Visual representation of the Recommended Phase 2 Dose (RP2D) comparison workflow.

3.6.4. Comparison of optimized sampling schedules

REE (as described by section 3.6.2) were used to evaluate the bias and imprecision of predicted nadirs in Paper IV. Three types of comparison were performed: (i) the true versus the estimated population nadir, (ii) the true versus the predicted individual nadir using the corresponding set of EBEs derived from the sampling times in the design, and (iii) the true versus the observed individual time to nadir value.
3.6.5. Maximum tolerated dose determination framework

The overall aim of Paper V was to compare the different dose escalation methodologies to the “3+3” design methodology regarding several features: (i) the precision and accuracy of the RP2D, i.e. absence of bias and replicability of the results (ii) the proportions of patients treated with sub-therapeutic and toxic doses, i.e. doses with probability of DLT <1/6 and >1/3, respectively (iii) the average number of DLTs per trial, and (iv) the escalation dose trajectory with respect to the number of included patients in the trial, i.e. how many patients accrued until convergence to the final RP2D. The framework of this comparison was carried through 10,000 trial simulations for each method, and is summarized in Figure 19.
4. Results

This chapter describes the results obtained in Papers I to V and follows the structure of the thesis’ aims. Most important results are presented per section; additional information can be retrieved in the original papers.

4.1. Statistical hypothesis testing

This thesis extended the existing methodologies for calculation of statistical power in model-based analyses by proposing the alternative MCMP method. The comparison of the novel method with the more established SSE-based approach is the subject of the following sections. Additionally, inspection of the operating characteristics for this new method is reported here as guidance for future end-users.

4.1.1. Performance of the MCMP method

The hypothesis of a possible covariate/drug effect in all explored models was tested by introducing in the simulated model, a covariate/drug effect relationship to one parameter of the model. This way, the population mean value of the parameter is altered by a certain estimated fraction due the inclusion or not of the categorical covariate COV (i.e. value of 0 or 1 according to a predefined allocation design).

Methods comparison

In all explored examples, the MCMP power and the calibrated simulation and estimation based power resulted in an overall good agreement between the two methods as shown in Figure 20. For power higher than 40%, the power estimate obtained with the MCMP method was never off by more than 15% compared to the calibrated SSE. As expected for SSE, actual type I error rates for small sample sizes were found to be above the nominal 5% cut-off value, resulting in up to ~30% power difference between SSEs and calibrated SSEs. Thus, the need of a type I error assessment was necessary, although computation demanding.

In terms of runtime, results generated for the Type 2 Diabetes Mellitus model in Figure 20, in which the estimation step dominated overall run-time,
was shortest for MCMP. The runtimes for the SSE and calibrated SSE results in the same figure were 168 and 1,773 times longer.

Figure 20. Outcome of predicted power study in 6 nonlinear mixed-effects models at varying sample size per study arm from stochastic simulation and estimation method (grey squares), stochastic simulation and estimation method (black triangles) calibrated with type I error rate (dark red diamond) and MCMP method (grey circle).

**Implementation considerations**

Due to the inherent properties of the MCMP method, a large dataset must be simulated for the ensuing multiple samplings of individual objective function values. In the estimation of the necessary size of this dataset, Figure 21 displays the relation between MCMP dataset size and the precision of a “true” number of patients to be included for 90% power ($N_{90\%}$). Dataset sizes above 2,000 and up to 10,000 individuals were found necessary to obtain a variation of this number of patients less than 10% and 5%, respectively. This is further confirmed by the decreasing standard error in 90% power prediction of 6.3%, 4.2%, 2.7%, 2.1%, 1.4%, 1.1% and 1.0% for increasing dataset sizes of 250, 500, 1,000, 2,000, 4,000, 8,000 and 10,000 individuals, respectively. Therefore, it was found in this evaluation that including 33 and 160 times the number of subjects needed to reach the desired power is sufficient if relative standard errors of 10% and 5% are acceptable for the study size prediction’s precision.
Figure 21. Relationship between relative standard error (RSE) of the estimate for number of patients needed to achieve 90% power \(N_{90\%}\) versus the ratio of dataset size \((N=250, 500, 1,000, 2,000, 4,000, 8,000 \text{ and } 10,000 \text{ total patients})\) and \(N_{90\%}\).

**η-shrinkage impact**

Reduction of samples per subject and increase in residual error for the IV bolus example model resulted in shrinkage of 52% for the CL parameter. From the MCMP method, at a sample size of 210 individuals, power prediction from a calibrated SSE was found to be 90.5% versus the power prediction of 90.1% from the MCMP method. Hence, the impact of higher shrinkage did not result in a different power prediction from a calibrated SSE-based one for power assessment in the 80–90% range.

### 4.1.2. Demonstration of application benefits

The MCMP method was further used to derive model-based sample sizes for two different scenarios: a pure POC design with a placebo and an active arm and a dose-ranging scenario. A comparison between conventional study power and pharmacometric model–based power was made essentially at 80% study power in all examples and scenarios, to demonstrate the benefits of applying pharmacometric methods.
POC studies
In the POC stroke example, using a two-sided *t*-test to detect a difference in the change from baseline and day 90 NIHSS score (using last observation carried forward) between placebo and the active dose group resulted in a study size of 388 patients (194 patients/arm), visualized in Figure 22a. Using a pharmacometric model–based approach, the 80% study power was reached with a study size of 90 patients (45 patients/arm), resulting in a 4.3-fold difference in total study size between the two methods.

In the diabetes example, the conventional power calculation resulted in a study size of 84 patients (42 patients/arm) and pharmacometric approach resulted in a study size of 10 patients (5 patients/arm), presented in Figure 22b, corresponding to an 8.4-fold difference between the two methods. Both investigated examples show thereby a several fold reduction in study sizes when employing a model-based analysis.

![Figure 22. Power curve comparison between the pharmacometric model-based power (gray triangles) and the *t*-test based power (black circles), for the POC scenario. Panel A displays the power curves for the stroke example and panel B the diabetes example.](image)

Dose-Finding studies
The dose-ranging POC study scenario also resulted in a several fold difference between the two analysis methods for both disease areas, as visualized in Figure 23. In the stroke example, the pharmacometric approach resulted in a total study size of 184 patients and the *t*-test based study size was 776 patients (i.e., a 4.2 factor difference), as displayed in Figure 23a.

In the diabetic example, using the *t*-test to detect a significant difference between the placebo and the active treatment resulted in a total study size of 168 patients (42 patients/arm) to reach an 80% power, as shown in Figure
23b. The sample size required to reach the same power using the pharmacometric model-based approach resulted in a study size of 12 patients (three patients/arm), resulting in a 14 factor difference in study size between the two methods. The reasons for the increased difference between the methods, as compared with the pure POC scenario, are the nonlinear exposure–response relation that is more informed by multiple dose groups, and the inclusion of a follow-up observation adding more support to the drug effect.

Figure 23. Power curve comparison between the pharmacometric model-based power (gray squares) and the $t$-test based power (black circles), for the dose-ranging scenario, with four parallel arms. Panel A displays the power curves for the stroke example, whereas panel B displays the diabetes example. The $t$-test was based on the difference between placebo and the highest dose group and the total study size was calculated with the addition of two equal sized treatment arms.

4.2. Anticipation of prospective constraints

The following sections describe the results obtained when using OD theory in the planning of trial designs with certain practical or drug safety constraints. An extension of the classical OD approach was explored for integrating prior knowledge of BQL data and the results are presented in section 4.2.1. Another work was also explored to integrate clinical constraints in the optimization of drug regimen, based on utility functions to improve toxicity management. The results of this research are presented in the section 4.2.2.
4.2.1. Application to limit of quantification constraints

In the first part of Paper III, designs with increasing BQL data were evaluated using one of the 7 approaches described in section 3.4.1 as methods D1-D7 without any design optimization. Next, the best methods were used in design optimization. The resulting expected information design content, method runtimes, parameter bias and imprecision, design robustness and predictability are presented here.

Expected design information

In Figure 24, the determinants of the expected and observed FIMs for different LLOQ levels can be interpreted as a summary metric of all parameter precisions across methods. It also displays the empirically obtained OFV as a reference, i.e. from 200 SSEs re-estimated with using the M3 approach.

![Figure 24. Comparison between determinants of FIM for the methods D1-D7 and determinant of the empirical inverse covariance matrix for the M3 method. The matrix determinant is a quantitative metric of the information content (i.e. overall parameter precisions) derived from the evaluation of the one-compartment model using the base design.](image)

With changing levels of censoring, several trends can be observed across methods. In general, with the exception of D1, all methods follow a downward \( \text{det}(\text{FIM}) \) trend with increasing levels of censoring (a decrease in information, as expected). For method D1 (ignoring LOQ), the predicted parameter standard deviations were always identical for any LLOQ levels (as expected), and lower than the empirical ones (an over prediction of the information content). Method D3 slightly overvalued the influence of the censored data and predicted much lower standard deviations for the fixed effects than the empirically observed. Method D4 clearly over-predicted the loss in
information due to the censoring and was strongly dependent on the magnitude of the additional RUV term. Methods D5 and D6 showed a very similar performance because of their similar censoring mechanism and predicted standard deviations and information loss close to the empirical ones, with a slight over prediction of the loss in information as the amount of LOQ data increases. Method D7 most closely mimics the trend seen in figure 24 for the M3 method, and always predicts the information to be larger than the M3 method as predicted by the CRB.

**Runtimes**
The runtimes for all methods relative to D1 are presented in Figure 25. All runtimes were reported as the time necessary to perform one design evaluation using the default settings of each method (i.e. different from the time needed to calculate one individual FIM for some methods). In terms of shortest runtimes method D2 was 1.27 times slower than D1, followed by the method D4 (6.6 times slower), D6 (8 times slower) and D5 (137 times slower). Finally, methods D3 and D7 were 21,000 and 37,000 times slower than D1, which rendered them impractical for design optimization.

![Figure 25. Runtime factors relative to the method D1 for each method and reported with their respective runtimes.](image)

**Optimized designs and parameter bias and imprecision**
Methods with reasonable FIM calculation time (D1, D2, D4, D5 and D6) were selected for design optimization. Four different sampling schedules for the four LLOQ levels characterizing the censoring levels of a two-compartment example model were obtained for each of the selected methods. The final optimal designs are graphically displayed in Figure 26.

For all methods but D1 (ignoring LOQ), sampling times were shifted earlier in time for increasing levels of LLOQ. Method D2 consistently shifted the last sampling point to the time when the population prediction equaled...
the LLOQ. For methods D4-D6, and for most experimental conditions, a
number of optimized sample times were placed in regions where the typical
response would be BQL (but there is still a chance of a non-BQL measure-
ment in any realized experiment). For method D4, all but the last design
point remained essentially the same for the different levels of LLOQ. The
last design point converged towards an asymptote of roughly 18 hours with
increasing RUV and/or increasing LLOQ levels. Optimization using meth-
ods D5 and D6 resulted in similar designs with the last two sample points
consistently being shifted to earlier times for higher levels of LLOQ but with
increased chance of BQL measurements with higher LLOQ values.

![Figure 26. Optimal sampling schedules (5 total sampling times) are represented for
each candidate method and for each LLOQ. The number 2 indicates a clustering of
two samples at the same sampling time. The dashed line linking individual sampling
time (dot) in each method panel corresponds to the trend of this sampling time
across all LLOQ levels. The grey shaded bars indicate where the median response is
above the LLOQ.](image)

In terms of fixed-effects parameter bias and imprecision, method D6 per-
formed the best showing low bias and imprecision in parameter estimates
until the largest investigated LLOQ (corresponding to 73% BQL data in the
original non-optimized design). For the other methods there was clear bias
and/or imprecision for at least one parameter even with the lower LLOQ
levels (with the exception of method D5 with the lowest LLOQ). Method D4
had consistent problems in estimating CL, with high bias and extremely low
precision. For the largest investigated LLOQ, all methods provided bias for
at least one of its parameter and more departures from normal values were
observed. For random-effects parameters, no clear conclusion could be
drawn across all tested methods, as most of the boxplots varied largely between censoring levels.

**Design robustness**
For all methods other than D1, optimized designs up to a LLOQ of 0.1 (corresponding to 41% BQL data in the original non-optimized design) were globally robust. For higher LLOQs, more apparent discrepancies were observed, as illustrated by 63 and 72 out of 200 trials being completed without termination error for methods D2 and D1 respectively, compared to 148 completed trials for method D6 at the highest LLOQ. In general, method D6 was found to provide the most robust optimized designs, regardless of the censoring level.

**Predictability**
The EPI described in section 3.6.2. served as a visual tool to report the predictability of an optimized design. In other words, given a specific design, one might want to assess with simulations the precision and the accuracy of the model prediction. EPI is therefore an elegant solution to reach this objective.

EPI for one LLOQ was explored across all methods selected for the optimization of a two-compartment model and resulting plots are depicted in Figure 27. EPI for method D6 matched best the “True Model” with tight intervals for the predicted fixed and random medians. Method D5 also reported comparable median predictions to the True model, with a more pronounced increase of the fixed effects’ EPI in the terminal phase. Methods D4a, b, c, d, and e provided quite similar EPIs regardless the magnitude used to define the extra additive error. They all showed reasonable prediction intervals for the random effects but an inflation of the fixed effects’ EPI again is observed for the elimination phase of the profile. For method D1, the uncertainty contours of the EPI for both fixed only and with random effects were inflated indicating large uncertainty in both parameter types. For method D2, the EPI involving both parameter types was even larger, with a widening toward the end, indicating that parameters governing the terminal phase of the profile are imprecisely estimated.

4.2.2. Application to dosing and sampling optimizations
Results from dosing schedule optimizations using utility-based criteria for maximization of typical nadir, AUC between grade 2 and 3, and minimization of time spent in grade 4 are reported for 3 and 5 doses in Figure 28.
**Figure 27.** Expected Prediction Intervals (EPI) obtained for LLOQ at 0.15 in the 2-compartment IV Bolus model example. The graphs visualize the influence of the parameter uncertainty due to different optimal designs for both fixed (grey area and red solid line) and random effects (blue area and black dashed line) and were generated from multiple simulations and re-estimations with the M3 method. The reference curve (“True Model”) was generated using a very rich design (70 samples).
Dosing regimen

Respective proportions of patients experiencing each toxicity grade show that the optimized schedules provided safer trials than the conventional treatment with an original 71.8% of patients experiencing grade 4. A 7-fold reduction of number of patients was possible using the nadir criterion, while AUC then $T_{G4}$ provided also improvement. Safer dosing regimen was found when 5 instead of 3 doses were allowed.

![Figure 28](image)

**Figure 28.** Proportion of patients experiencing each toxicity grade per design based on the three different criteria for 3 doses (left) and 5 doses (right) summing up to at least 100 mg/m$^2$ for 1 treatment cycle. 1,000 simulation-based individual nadir values were extracted for each dosing regimen and reported in each respective toxicity grade. Design “Baseline” represents the original single dose of 100 mg/m$^2$.

Figure 29 displays the optimal dose sizes for 3 doses with fixed increasing time of administration intervals when using the criterion D (maximizing dose size while keeping nadir at or above grade 4). When dose sizes were permitted to be unrestricted, a single dose of 70 mg/m$^2$ was found optimal for 50% of patients experiencing grade 4 neutropenia (FO approximation), a reduction by 30% of the conventional dose. As dosing intervals increase, total dose sizes were possible to escalate up to a total of 150 mg/m$^2$ for 3 doses given every 10 days, which increases the amount administered in the conventional first treatment cycle by 1.5-fold.

Sampling schedules

The comparisons between optimization strategies to predict accurately and precisely a population nadir value, an individual nadir value and to capture the true nadir value with sampling times were performed for the 9 sampling times scenario. D-optimality and SRS ETA NOEPS provided the less unbiased population nadir values while C-optimality provided the worst prediction with high imprecision. MAP-optimality and SRS ETA EPS slightly under-predict the true nadir but with better precision. For individual prediction of nadir values, most methods perform equally unbiased, with C-
OPT FIX and OBS, and SRS ETA NO EPS being the least biased. Nonetheless, large imprecision (> 25%) was observed across all optimization strategies. Finally, sampling designs to best capture the true nadir value (to “observe” the true nadir) were derived from both methods SRS, with SRS ETA EPS more precisely than SRS ETA NOEPS. High departure from the true time to nadir was found for C-optimality’s sampling designs.

4.3. Dose-escalation oncology trials

Results of Paper V are reported in this section. First, the final PKPD model built from the clinical data described in section 3.1.2 is reported. Then using this PKPD model, individual toxicity profiles were derived to determine a model-based RP2D using the methodology described in section 3.5.1. for study 1 (abexinostat + free formulation of doxorubicin) only. The final section represents the comparison of the “3+3” RP2D found clinically, the model-based RP2D and the RP2Ds derived from the Bayesian methods.

4.3.1. PKPD model

The final PKPD model structure is illustrated in Figure 30 and the time course of observed platelet counts superimposed on the corresponding predictions of the final PKPD model is shown in the pcVPC, stratified by treatment cycle (Figure 31). The pcVPC shows that simulations from the final model capture both the initial drop of platelet counts following the combined
treatment and the subsequent return to baseline. The variability was however in general slightly overpredicted.

**Figure 30.** Schematic of the final semi-mechanistic PKPD model linking the PKs of abexinostat and PLD/doxorubicin to a PD thrombocypenia model. Grey highlighted compartments correspond to compartment for which data was available. The estimated system related parameters are: baseline platelet count (BASE), mean maturation time (MTT), the feedback parameter on the proliferation rate ($\gamma$) and the release rate from PLD to doxorubicin ($K_{\text{release}}$). The estimated drug effect parameters are the slope for the linear effect of abexinostat ($\text{EFF}_{\text{HDACi}} = \text{SLOPE}_{\text{HDACi}} \times \text{Conc}_{\text{HDACi}}$), and the slope and the power coefficient for the power effect of doxorubicin ($\text{EFF}_{\text{DOXO}} = \text{SLOPE}_{\text{DOXO}} \times \text{Conc}_{\text{DOXO}}^{\text{pow}}$).

For the final structure of the PK model, the PK of abexinostat and doxorubicin were well described by the published models. For the PK of PLD released into DOXO, a first central compartment with first-order elimination best described the initial concentration of PLD encapsulated drug in the blood circulation. The release of free DOXO from the liposomes was best modeled by a time delay represented by an effect compartment (“release compartment” in Figure 30) with a first-order release rate constant $K_{\text{release}}$, estimated at 0.00674 hr$^{-1}$ equivalent to a half-life of 4.3 days. No PK interaction was identified between abexinostat and DOXO/PLD in the combination therapy. The platelet counts were well described by the base PKPD model with a platelet baseline value estimated at 264 x10^9/L, a mean transit time of
100 h and a feedback factor of 0.268. The drug effect of abexinostat was expressed as a linear function dependent on the drug concentration and the drug specific parameter $\text{SLOPE}_{\text{HDACi}}$. The drug effect of doxorubicin was best described by a power function linking drug concentration to the effect by two drug specific parameters $\text{SLOPE}_{\text{DOXO}}$ and coefficient $\text{pow}$. Inter-individual variability was included in the final model for 3 parameters; platelet baseline (34.5%), $\text{SLOPE}_{\text{HDACi}}$ (58.5%) and feedback factor $\gamma$ (28%). IOV was included on the coefficient $\text{pow}$ of doxorubicin drug effect and estimated to 15.1%.

The model parameter estimates were estimated with good precision (relative standard errors for fixed effects parameters < 5% and for random effects < 35%) with low $\eta$-shrinkage for all random effects parameters, except for the IIV and IOV of the DOXO drug parameter $\text{pow}$ coefficient (> 20%). No significant correlations were found between drug effect parameters and system parameters. The log-transformation of the data allowed a better agreement of the residual error distribution to a normal distribution, therefore the modeling of the residual error was best described by an additive error on log scale (approximately a proportional error of 18.6 % in normal scale), with an IIV estimated at 32.2%.

4.3.2. Comparison between dose-escalation methods

Figure 32 displays the simulated probabilities of selecting the appropriate dose level for the RP2D in the explored methods. Throughout all methods, the “3+3” design predicted the lowest RP2D with a median at 75 mg/m$^2$ (3 dose levels below the PKPD model-based RP2D found to be 120 mg/m$^2$) estimated with a large imprecision of 90% CI [30-120] mg/m$^2$. The superposition of the clinical RP2D derived from the two clinical studies to this “3+3” design RP2D distribution shows that the clinical RP2D is clearly under-estimated (4 dose levels below the PKPD model-based RP2D). Interestingly, the 3+3 method recommended 3.4% of the simulated trials to stop at the starting dose of 30 mg/m$^2$ which is known to be a safe dose by main investigators. The highest RP2D was selected by the CRM with a distribution centered at 105 mg/m$^2$ (1 dose below the PKPD model-based RP2D) and a large 90% CI [75-165] mg/m$^2$. A high percentage (16.8%) of trials recommended the highest dose level of 165 mg/m$^2$ which was known to be toxic from abexinostat single-agent trial data, while 21.3% of the simulated trials recommended the RP2D at 105 mg/m$^2$. The RP2D distributions of EWOC and BLRM were centered at 105 and 90 mg/m$^2$, respectively. They both displayed a reduced 90% CI with less frequent trials recommending toxic doses such as in the CRM. Finally, the RP2D distribution using the mTPI method was quite similar to the “3+3” design, with a centering at 75 mg/m$^2$ and a large 90% CI. Unlike the latter one, more trials recommended this RP2D (30 versus 21.6%) and higher dose levels.
Figure 31. Prediction-corrected visual predictive check (80% predictive interval) of the thrombocytopenia PKPD model for the platelet counts following 6 treatment cycles of abexinostat co-administered with Pegylated Lyposomal doxorubicin (PLD) (top) and co-administered with doxorubicin (DOXO) (bottom). Observed data and their corresponding median and 10th and 90th percentiles are represented by black circles and solid and dashed black lines, respectively. 95% confidence intervals of the simulated median (dark grey), and of the 10th and 90th percentiles (light grey) are displayed by the shaded area.

Figure 33 illustrates the proportions of patients being either under/over-dosed or in the target interval of probability of DLT risk, i.e.(0.17, 0.33], for each dose escalation methods. Figure 34 summarizes the rate of patient inclusion for each dose level for 10,000 simulated trials. Across all methods, the “3+3” design shows the least frequent risk of DLT, which is mainly explained by 61% of patients being treated to nontoxic dose levels. The trajectory plot is coherent with this high allocation of patients to low doses since most patient inclusions stagnate at a dose level 60 mg/m² before trials mostly stop at the dose of 75 mg/m². A Last Observation Carried Forward (LOCF) procedure was conducted to map the trajectory up to the maximum sample size of 36 patients (which some rare “3+3” design trials still reach this number of patients inclusion). Among the Bayesian methods, CRM provided the highest risk of DLT, by exposing about 8% of the total patients to toxic doses. Although safety could be an issue, CRM also exposes more than half of the patients to target doses. Additionally, the patient inclusion was much faster at low doses and allows more dose levels to be explored (7 out of 10 total doses). EWOC, BLRM and mTPI show equivalent performance in un-
der-/over-dosing patients, with BLRM being a bit less effective to treat patients at target doses. mTPI was found to be the safest method after the “3+3” design with a probability of DLT risk per trial of 0.156.

![Figure 32](image.png)

*Figure 32. Percentage of trials with selected RP2D for dose escalation methods “3+3” design, CRM, EWOC, BLRM and mTPI. Dashed line represents the PKPD model-based RP2D, and the clinical RP2D was found at 60 mg/m².*

Significant differences were observed in the trajectory pattern. Similar to CRM, BLRM and EWOC quickly escalated the low doses up to 75 mg/m² before stagnating for the next or two subsequent levels, respectively. In mTPI, the inclusion rate was slower with larger escalation steps at 60 and 75 mg/m². Contrary to the “3+3” design, trials were allowed to continue patient enrollment until maximum sample size had been reached. This difference potentially explains the last cohort of patients exploring the next dose level of 90 mg/m².
Figure 33. Percentage of patients in under-dosing, target interval (0.17-0.33], over-dosing category, and probability of DLT risk for dose escalation methods “3+3” design, CRM, EWOC, BLRM and mTPI.

Figure 34. Summary of 10,000 simulated dose-escalation trajectories with respect to the included number of patients for dose escalation methods “3+3” design, CRM, EWOC, BLRM and mTPI. Solid line represents the median, lower and upper dashed-dotted lines represent the minimum and the maximum, and shaded area in grey is delimited by first and third quartiles. LOCF: Last Observation Carried Forward.
5. Discussion

5.1. Statistical hypothesis testing

Formulating and testing hypotheses are integral part of any scientific research. In drug development, hypotheses are generally generated at every stage to learn from the drug and the biological system, but notably to inform decision-makings at critical stages such as POC and DF studies. Therefore, the subject of the first part of this thesis was to expand traditional model-based power assessment to a faster methodology that requires less computation burden. This newly developed Monte-Carlo Mapped method (MCMP) was applied to two complex nonlinear mixed-effects models to assess the sample size necessary for detecting a drug effect for these trials.

POC trials in phase IIa are often categorized as the first confirmatory trial in a drug development program and it is not uncommon that the primary analysis is similar to the analyses used in the phase 3 trials. However, the informativeness of the trial data is often diluted by traditional handling of the data analysis. Using only end of study observations and discarding the rest of the data are typical characteristics of pairwise comparisons. Additionally, although POC trials are often executed with multiple treatment arms to fulfill secondary objectives such as exploring dose–response relations, these comparisons are typically conducted between one active dose and placebo, discarding therefore the other dosing arms. Such approaches, e.g. t-tests, make interpolations between treatment arms difficult and reduce the ability to propagate knowledge about dose/exposure–response to future studies.

As these results show, the use of a pharmacometric model–based approach within drug development has the potential to reduce study sizes of clinical trials. One of the main reasons for this is the use of longitudinal data. The pure POC example contains minimum information about the drug effect, involving only one active treatment arm and placebo, nevertheless by including all data available (i.e., repeated measurements and simultaneous different endpoints) the pharmacometric approach results in a several fold reduction in study size. Mixed-effects modeling is also flexible when dealing with unbalanced repeated measurements which is often the case in clinical trials, and the utilization of a pharmacometric model–based analysis ensures that all available data are used in the primary analysis thereby increasing the information content of the trial, as these examples clearly illustrate. Another incentive to use this approach is the detection of a drug effect which rela-
relationship with respect to the dose/exposure is nonlinear. As the results from the diabetes example indicate, this relationship described in POC and DF studies resulted in further drastic reduction between the two scenarios. This is mainly explained by the fact that nonlinearity is more informed if multiple levels of doses are tested. On the other hand, if the drug effect was linear such as in the stroke example, study sizes would result in the same factor difference in between the POC and dose-range scenarios. Naturally, a collateral increase of precision in the drug parameter will also most likely occur with the addition of more dose levels (not explored in the present investigation). Finally, a pharmacometric model–based power can be combined with a formal optimal design to pin down the most informative clinical trial design in terms of both study power and parameter precision. Ueckert et al. performed an explicit optimization for statistical power in the planning of a study in Alzheimer’s disease, resulting in a 30% smaller study size when targeting 80% power. Naturally, there are many methods available for the statistical analysis of clinical trials and a few other pharmacometric model–based analyses, however, both Jonsson and Sheiner and Hooker et al. have presented results that indicate that model-based methods lead to a reduction in study sizes and should be use more profusely. Nevertheless, one of the major drawbacks of the model-based power assessment is the dependence over the availability of a pharmacometric model. If the information about a model is very limited, a model-based power calculation may not be sensible. Alternatively, this “lack” of a proper model can be salvaged by developing a “best guess” model from preclinical data, or a structurally similar model from a predecessor compound. Another downside relative to the model selection is the assumptions of no model misspecification and the detection of a drug effect different from zero. It is reasonable to believe that model misspecifications will lead to imprecision in the statistical power, hence an uncontrolled type I error rate that will necessitate a calibration of the cut-off value for the test statistic. Nonetheless, in this project, both scenarios did not suffer from the error associated with model misspecification, since the simulated and analyzed models were identical. In real life, this might rarely be the case, since no “true” model can be fully captured.

The study power calculations based on pharmacometric models historically often rely on simulation and estimation exercises which can be very time consuming and, therefore, not extensively used. The newly developed MCMP method for calculating the study power has the advantage of being a much faster method than the traditional simulation and estimation procedures, making a pharmacometric model–based power calculation more accessible. Tested on several types of data and for real-life PK and PD models, the methodology has demonstrated a good agreement in power prediction and requires only one simulation and estimation step, hence leading to an important reduction in time and computation load. This substitution is ex-
plained mainly by the fact that the overall objective function value specific to a given model, design and dataset can be described by the sum of individual objective values, allowing iOFV values to be sampled instead of OFV values from simulated studies. Nonetheless, two main technical aspects to be considered are the number of random iOFV samples and the minimum size of the MCMP dataset to have acceptable precision in the study size estimates for a given power of interest. In all explored examples, a number of 10,000 stochastic samples results in less than 1% of relative standard error in the number of subjects’ value to reach 90% power, while the size of the MCMP dataset must include ca. 33 and 160 times the number of subjects needed to reach the desired power if 10% and 5% of relative standard errors are acceptable for the study size prediction’s precision. Naturally this size is not known before the first MCMP dataset size is chosen, so if a too small study size was chosen, a repeat evaluation with a higher MCMP study size may be necessary to reach desired precision. This ratio is also expected to be effect size- and model-dependent, but from all explored examples, a 50-fold of the number of subjects needed in the study for a 90% power assessment provided acceptable relative standard error values.

In terms of statistical inference, the log-likelihood used in this method allows making stronger inference based on each parameter change of the model and in their respective correlations, unlike power calculation methods derived from the linear Wald test, which generally assumes symmetric confidence intervals and unbiased parameters estimates. Additionally, the LL-based methods, unlike the Wald test again, allows the use of estimation models that are different from the simulation model, the inclusion of different possible sources of bias in the model and the test of different estimation methods. Recently, Plan et al. demonstrated the difference in power assessment if simplified estimation models were used to characterize events with graded severity \(^{163}\).

From the investigations on several models, the MCMP method was found to be a good approximation of the outcome of a calibrated SSE for power in the main region of interest (i.e. 80–90%). However, it may be less precise in regions for powers lower than 20%. A possible explanation of discrepancies at these low powers is suggested by the omission of the estimation step for each sample size of the MCMP power curve, which unlike nonlinear mixed-effects maximum likelihood estimators, does not acknowledge the asymptotical normality in its estimates with respect to the number of individuals. This is nothing that is unique for the MCMP method; indeed every method that uses this type of scaling without estimation, e.g. Fisher Information methods, will suffer from this unwanted property. Furthermore, scaling sample size without estimation will assume the same bias (size and direction) as the bias from estimation with the big data set. In this case, the effect on the power due to this assumption is much harder to predict because the bias might change sign and/or size differently between sample sizes and parame-
ters. Nonetheless, the reduction of number of samples per individual and the increase of residual error, hence the impact of higher shrinkage, did not result in a different power prediction from a calibrated SSE-based one for power assessment in the 80–90% range.

Regarding the need for type I error assessment, the dataset sizes at which the MCMP is run in estimation are in the region where the calibration indicates a close to nominal type I error magnitude. This is the reason why no type I error calibration was necessary with the MCMP. However, the MCMP method only claims to remove the dependence on sample size of the type I error and does not acknowledge other reasons for why a type I error rate can deviate from the nominal such as model misspecification. Additionally, the MCMP method does not inform on any design flaw that will make a model numerically unidentifiable. Performing a few simulations and re-estimations with the decided sample size from the MCMP method could be used as a confirmation of a proper model and numerical identifiability of this model.

In conclusion, the development of more effective methodology for power calculations applied to nonlinear mixed-effects models are hence believed to lead to more informative and efficient clinical trials, such as indicated in these few examples.

5.2. Anticipation of prospective constraints

In this second part of the thesis, the anticipation of BQL observations and dose-limiting toxicities in the planning of a prospective trial was addressed using optimal design. Generally, a single “true” model in the optimal design performed in Paper III and IV was assumed and therefore, the methodologies presented here are adapted to clinical development stages with sufficient confidence in the pharmacometric model. Nonetheless, global optimization using E-family optimality criteria and the same principles developed in these two papers can mitigate this issue in case of less refined models.

The premise behind Paper III and IV was to advocate the importance of incorporating as much available information as possible in study design evaluation and to include upfront design constraints in the planning of a trial. This may in return allow scientists to wean “rescue” analyses off from failed trials that have not been optimized for the analysis.

In Paper III, four new methods were developed to handle LOQ data in optimal design calculations for trials with BQL data. These four new methods and three methods previously seen in the literature were then implemented and compared in a conventional optimal design software PopED. The methods’ evaluation was carried out across a range of LOQ censoring to ensure discrimination between methods.

Clearly in all tested scenarios, trivial strategies to ignore the LLOQ (i.e., to plan the study without considering the limit – method D1) or to complete-
ly avoid BQL observations (i.e. chose sampling times that essentially have 0 % probability to be BQL – method D2) are non-optimal. The best performing methods if censored data were unavailable numerically, rely on strategies that are between those two extremes, i.e. the information content of a sample below or above LOQ is not zero but is also not as informative as a normal measurement. Consequently, methods D3- D7 propose different strategies in weighting sample times’ information content.

For the scenarios investigated, method D5 (simulation and FIM rescaling) and D6 (integration and FIM rescaling) provided the most accurate and precise parameter estimates and showed the best compromise in terms of runtime and efficiency for design evaluation and optimization. In fact, method D5 scales the population FIM with the probability of BQL predicted from simulation, whereas method D6 scales it with the probability of BQL calculated from the FO approximated joint density with an assumption of normality of the responses. A close correspondence in the predictive performance between D5 and D6 was expected and further observed in the investigated examples, which show that the normality assumption in method D6 was not an issue. However, based on the difference in runtime, method D6 may be more often selected since it performs faster than D5, the later relying on simulation (this number was evaluated to 10,000 samples). Contrary to general belief, method D4 (additive error) was found inadequately insensitive to the range of censoring levels tested in design optimization, and definitely inappropriate in design evaluation. This is inherently due to the fact that predictions with this method are poor because the additional additive error is included into all sampling points, regardless of their censoring probability, inflating thereby the overall parameter uncertainty. Method D3 (individual response set to zero) and D7 (integration of BQL data in its likelihood calculation) are also both middle ground methodologies, that particularly inform on the individual probability of BQL data. In Method D3, the FIM is downscaled after removal of individual sampling times due to censoring, but simulations of these profiles render the runtime impractical. On the other hand, method D7 acknowledges the information content present in individual BQL data in a way similar to the M3 estimation, therefore, standard deviations and determinants of the FIM derived from method D7 and M3 showed a similar trend across all LLOQ levels. Nonetheless, extremely long runtimes due to its algorithmic complexity also render method D7 impractical for optimization, but possible for design evaluation if the goal is to obtain precise expected estimates. Finally, method D1 and D2 represent the extremes of handling BQL data, by either ignoring or omitting them in the analysis. Although both methodologies are shown to be bias, they might constitute quick alternatives for designs with low percentage of BQL data (e.g. < 20%). In fact, the impact of such condition has been shown to have minimal impact on the ensuing data analysis66,67, which makes method D1 and D2 attractive for optimization in these situations. Additionally, when method D2 is used in
design optimization, the method consistently assigns one sample time at the latest time point of the experiment. The latter can be considered in some cases optimal in the restricted space limited by the LLOQ, e.g. late sample time for the clearance value.

One possible limitation of the evaluation of these methods is the use in the M3 method of the Laplace estimation method as reference. It has been previously highlighted by Plan et al. that under certain condition the Laplace estimation method might show considerable bias. Alternative estimation algorithms can therefore be tried such as SAEM or Gaussian quadrature, which, in some cases, have been demonstrated as superior. The choice of a simple one-compartment model for the design evaluation was one measure to reduce the contamination through bias.

In conclusion, this is the first study evaluating the relative merits of including BQL data using population optimal design theory in the planning of a prospective trial design. Nonetheless, if BQL data may be retrieved from laboratory analysts, they should be preferred and used before reverting to the methods developed in this Paper III.

In Paper IV, methodologies for dosing and sampling optimally a trial with common hematological adverse events were developed. Based on the Quar-tino et al. model and the anticancer agent docetaxel, the goal was to optimize the trials to expose as less as possible the number of patients to the experimental drugs’ side-effects. For that purpose, clinically relevant optimal criteria were defined based on several components of the toxicity profile instead of the classical nadir value below a pre-specified toxicity grade threshold. The new criteria based on both time course of the toxicity biomarker and drug effect, found to be highly correlated with the prediction of Febrile Neutropenia (FN) were also used to explore other aspects of neutropenia such as the duration of grade 4, an absence of at least baseline value at the end of the treatment cycle (which mostly result in dose administration delay) and a therapeutic window delimited by grade 2 and 3, deemed to be a proper trade-off between toxicity and efficacy. Thus, the quantitative knowledge of the full neutropenia time course offers a multitude of possible optimizations that could be customized according to the objectives from different stakeholders.

The application of OD to optimize the dosing schedule of docetaxel resulted in a 7-fold decrease of patients experiencing grade 4 neutropenia and a 1.5-fold increase of the total amount of drug usually administered under a standard first treatment cycle if using an optimized dosing schedule. Resultant designs were determined with some practical constraints (not all constraints included), ranging from discretization of the doses to off-treatment periods. In addition the total amount of drug within a cycle of 3 weeks was kept at least at 100 mg/m², the same amount as currently used in clinical practice. Considerations from other side-effects such as fatigue/asthenia were as well reflected in the allowance of 1-week rest interval.
Most of the criteria did not take into account the inter-individual variability (the use of FO approximation which optimizes for the typical patient) which is stated to be a serious impairment in clinical practice for docetaxel\(^{154}\). A possible improvement could therefore be the use of the FOCE method, which will simulate out profiles from the distribution of parameters and therefore allows selecting a design that will in average satisfy the criterion. Another limitation of these approaches was the indiscrimination between a grade 4 with nadir value at \(0.5 \times 10^9/\text{L}\) and at \(0.1 \times 10^9/\text{L}\), which clinically may be more related to lethal side-effects. This could be improved by assigning weights to restrict attainment of nadir values in such zones. Finally, all optimization were only performed in terms of toxicity and not taking into account the drug’s efficacy. This could correspond well to a classical Phase I oncology trial for cytotoxic agents, which mainly focus on the safety profile of the drug. However, if efficacy should be influencing the choice of a dosing regimen, a Clinical Utility Index (CUI) could be utilized where a utility function could better reflect the dual approaches. The latter may relate well with the surge of targeted therapies and for drugs with narrow therapeutic windows.

In sampling schedule optimization, 9 total sampling times and a period of on-treatment followed by an off-treatment permitted the full illustration of the characteristics of each developed methods. In general, D-optimality provided the most accurate typical nadir values, while sampling the “observed” nadirs was better captured using the SRS methods. The MAP-optimality based on minimizing the variance of the EBEs should \(a\ priori\) provides the most accurate and precise individual estimated nadir values. However, none of the methods performed significantly better and all presented very high imprecision. This is mainly explained by a high \(\eta\)-shrinkage for some random-effect parameters in the original model, which thereby did not allow obtaining better precision. Nonetheless, a possible improvement to limit this shrinkage could be to increase the number of sampling times per individual, which might however exceed the number of sample times feasible in practice.

In conclusion, OD methodology can be applied for toxicity monitoring in oncology studies by using clinically relevant optimality criteria within clinical constraints. Pre-determination of safety limits in the planning of a dosing regimen (utility functions + Newton-based algorithms) can impact the optimal dose scheduling and allows less toxicity outcomes.

### 5.3. Dose-escalation oncology trials

The last part of this thesis specifically focuses on dose-escalation algorithms used in phase I oncology trials. In Paper V, the overall scope of the dose escalation methods comparison was to provide insight into the capability of
the selected methods ("3+3" design, CRM, EWOC, BLRM, mTPI) to correctly determine the MTD (then the RP2D), and to achieve desirable operating characteristics that make a trial efficient and ethical.

The foundation of this work relies mainly on the use of a model-based in silico framework that provided the knowledge of a "true" RP2D and the possibility to stochastically sample individual toxicity profiles, entering each of the simulated trials according to the used method. Additionally, this framework permitted to derive a model-based PKPD RP2D (referred to as the "true" RP2D) obtained from selecting the exact dose that gave a toxicity risk of 33%, based on simulations using a PKPD model developed here from the trial data.

The developed PKPD model in this work represents the physiological thrombopoiesis process from the bone marrow, to the differentiation and the circulation of platelet in the blood circulation. Drug-related thrombocytopenic toxicity, notably grade 4 as a nadir value below $0.5 \times 10^9 / L$, was introduced by the use of a combination treatment of abexinostat and doxorubicin into two different formulations, a free and a pegylated liposomal form. Both formulations were integrated in the final model, by connecting them via a release compartment with a release rate estimated at $0.00674 h^{-1}$, found to be in agreement with literature in vitro release rate values\textsuperscript{166,167}. The developed PKPD thrombocytopenia model was finally validated by the use of pcVPC, which showed a good agreement between observed and simulated observations, while PPC qualified the model to adequately predict the nadir value at cycle 1 as it is usually carried out in clinical DLT assessment.

Using the simulations based on this PKPD model, RP2Ds were derived from the different methods then compared. Comparison between dose escalation methods essentially showed that the traditional "3+3" design was not the optimal dose escalation strategy. The resulting "3+3" RP2D was imprecisely estimated and was found too conservative. The results here were not unforeseen although 95% of all trials are still conducted using this methodology despite its poor operating characteristics\textsuperscript{15}. Instead, Bayesian methods have been shown here to be potential alternatives in this comparison. The CRM method recommended the highest and closest doses to the PKPD RP2D and allocated the most patients at the targeted dose level. Furthermore, the number of dose levels to be explored was larger and dose escalation much quicker. Nonetheless, about 8% of the patients were exposed to toxic doses, and mostly resulted in higher number of trials recommending those doses as RP2D. This caveat was corrected by the inclusion of uncertainty in the model parameters using EWOC and BLRM methods, which improved considerably the over-dosing risk of toxicity by more than 7-fold. These results concurred with the expected performances of these two methods, since EWOC and BLRM actively maximize the probability of targeted toxicity similarly as the CRM, yet controlling the probability of excessive or unacceptable toxicity.
Although Bayesian methods have been proven to have better operating characteristics than the “3+3” design in general, several pragmatic issues have limited the use of these methods. One of them is the non-intuitive decision-making behind such methods. The modified Toxicity Probability Intervals (mTPI) offers a middle ground in the spectrum of the “3+3” design and the fully model-based designs. However, the RP2D distribution and dose trajectory using this method did not differ much here from the “3+3” design’s, and its superiority over the Bayesian designs was not fully demonstrated in this example.

When it comes to using Bayesian strategies, a sensitivity analysis should be conducted on the assumption of the dose-toxicity model and the prior distributions. In fact, the choice of prior distribution is of utmost important and may be rather tricky, since distributions are typically elicited from experts familiar to preclinical data or previous knowledge of the drug. Nonetheless, a clear misspecification of this distribution may lead to poor performance of Bayesian methods, which will dilute the superior characteristics found in these methods over the “3+3” design.

Another drawback of all the methods stated above is also the underlying assumption of the relationship between efficacy and toxicity described typically as monotonically increasing with increasing doses in most cytotoxic drugs. Consequently, these methods have used toxicity as the primary endpoint. However, with the rise of molecularly targeted agent therapies, the dose-efficacy and dose–toxicity curves may differ from those for cytotoxic agents, and efficacy may occur at doses that do not induce clinically significant toxicity. Thus, for trials involving this kind of agents, the primary endpoint may instead be a target inhibition in tumors or surrogate biomarkers or a biologically relevant pharmacokinetic level. Besides, characterization of a proper dose-toxicity and dose-efficacy relationships are in this case of primordial importance. In this perspective, the MBDD approach can integrate both of these profiles by acknowledging the underlying biologic rationale behind the toxicity and the efficacy of an agent. Additionally, in case of a combination treatment, PKPD models allow to assess the true nature of two agents’ interaction, e.g. overlapping, additive, complementary, or synergistic, and provide the dynamics of a toxicity variable based on both effects. This approach may therefore provide an alternative way to the currently proposed dose escalation methods for drug combination therapy, typically based on empirical exploration of tolerable combination dose levels, informed from single-agent dose escalation studies.

Finally, the methodologies described in this article does not account of patient’s time to toxicity. They do not thereby accommodate for late-onset or cumulative toxicities which are common in oncology such as in radiotherapy. Furthermore, assessment of DLT usually requires that the toxicity outcome is only considered in the first treatment cycle so that by the time of the next dose assignment, all toxicity outcomes of currently treated patients are
completely observed. The RP2D selection based on these practices might therefore be “optimistically” biased and might possibly lead to later dose reduction for patients treated at this dose in Phase II/III trials. In these cases, the conservatism of the “3+3” design might have thus selected a more adequate RP2D for long-term treatment. Nonetheless, one obvious solution to this problem would be to increase the DLT time monitoring, although longer trial length is generally not highly desirable. Alternatively, Cheung and Chapell\textsuperscript{168} offered an extension of the CRM method known as the Time-To-Event CRM (TITE-CRM). A third alternative proposal might be again the use of a PKPD model which allowed longitudinal toxicity data to be monitored.

In conclusion, this work was in line with the methodology shift advocated by regulators, industrials and academics in phase I oncology studies. The favorable operating characteristics of Bayesian methods should preferably be used instead of the “3+3” design and the integration of early PK and PD data trials in a PKPD toxicity model may offer a powerful tool to diagnose ill-determined MTD.
6. Conclusions

Drug development and regulatory decisions are driven by information that is compiled primarily from clinical trials. Making studies more likely to succeed and decrease the time between discovery & “confirmatory” studies are imperative components of the sustainability of the drug development system. In reaching such standard, one might need to reconsider several aspects of the conventional trial design, such as pairwise comparison, single endpoints, single imputations of missing data and rule-based designs. Instead, employing the development of pharmacometric models coupled with optimal design theory in this thesis has helped mitigate some main drawbacks of these methodologies.

Novel pharmacometric approaches were developed to usher new treatment through a more rapid, minimalist, informative, effective and ethical clinical trial workflow. The MCMP method was able to provide relevant power information at less than 1% of the runtime of a conventional model-based power assessment. Additionally, statistical inference using nonlinear mixed-effects models was performed across all dose group levels, integrates the drug effect over the complete time course of the study and incorporates all data available, which inevitably resulted in several fold smaller study sizes in critical decision-making stages such as phase II proof-of-concept and dose-finding trials. Loss of information was also improved by incorporating forefront at the trial design level preconceived analysis elements such as censored data. Thus, by anticipating BQL observations in the planning of a prospective trial, the newly developed optimal design methods permitted to design trials that show better estimation properties and robustness for the ensuing data analysis. Effectiveness and ethical conduct of trials were also emphasized in the evaluation of new oncology dosing strategies with primary clinical metrics and practicability constraints being at the center of the trial optimization exercise. Furthermore, novel exploration of dose-escalation oncology trial designs has allowed favoring Bayesian approaches over traditionally rule-based algorithms, while steering the use of pharmacometric models as an effective companion diagnostic tool to guide dose-escalation studies.

In conclusion, pharmacometric model-based approach and optimal design theory have permitted to minimize time and computation burden, number of individuals, loss of information, patient risks and suboptimal treatments.
Eventually, we hope that drastic streamlining of clinical trials is afoot with more model-based support.
Acknowledgments

The work presented in this thesis was carried out at the Department of Pharmaceutical Biosciences, Faculty of Pharmacy, Uppsala University, Uppsala, Sweden. Institut de Recherches Internationales Servier, Suresnes, France, is gratefully acknowledged for financial support of this PhD thesis. I am very thankful for the opportunities of educational traveling to workshops and conferences mostly from Apotekarsocieten and Smålands nation; Furthermore, for parts of the research in this thesis to be funded by the DDMoRe initiative.

*If I have seen further it is by standing on the shoulders of giants.*

Isaac Newton

Given my size, I think Sir Newton was quite early in time to figure out that I literally could have been carried by everyone that I was fortunate enough to encounter during these PhD years in Sweden. Nonetheless, more than “standing” on shoulders, I was granted with immense help, support and guidance, merely an approximation of all the wonderful feelings passing through my memory –Nothing of this journey would have ever happened without:

*Prof. Mats Karlsson,* my primary supervisor, for being the first M in MCMP and for embedding my role model of a brilliant scientific leader steering this virtuous cycle that everyone of us each day benefits from! Thank you for catapulting 1000 ideas per minute and made me all wonder how a human brain can do that. And many more thanks for the later non-scientific sessions, there will always be a very grateful heart for all this.

*Assoc. Prof. Andrew Hooker,* my second supervisor, for always being enthusiastic to all my trial-and-errors in OD and for always supporting my choice in projects I was interested in even if it leads to unexpected paths in my PhD research plan. You have been my cheerleader Number 1 since the “how to be a good PhD student” session, thank you!

*Assoc. Prof. Lena Friberg,* my third supervisor, for teaching me very patiently modeling, for your thoughtful and always detail-oriented supervision and support for PAGE and for always keeping an open door to all my (ir)relevant
questions. Many thanks for your constant kindness and for having me hooked to the oncology field!

*Dr. Marylore Chenel*, my fourth supervisor, for your trust received when being accepted as a PhD student at SERVIER. Many thanks for giving me the opportunity to have a different perspective of my work under the light of drug development in an industry setting and for all the pep talks during my illness. Your optimism in turning all problems into possibilities is very much contagious!

*Prof. Margareta Hammarlund-Udenaes*, for making this department what it is, i.e. an amazing place to work and to be. For unifying us every Monday with the news of the week at the sound of your voice “Monday again”!

*Dr. Kajsa Harling and Rikard Nordgren* for all the technical support in new functionalities of PsN such as your help in implementing the *mcmp* function, in deleting my recurrent “ghostly” screens on the cluster, and for one of the best spex I’ve read, Kajsa!

*Dr. Anja Henningsson*, for saying out loud the unfortunate truth about the reading rate of our kappa (*nobody reads it anyway*), and having the brilliant suggestion to rather add a tag line of my thesis in the most-rated acknowledgments section. So People, if you have to remember of one thing, here it is: “Think about your study design”. Done!

*Dr. Siv Jönsson*, for lending the most caring ear to all my recent life troubles, to have brought renaissance into my feelings for this PhD and to constantly encouraging me to be proud of who I am and what I’ve done. You have helped in such a way that is beyond words.

My main co-authors, *Dr. Martin Bergstrand*, for always believing in me more than I do in myself, for being the second M in MCMP, for ice cream & Life discussions, and for letting me & Paolo register the first trademark of "Individualized prediction- and variance-corrected trips®", *Dr. Kristin Karlsson* for always leaving an open door to my statistical questions and for all the great teaching in power calculation, *Dr. Niclas Jonsson* for your big ideas and hopes on pharmacometrics, *Dr. Joakim Nyberg* (I cannot see any Matrix films without seeing you as Neo anymore) for introducing patiently the FIM to a snail learner and for rooting my interest in study designs. Thanks for all the Greek equations and for the humbleness behind such a talented scientist! *Dr. Sebastian Ueckert*, for always bringing new ideas to all my projects and for all your advices to write a thesis and a paper in a non-schizophrenic way (I quote you!) :P I have an excellent memory of the most “optimized” trip from Sils Maria due to the mess done by the “It-Who-Can-
Not-Be-Named” volcano, Sylvain Fouliard, for always being so patient and always available for an interesting discussion in modeling. Thank you for following me in the crazy twists of our project: we for sure have gone far far away from our initial 8 patients! And Dr. Quentin Chalret du Rieu, gosh we are finally finishing! Thank you for being my partner in crime at SERVIER!

Dr. Ron Keizer and Dr. Angelica Quartino, for being my half-time opponents – Ron, for your piano encouragement and inspiration for the thesis cover and Angelica for teaching me to like pink and all the girly talks. I really wish San Francisco could be closer!

Magnus Jansson & David Khan, for being the best co-founder team ever of the Farmbio Movie Club and for letting me (sometimes) choose a non-guy-ish film!

Brendan Bender, for liking very much to contradict you ALL THE TIME, for my desperate moment of spex editing (btw, you nailed it, thinking of a career change?) and for the great time in Grenoble when you introduced me to your top 5 pharmacometrician couples ranking (wink, wink)

The past French Team: Brigitte Lacroix for the Ying the Yang and the Sushi, our own version of Spaghetti Western, Dr. Alexandre Sostelly for fighting with me the “evil” FOCE FIM et les pauses “cigarettes”, Dr. Emilie Hénin, for teaching me skiing and always making the best bakeries in town, Dr. Paul Baverel, for lots of starting encouragements, Dr. Audrey Deligny & Dr. Céline Monard, for helping me to remember how to speak French during the French Connexion!

Oskar Clewe The Best West Coast for accepting the bet of the most being cited in Thesis acknowledgements and that you are clearly losing now ☹… Btw, I love your fish yellow sweater, very Andy Warhol! Ida Netterberg, for all your superb “management” skills and your wonderful jamming energy, Dr. Irena Loryan for the immense care and support for me and my family and all your constant encouragement, Dr. Julia Korell, for always being so cheerful and lively, Henrik Bjugård Nyberg for so nicely giving me my last “home” in Uppsala, Ana Kalezic for saving me and my father twice from being homeless in Uppsala and for your endless kindness, Emilie Schindler & Anne-Gaëlle Dosne, for all the help in “lost in translation” between mom and the PM group, Benjamin Guiastrenec, for your R-hero cape, Chunli Chen, for tempting our bellies to the great food of Uppsala and “making my week” and Dr.Waqas Sadiq, you are the smiley face of this department!

Dr. Yasunori Aoki, for reminding me to never let “others drown my inner voices”, for all OD and use of pharmacometrics discussions, for sharing the
same passion for TED talks. Little disappointment for you though… I am still not converted to the MAC yet :P

*Dr. Anna Largajolli* for your lovely smile and for being WonderWoman with the red sofa and with all my suitcases. Your friendship has made everything easier and sweeter, go my OFV girl ☺

*Eric Strömberg and Caroline Frisk,* Alicante, there we went and so much fun!!! Many many thanks *Eric* for allowing me to be Camille the Film School version again by sharing with me discussions about Sci-Fi & Game of Thrones, and being the best Powerpoint buddy ever!!! And thumbs-up for the exploding sphincter Nutella cake!

*Dr. Jan-Stefan van der Walt* aka the JS, I am very proud to be your “pluripotent” friend and for having the chance to receive so many Bear hugs. You’re my Sweetie this time, love you 😊

*Dr. Elodie Plan,* my dear “Grande Soeur” from Day 1, you kept my mind clear by constantly reminding me that I am the master of my own ship and for all the valuable advice on Life and career. And you fill it with beautiful thoughts from our mutual passion for graphic arts, good films, and humanitarian work. Merci pour toute ton inspiration!

*Elin Svensson and Dr. Thomas Dorlo,* for the constant inspiration that you both carry through your endeavor in making this world more fair to everyone, for all the great veggie dinners and discussions about DNDi, MSF and humanitarian documentaries. We need more individuals like you on this planet!!!

*Dr. Coen van Hasselt,* my dear friend, thank you for the title of this thesis, for always taking the time to reply to a quick email and for all the great scientific discussions in AOD. I always put your opinion on a pedestal and will continue now and then. You are one of the smartest ones I’ve got to meet in this lifetime!

*Dr. Ami Mohamed,* you have made me experienced so many things! Your first driving on the right-hand side in Korea, our food binging in Japan, my longest “paparazzi” job ever for an inert statue in Taiwan, all the sarang hae and the crazy driving in the middle of the night to see the sunrise on Seongsan Ilchulbong. I miss you very much!

*My dear Anders Kristoffersson,* everything is way too silent now without your whistle!!! You are forever my Mister Wikipedia and it was my greatest luck to have shared an office with you! Thank you for all the great OD dis-
cussions, for the reindeer cookies on my birthday, for teaching me how to make “wine” with the berries in front of BMC and for the idea to have the “world’s first professor of pharmacometrics from Götland” (ditto Wikipedia) as one of the welcoming portraits at the Arlanda airport!

Dr. Paolo Denti, we are forever the best advertisers for the once-in-a-lifetime "Booze Cruise to the Baltic” with our team of Brazilian masseurs Robertinha and Rejinaldo! Thank you for the fantastic memories of Greek islands and Andalucia; and I promise you that beautiful Cape Town will be my first destination when I will be fit enough again to travel there!

Dr. Marilee Andrew, my dearest Big Sister Cat, thank you for making my first steps in this PhD less scary and for teaching me to embrace this vulnerability that you and I both share. Your beautiful soul has left a handprint in my heart, I dearly miss you!

Dr. Elke Krekels, my dearest friend, this whole PhD would have not meant the same if you’ve not been my “reasoning brain”. Thank you for keeping my dad on track at the hospital (you are forever his champion), for all the constant support and help that you have given me. A lifetime of chocolate supplies would not even be enough to give you back for all of this 😊

Dr. Åsa Johansson, there there (Sheldon way) my crazy ice-cream-lover friend, how can I ever thank you for always being the safe harbor of all my hardships and heartaches? I would have never made it through this PhD without having you by my side, know it in your heart! I also hope that this won’t be “vår sista dans”, but the continuation of a lifetime friendship, the longer the better! Puss och kram!

All the fantastic and wonderful colleagues at the department – past and present - for making this group a myriad of inspiring, dynamic and talented people and for making this place my second home. I feel extremely fortunate and thankful to have been part of this “family”. Special thanks to Marina Rönngren for all your help & concern!

All the fantastic and wonderful colleagues at SERVIER: Myriam, Roeline, Karl, Charlotte, Maud, Marie-Hélène, Isabelle, Clémence, Céline, Kathryn, Katie, Tanguy, Sophie P, Sophie G, Jacques, Etty, Manon, Olivia, Klervy, Adrien, Simon, Alexia, Noella, Clement and special thanks for Anne Dubois and Cyrielle Dumont – Merci les filles de m’avoir épauleté tant de fois!!!

Margarida, Chris and Johan, for always being here - up and down, sun and rain, day and night. You’ve made my life in Uppsala so wonderfully great,
thank you for all our evening dinners when we mutually mute our fears for the future and make any banal conversations into a giggling exercise!

*Juan,* my dear “garde-boue” friend, for taking care of me after a late shift at Fyrisbiografen and for chasing together the best night shots in Uppsala and *Vladimir,* for all the best hot chocolates in Uppsala and Paris and for the board game nights. You guys rock!

*Damien,* for always catching up with me at PAGE as if it was still yesterday, thank you for keeping in touch after so many years! *Matt,* always grateful that we both waited to push the same stop button in the bus to Valbonne and since then, our friendship has never ceased to boom, *Cristina* for all the strength and support you gave me, merci à toi, la plus FORTE personne que je connais! *Siv & Pierre* for all the great board game afternoons and the “chasses aux indices” in Paris, *Diem* for always being here taking care of me and my parents and for being the big bro I’ve never had, *Irina, Alexandra, Sam and Mandy,* my dear cinema friends, thank you for always reminding me my great passion for filmmaking – your talent and resilience are priceless!

*Flo & Isa,* you are my oldest childhood friends and the rare ones I could count on the fingers of my hand. Thank you for never forgetting me even when I decided years ago to “visit” the world and had never really returned yet. Thank you for all these memories that lie dearly in my heart. Always will be your dear Coin-Coin 😊

*My “British Grandma” Sheila, Paul and Colin* – for always welcoming me at “home” with a nice cup of tea, mash peas and fish and chips. Will always always hold close to my heart our time spent together in Ascot!

*My family in Canada, Vietnam and California,* for always welcoming me and for being proud of me. Stories of your generation have taught me that nothing comes without a hard work worth doing! *Petite Mimi & Wang Li,* merci d’être toujours près de Tjé-Tjé 😊

*Mes parents, Papa et Maman,* pour avoir toujours marché à mes côtés qu’importe les distances, pour m’aimer à chaque instant et me prendre dans vos bras quand le monde s’écroule autour de moi. Merci de m’avoir laissé voler très bas et très haut, d’avoir testé mes propres ailes aux caprices du vent de la Vie. Votre amour et soutien ont rendu possible l’impossible! 999 999 999 bisous

Camille
References


37. Mentré, F., Chenel, M., Comets, E. & Grevel, J. Survey on the current use of optimal design approaches and the developments needed in adaptive optimal design for model based analysis performed amongst DDMoRe’s. in *PAGE 21 Abstr 2337* (2012).


60. Keizer, R. J. *et al.* Incorporation of extrapolated concentration data below the limit of quantification in population PK analyses. 5


90. Food and Drug Administration. *FDA Adverse Event Reporting System.*


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy 192

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”.)