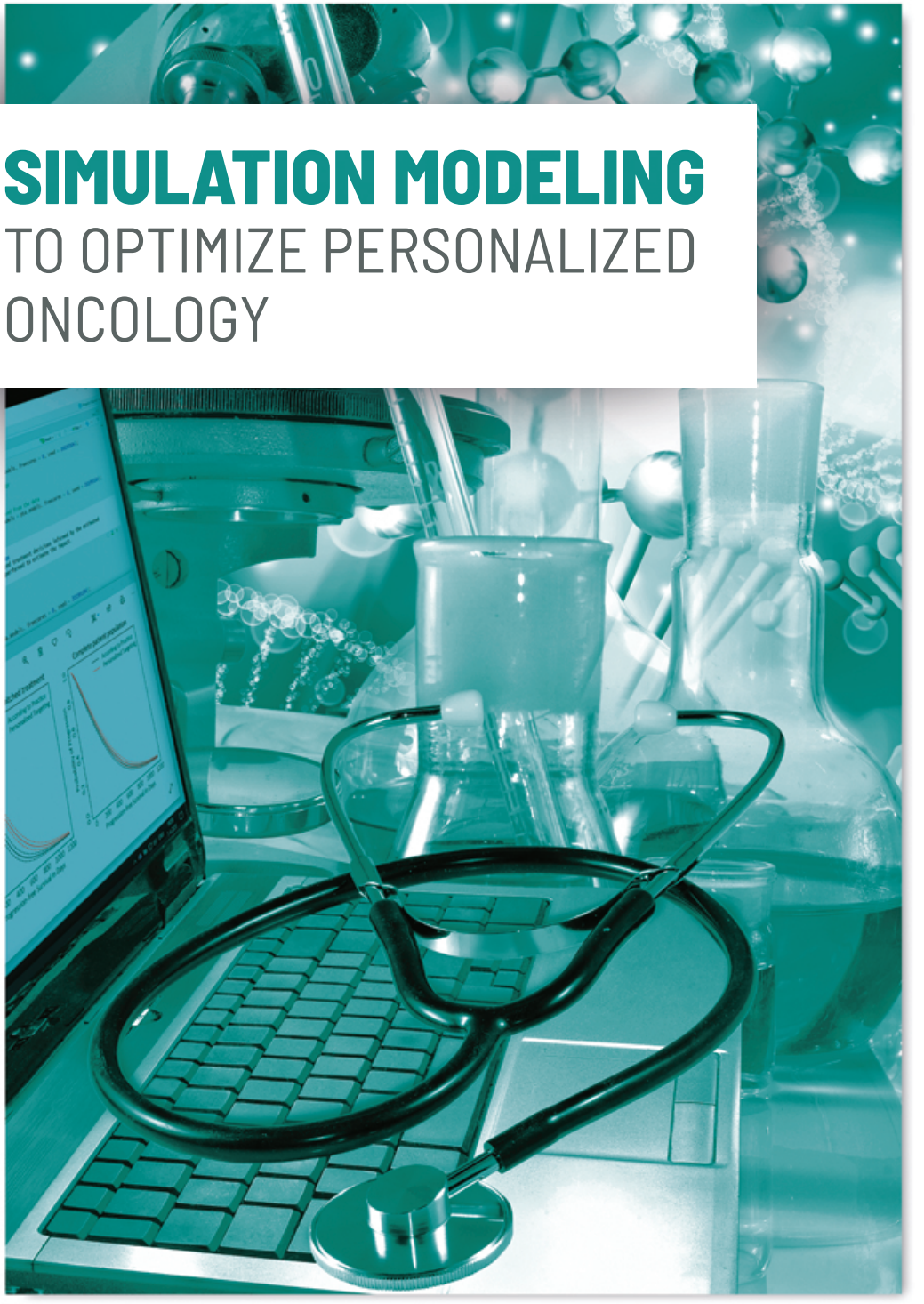
SIMULATION MODELING
TO OPTIMIZE PERSONALIZED ONCOLOGY



KOEN DEGELING

SIMULATION MODELING TO OPTIMIZE PERSONALIZED ONCOLOGY

Koen Degeling

SIMULATION MODELING TO OPTIMIZE PERSONALIZED ONCOLOGY

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. T.T.M. Palstra,
on account of the decision of the Doctorate Board,
to be publicly defended
on Thursday the 4th of July 2019 at 14:45 hours

by

Koen Degeling

born on the 20th of March 1991
in Hengelo, the Netherlands

This thesis has been approved by:

Supervisor:
prof.dr. M.J. IJzerman

Co-supervisor:
dr.ir. H. Koffijberg

Cover design: Jip van Bodegom
Printed by: Ipskamp Printing, Enschede
Lay-out: Koen Degeling
ISBN: 978-90-365-4778-9
DOI: 10.3990/1.9789036547789

This thesis is part of the Health Sciences Series, 19-27, Department of Health Technology and Services Research, University of Twente, Enschede, the Netherlands. ISSN:1878-4968.

©2019: Koen Degeling, Enschede, The Netherlands. All rights reserved. No parts of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author. Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd, in enige vorm of op enige wijze, zonder voorafgaande schriftelijke toestemming van de auteur.

GRADUATION COMMITTEE:

Chairman/secretary	prof.dr. Th.A.J. Toonen	University of Twente
Supervisor	prof.dr. M.J. IJzerman	University of Twente
Co-supervisor	dr.ir. H. Koffijberg	University of Twente
Referee	dr. T.L. Feenstra	University of Groningen
Members	prof.dr. A. Manca	University of York
	prof.dr. A. Briggs	University of Glasgow
	prof.dr. M.M. Rovers	Radboud University
	prof.dr. J.J. Kolkman	University of Groningen
	prof.dr.ir. E.W. Hans	University of Twente
	prof.dr. L.W.M.M. Terstappen	University of Twente

TABLE OF CONTENTS

Chapter 1	General introduction	11

Part 1: Simulation modeling of health and economic outcomes for personalized clinical pathways

Chapter 2	A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models	21
Chapter 3	Matching the model with the evidence: comparing discrete event simulation and state-transition modeling for time-to-event predictions in a cost-effectiveness analysis of treatment in metastatic colorectal cancer patients	43
Chapter 4	Comparison of timed automata with discrete event simulation for modeling of biomarker-based treatment decisions: an illustration for metastatic castration-resistant prostate cancer	59

Part 2: Methodological guidance for developing patient-level simulation models

Chapter 5	Accounting for parameter uncertainty in the definition of parametric distributions used to describe individual patient variation in health economic models	91
Chapter 6	Comparing strategies for modeling competing risks in discrete-event simulations: a simulation study and illustration in colorectal cancer	131
Chapter 7	Comparing modeling approaches for discrete event simulations of competing events based on censored individual patient data: a simulation study and illustration in colorectal cancer	171

Part 3: Metamodeling to reduce computational burden of analyzing simulation models

Chapter 8	A scoping review of metamodeling applications and opportunities for advanced health economic analyses	229
Chapter 9	How to use metamodeling for reducing computational burden of advanced analyses with health economic models: a structured overview of metamodeling methods in a six-step application process	247
Chapter 10	Maximizing the benefits of colorectal screening: from evaluation to optimization of diagnostic strategies using a metamodel	267

Part 4: Simulation modeling to evaluate and optimize targeted treatment strategies

Chapter 11	Simulating progression-free and overall survival for first-line doublet chemotherapy with or without bevacizumab in metastatic colorectal cancer patients based on real world registry data	289
Chapter 12	General discussion	325
Summary		335
Samenvatting		341
Acknowledgements		347
Curriculum Vitae		353
Publications		357

CHAPTER 1

General introduction

GENERAL INTRODUCTION

Improved understanding of cancer disease mechanisms and how disease onset and progression are affected by certain patient characteristics, has resulted in the development of many novel therapies and biomarkers over the last two decades. The treatment landscapes in metastatic cancers now typically comprise different therapies across multiple lines of treatment, with several predictive and prognostic biomarkers available to stratify patients in subpopulations to maximize clinical efficacy and patient health outcomes. Further efforts to explain variation in clinical efficacy and health outcomes based on patient characteristics, and to accentuate consideration of patient and physician preferences into a shared decision making process, have paved the way for personalized treatment decisions aiming to provide patients with the best possible care. Together with advances in the prevention and detection of cancer, these advances in disease management have contributed to improved health outcomes, e.g. in terms of cancer survival [1].

Although technological developments typically add value in terms of clinical efficacy, health outcomes, patient experience, or organizational benefits, they mostly do so at a higher cost compared to existing technologies. In 2009, drug costs were estimated to account for 27% of cancer-related healthcare costs in European countries [2] and prices of cancer drugs continue to increase [3, 4]. Given that a limited amount of healthcare resources needs to be allocated across multiple disease types and alternative interventions in an attempt to optimize clinical outcomes, increasing costs of cancer care together with increasing cancer incidence create a substantial challenge regarding the sustainability of healthcare systems [5]. Research showing a negative correlation between the benefits and costs of cancer drugs [6], suggesting that more expensive treatments are not associated with better outcomes, further highlights the importance of carefully allocating resources across treatment options. Assessment of novel health technologies is essential to inform these resource-allocation decisions, by evaluating healthcare interventions in direct comparisons with other interventions based on clinical efficacy, ethical considerations, cost-effectiveness, budget impact, and social and organizations aspects.

Breakthroughs in cancer medicine and approval and reimbursement of a variety of new cancer therapies, also create complexity with regard to the selection and combination of therapies and biomarkers across multiple lines of treatment [7]. Unless practice is properly guided, this complexity is likely to result in suboptimal use of therapies and biomarkers, inducing real-world practice variation and ineffective and inefficient treatment. This is not only burdensome to society, but also to patients and their families in both health and economic terms, threatening the accessibility of technological developments for future generations [8]. Direct treatment comparisons of healthcare interventions are limited in their ability to address treatment sequencing challenges, as these comparisons typically focus on two or a limited number of alternatives for a specific indication and line of treatment. These comparisons aim to identify the intervention that is to be preferred for a specific setting, without considering use of alternative interventions in successive lines of treatment or how established treatment options in successive lines are affected by changes in disease management. Hence, there is need for translational research to go beyond direct comparisons of specific interventions and

investigate how different therapeutic options can be combined and targeted over multiple lines of treatment based on patient and disease characteristics, including biomarkers.

Randomized clinical studies have been the preferred source of evidence for evaluating healthcare technologies and translational research. However, traditional clinical trial designs are challenged by today's complex clinical context due to their inflexibility, and adaptive trial designs have been proposed and used to assess the efficacy of (larger numbers of) interventions more efficiently [9]. Despite developments in clinical trial designs, real-world data from observational studies have been recognized as valuable source of evidence when results from randomized studies are not available [10]. Translational research in oncology may need to increasingly rely on such real world data, e.g. from clinical registries, as it is unlikely that evidence on clinical efficacy of all relevant treatment combinations and sequences for specific patient subgroups can be obtained from controlled clinical studies only. Furthermore, real-world data may address limitations regarding the external validity of randomized controlled studies and provide insights into the real-world cost-effectiveness of healthcare technologies [11]. However, regardless of its design, a single clinical study rarely provides all evidence required to evaluate a healthcare technology, indicating a need for simulation modeling methods to combine evidence from different (types of) sources [12]. Other reasons for employing simulation modeling methods are, for example, to extrapolate beyond time horizons of clinical studies, to inform and optimize clinical study designs or, more specifically for translational research, to perform what-if scenario analyses.

In order for simulation models to support decision making, they need to adequately represent clinical practice and reflect the true nature of the evidence used to define them, including all types of uncertainty in this evidence [13, 14]. Neglecting to do so may result in biased results and, hence, suboptimal resource allocation and treatment decisions. Given the advances in personalized oncology, simulation modeling studies nowadays are challenged to reflect patient heterogeneity and preference-sensitive treatment decisions on (combinations of) treatments and biomarkers over multiple lines of treatment [15, 16]. Given this complexity, it is being questioned whether application of traditionally used cohort-level simulation modeling methods, such as decision tree analysis and Markov modeling, is still appropriate for cost-effectiveness analyses to inform reimbursement decisions. Although theoretically possible to some extent [17], representing today's dynamic clinical pathways using cohort-level simulation modeling methods will most likely result in (too) complex model structures, especially when representing the level of detail required to inform treatment decisions.

Alternative simulation modeling methods that are able to represent dynamic processes on an individual patient level, have been proposed to model today's dynamic clinical pathways based on patient-level histories and (disease) characteristics [18-20]. Examples of these so-called patient-level simulation or microsimulation modeling methods are patient-level state-transition modeling and discrete event simulation [21, 22]. These methods have increasingly been used in health economic modeling studies, evaluating personalized interventions and informing policy decisions on resource allocation. Due to their ability to model the impact of different treatment sequencing strategies for specific patient subpopulations, these methods

have potential to go beyond traditional evaluations of healthcare technologies and optimize personalized oncology by improving selection and combination of therapies and biomarkers across multiple lines of treatment based on patient and disease characteristics

Although promising in their application, patient-level simulation modeling methods may be perceived as more complex compared to cohort-level methods with regard to several aspects [18, 20]. Because analyses of patient-level simulation models involve simulation of individual patients, these are more computationally demanding compared to analyses of cohort-level models, which may pose challenges regarding the time required to perform these analyses. Also relating to modeling individual pathways, data analyses performed to provide input for patient-level simulation models tend to be more complex than those performed to populate cohort-level models. Furthermore, whereas cohort-level simulation models can typically be implemented in well-known software environments, using spreadsheets for example, patient-level simulation methods like discrete event simulation are often implemented in dedicated simulation software environments that require more technical skills. Although methods such as discrete event simulation provide more flexibility in terms of model structure and evidence format used compared to cohort-level methods, this also implies that more design choices need to be made when implementing these methods. Finally, due to increased model complexity, patient-level simulation models may be more challenging to validate and communicate to stakeholders, complicating review of such models.

Given these challenges regarding their use and, contrariwise, their clear potential to reflect patient heterogeneity and preference-sensitive treatment decisions on (combinations of) therapies and biomarkers over multiple lines of treatment, several questions regarding the use of patient-level simulation modeling methods arise. For example, in what scenarios is use of patient-level simulation modeling methods preferable over their cohort-level equivalents? How should the (type of) evidence be considered in selecting between these two categories of simulation modeling methods? Furthermore, when implementing patient-level simulation models, how can the challenges associated with their use be addressed? Also, which design choices are to be made, which options are there to these choices, and which option is preferably used for a specific scenario? Finally, how complex can these simulation models become so that sufficient confidence in them can still be established for their output to be meaningful to different stakeholders?

This thesis contributes to answering these questions regarding the use of patient-level methods to model personalized oncology pathways. Although these questions are applicable to all patient-level simulation modeling methods in general, focus in this thesis is on discrete event simulation. Results presented in this thesis will contribute to improved understanding regarding the implementation of patient-level methods and increased awareness of their potential to evaluate and optimize personalized oncology pathways. By providing methodological guidance and making source codes and modeling tools publicly available, modelers are equipped with essential resources that empower them to appropriately apply these methods and utilize their full potential.

After this general introduction, this thesis continues with an overview of the current status of simulation modeling in personalized medicine, including comparisons of different simulation modeling methods. A review is performed to investigate which simulation modeling methods have been used for simulating health and economic outcomes of personalized clinical pathways and the extent to which challenges regarding simulation modeling in this context have been addressed (Chapter 2). Based on a case study in metastatic colorectal cancer, a comparison of state-transition modeling and discrete event simulation is made to inform selection between these two methods, where focus is on the structure of the evidence to be used (Chapter 3). Additionally, a case study in metastatic castration-resistant prostate cancer is performed to compare discrete event simulation to timed automata, a modeling paradigm that originates from computer science and has not yet been applied for modeling health and economic outcomes (Chapter 4).

Following the discussion of previous applications and selection of simulation modeling methods for estimating health and economic outcomes of personalized clinical pathways, the second section of this thesis provides essential methodological guidance for the development of patient-level simulation models. Different approaches to account for parameter uncertainty (i.e., second-order uncertainty) in any type of patient-level simulation model are illustrated and compared based on a simulation study and a case study in metastatic colorectal cancer (Chapter 5). Regarding discrete event simulation specifically, one advantage of this method is that it provides great flexibility with respect to how competing events are implemented and two studies are performed to provide guidance on selecting between different approaches to do so. First, a study is presented that compares different approaches to implement competing events based on uncensored individual patient data (Chapter 6). Second, different approaches are compared for studies informed by censored individual patient data (Chapter 7). Both studies involve a simulation study and a case study based on data from a randomized controlled metastatic colorectal cancer trial.

Because patient-level simulation models tend to be more computationally demanding compared to their cohort-level equivalents, performing analyses that require many simulation evaluations, such as probabilistic sensitivity analyses and value of information analyses, may not be possible within feasible timeframes. To address this challenge, the third section of this thesis discusses the use of metamodeling methods to reduce computational burden of such advanced analyses. Previous studies utilizing metamodeling methods in combination with health economic simulation models are reviewed to identify potential applications and methods used (Chapter 8). Following an identified need for increased awareness of the potential of, and guidance for applying metamodeling in a health economic context, an overview of metamodeling methods and directions for their selection and use are provided (Chapter 9). To further illustrate their potential, a metamodel is developed to address computational issues with a colorectal cancer screening model, enabling optimization techniques to be applied to identify an optimal screening strategy for the Dutch population (Chapter 10).

Illustrating how simulation modeling methods can be used to optimize treatment targeting in personalized oncology, guidance provided in this thesis is used to develop a discrete event

simulation model that estimates clinical outcomes of different treatment strategies for metastatic colorectal cancer (Chapter 11). This model is developed based on real-world data from a clinical metastatic colorectal cancer registry and shows that clinical outcomes may be improved substantially by providing selected patients with alternative, already available treatments based on patient and disease characteristics.

Finally, this thesis concludes with a discussion of its findings, presents remaining challenges and directions for further research, and summarizes its main conclusions (Chapter 12).

REFERENCES

1. De Angelis R, et al. Cancer survival in Europe 1999-2007 by country and age: results of EUROCARE-5-a population-based study. Lancet Oncol. 2014;15(1):23-34.

2. Luengo-Fernandez R, et al. Economic burden of cancer across the European Union: a population-based cost analysis. Lancet Oncol. 2013;14(12):1165-74.

3. Savage P, Mahmoud S. Development and economic trends in cancer therapeutic drugs: a 5-year update 2010–2014. Br J Cancer. 2015;112(6):1037.

4. Howard DH, et al. Pricing in the market for anticancer drugs. J Econ Perspect. 2015;29(1):139-62.

5. Sullivan R, et al. Delivering affordable cancer care in high-income countries. Lancet Oncol. 2011;12(10):933-80.

6. Del Paggio JC, et al. Delivery of meaningful cancer care: a retrospective cohort study assessing cost and benefit with the ASCO and ESMO frameworks. Lancet Oncol. 2017;18(7):887-94.

7. Messner DA. Evaluating the comparative effectiveness of treatment sequences in oncology: a new approach. J Comp Eff Res. 2015;4(6):537-40.

8. Wait S, et al. Towards sustainable cancer care: reducing inefficiencies, improving outcomes—a policy report from the All.Can initiative. J Cancer Policy. 2017;13:47-64.

9. Pallmann P, al. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med. 2018;16(1):29.

10. Hershman DL, Wright JD. Comparative effectiveness research in oncology methodology: observational data. J Clin Oncol. 2012;30(34):4215-22.

11. Katkade VB, et al. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. J Multidiscip Healthc. 2018;11:295-304.

12. Buxton MJ, et al. Modelling in ecomomic evaluation: an unavoidable dact of life. Health Econ. 1997;6(3):217-27.

13. Briggs AH, et al. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group–6. Medl Decis Making. 2012;32(5):722-32.

14. Miller JD, et al. Current challenges in health economic modeling of cancer therapies: a research inquiry. Am Health Drug Benefit. 2014;7(3):153-62.

15. Annemans L, Redekop K, Payne K. Current methodological issues in the economic assessment of personalized medicine. Value Health. 2013;16(S6):S20-S6.

16. Phillips KA, et al. Challenges to the translation of genomic information into clinical practice and health policy: Utilization, preferences and economic value. Curr Opin Mol Ther. 2008;10(3):260-6.

17. Zheng Y, et al. Modeling treatment sequences in oharmacoeconomic models. Pharmacoeconomics. 2017;35(1):15-24.

18. Brennan A, et al. A taxonomy of model structures for economic evaluation of health technologies. Health Econ. 2006;15(12):1295-310.

19. Marshall DA, et al. Transforming healthcare delivery: integrating dynamic simulation modelling and big data in health economics and outcomes research. Pharmacoeconomics. 2016;34(2):115-26.

20. Stahl JE. Modelling methods for pharmacoeconomics and health technology assessment. Pharmacoeconomics. 2008;26(2):131-48.

21. Karnon J, et al. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–4. Med Decis Making. 2012;32(5):701-11.

22. Siebert U, et al. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–3. Med Decis Making. 2012;32(5):690-700.

PART 1

Simulation modeling of health and economic outcomes for personalized clinical pathways

CHAPTER 2

A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models

Koen Degeling
Hendrik Koffijberg
Maarten J. IJzerman

This chapter has been published as:
Degeling K, Koffijberg H, IJzerman MJ. A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models. Expert Review of Pharmacoeconomics & Outcomes Research. 2017;17(1):17-25.

ABSTRACT

Introduction: The ongoing development of genomic medicine and the use of molecular and imaging markers in personalized medicine (PM) has arguably challenged the field of health economic modeling (HEM). This study aims to provide detailed insights into the current status of HEM in PM, in order to identify if and how modeling methods are used to address the challenges described in literature.

Areas covered: A review was performed on studies that simulate health economic outcomes for personalized clinical pathways. Decision tree modeling and Markov modeling were the most observed methods. Not all identified challenges were frequently found, challenges regarding companion diagnostics, diagnostic performance, and evidence gaps were most often found. However, the extent to which challenges were addressed varied considerably between studies.

Expert commentary: Challenges for HEM in PM are not yet routinely addressed which may indicate that either (1) their impact is less severe than expected, (2) they are hard to address and therefore not managed appropriately, or (3) HEM in PM is still in an early stage. As evidence on the impact of these challenges is still lacking, we believe that more concrete examples are needed to illustrate the identified challenges and to demonstrate methods to handle them.

INTRODUCTION

With the advent of personalized medicine (PM), the delivery of health care is shifting towards selecting and monitoring the best available treatment for each individual patient based on patient characteristics and diagnostic information. More specifically, 'Personalized medicine seeks to improve stratification and timing of health care by utilizing biological information and biomarkers on the level of molecular disease pathways, genetics, proteomics as well as metabolomics' [1]. Personalized clinical processes typically involve multiple diagnostic tests and treatments over time, and the sequences of tests and treatments may differ between individual patients. Furthermore, treatment decisions are becoming increasingly preference sensitive [2], due to the fact that there is no longer a 'one-size-fits-all' approach and the need to appraise the combined information from multiple sources, such as numerous test results, patients' characteristics, and medical histories. Examples of personalized clinical processes include treatment targeting based on risk stratification by patient characteristics and response monitoring using biomarkers [3,4]. The shift towards more interactive and dynamic, and therefore more complex, clinical treatment processes is associated with challenges not only regarding the delivery of health care, but also regarding the health economic evaluation of medical technologies [5].

For instance, the use of randomized controlled trials (RCTs) for collecting evidence, and to inform health economic evaluations, is increasingly being questioned in a PM context [5–9]. RCTs are designed to draw conclusions on a population-level, while PM focusses on patient-level outcomes. The resulting data gaps characterize the increasingly common work setting in which evaluations of health care interventions need to be performed, initiating the need for accumulating other types of data, e.g. observational (big) data, or expert elicitation techniques [10,11]. Furthermore, new challenges occur with respect to analyzing trial data at the individual level and identifying and handling multiple subgroups [11]. Therefore, health economic evaluation in the context of PM increasingly relies on modeling approaches and new methods, such as dynamic simulation modeling and the
use of machine learning or other statistical approaches [12].

However, PM also challenges the field of health economic modeling (HEM), as there is a need for models to accurately capture the interactions and dynamics present in personalized clinical processes [13,14]. Annemans et al. [13] report on 10 methodological challenges that need to be considered when 'designing and conducting robust model-based economic evaluations in the context of personalized medicine'. In general, these challenges raised by Annemans et al. focus towards the need to appropriately represent the dynamics of personalized treatment decisions with a wealth of diagnostics and surrounding uncertainty. More specifically, these challenges can mostly be translated into appropriate handling of the diagnostic performance of tests, combinations of tests, greater uncertainty due to more complex analysis, and data gaps. Phillips et al. [15] complement these challenges by taking into account patients' and physicians' preferences, patients' characteristics (such as age, gender, comorbidities, and medical history), and considering the impact of drug therapies and companion diagnostics simultaneously. Other challenges relate to the absence of implemented guidelines, criteria, and standards for the evaluation of new technologies in PM [16,17].

Given these methodological challenges, the appropriateness of the commonly used modeling method for HEM, i.e. Markov modeling, is being questioned [14,18], as this modeling approach may not be able to fully capture the complex treatment processes associated with PM [19–22]. Consequently, the use of more advanced modeling methods, such as discrete event simulation, agent based modeling, and system dynamics, might be more appropriate in this personalized context [11,23–25].

It is unknown how current models address these challenges reported by experienced modelers and which modeling methods are being used to do so. Therefore, we assess the level of support for the methodological challenges described in literature by (1) identifying if and how the methodological challenges regarding modeling in the context of PM are being addressed, (2) exploring the different modeling approaches in PM in use to date, and (3) determining which alternative modeling methods may be appropriate to handle the specific issues in PM. Although several reviews have been published on modeling in PM [26–29], these focus on patient stratification using pharmacogenetics. We contribute to this literature by researching the challenges for HEM in PM in general, including patient stratification by other means than pharmacogenetics, e.g. the use of imaging technologies or risk stratification by patient characteristics.

LITERATURE REVIEW

We performed a search in PubMed, employing primary search terms on Personalized Medicine and Precision Medicine combined with additional search terms on modeling and simulation, both in the title or the abstract of the publication. The search strategy was further specified by adding well-known key words used in health economics. The exact search algorithm can be found in Appendix 1. The narrow primary search terms were required to include all sorts of patient stratification, e.g. pharmacogenetics, imaging technologies, and stratification by patient characteristics. No specific start date was applied and the search was updated until the 27 November 2015. The final sample was enriched by cross-referencing to include as many relevant publications as possible [30].

To maintain the broad perspective of this review, and therefore prevent erroneous excluding of publications, only duplicates and animal studies were removed from the initial search results. The unique sample was first assessed based on title and abstract by one reviewer (KD). Next, screening the full text of the remaining publications resulted in inclusion into, or exclusion from, the final sample. Only publications relating to HEM studies in PM were included for full text screening. Publications were considered to meet the PM criterion when tests or prediction models were used to stratify patients into subgroups, for screening or targeting purposes, or when tests were used to monitor treatment effectiveness and thereby support patient-level treatment decisions. The refined sample was enriched by cross-referencing. Cross-references were included based on full text screening and until theoretical saturation was reached when inclusion of additional publications did not result in additional insights in a specific disease area, for example the screening for breast cancer. A second reviewer was consulted if it was unclear whether a publication should be included or excluded. The final

decision to include or exclude a publication was made based on consensus between authors KD and HK. Reasons for exclusion were categorized into not mutually exclusive categories, as specified in Appendix 2.

Table 1. *Checklist presenting the main challenges described in literature and used for analysis of the final sample. A positive score on challenges 1-7 indicates that the corresponding challenge is tackled in the publication. A positive score in challenges 8-10 indicates that the authors mention they experienced difficulties regarding the corresponding challenge. Negative scores indicate that the corresponding challenge is not tackled or mentioned by the authors.*

	Challenge	**Identification**
1.	Modeling patient-level processes	*Is the model defined on a patient level?*
2.	Modeling patients' preferences	*Are patients' preferences modeled to take their effect on the outcomes into account?*
3.	Modeling physicians' preferences	*Are physicians' preferences modeled to take their effect on the outcomes into account?*
4.	Taking into account the diagnostic performance of tests	*Is the effect of the sensitivity, specificity, positive predictive value, and/or negative predictive value on the outcomes taken into account?*
5.	Modeling combinations of tests	*Does the modeled process includes combinations of tests and/or prediction models?*
6.	Modeling companion diagnostics	*Does the modeled process includes combinations of test(s) and treatment(s)?*
7.	Study-specific outcome measures	*Does the modeled process includes study-specific outcomes, such as disease-specific adverse events?*
8.	Data gaps	*Do the authors mention any evidence gaps? If so, do they mention that these evidence gaps exist due to stratification of patients based on risk models and/or test results?*
9.	Greater uncertainty due to more complex analysis	*Do the authors mention greater uncertainty with respect to the outcomes, due to more complex analysis, as a result of personalization of the model?*
10.	Absence of guidelines	*Do the authors mention any difficulties related to the absence of guidelines for HEM in the context of PM?*

SCORING CHECKLIST REPRESENTING THE MAIN CHALLENGES

In order to extract the data from the final sample, the general study characteristics and information on the used modeling method were summarized first. This includes the target disease, the description on how the treatment process was personalized, the used modeling method, the model structure, and the performed analyses. The treatment processes were characterized by the purpose of the stratification (screening, targeting, or monitoring) and whether the stratification was prognostic or predictive [31]. Screening is referred to as the process of diagnosing a patient with a specific disease, whereas treatment targeting relates to selecting a treatment, from a set of treatment options that is expected to be most beneficial to a specific patient, based on patient-specific characteristics or diagnostic information. Treatment monitoring concerns the process of repeatedly assessing a patient's response to the current treatment, in order to stop this treatment or switch to another treatment, if the patient is not benefiting from the current treatment.

Next, a checklist presenting the main methodological challenges for HEM in the context of PM was developed using the literature, where the publications by Annemans et al. [13] and Phillips et al. [15] served as the reference. After reviewing and classifying relevant papers, 10 different items in the checklist were used for the analysis of the final sample, based on the challenges derived from literature [13,14,17,20] (Table 1). The checklist was used to highlight the challenges in the final sample. For the first seven items in the checklist a positive result indicates that the authors addressed the challenge in the model, whereas for the remaining items eight to ten a positive result indicates that the authors identified that specific challenge and reported this in the publication. When a challenge was addressed or identified by the authors, this was scored as '+', when challenges were not addressed or identified this was scored as '–'. If applicable, additional information on the extent to which challenges were addressed or mentioned was recorded.

RESULTS

The search strategy yielded 2245 publications on PubMed (Figure 1). From this initial sample, five duplicates and 775 animal studies were excluded. The abstracts of all remaining 1465 publications were read and finally resulted in the exclusion of 1442 publications. These excluded publications were predominantly publications of a qualitative nature (n = 534, 37%) and experimental nature (n = 421, 29%). Several other excluded publications were publications on mathematical or statistical models developed to support medical decision-making (n = 168, 12%). A total of 128 publications were excluded from the final sample because they covered adjacent topics, such as personalized speaking tools or personalized computer systems (9%). The detailed reasons for exclusion are presented in Appendix 2. After exclusion based on title and abstract, 23 publications remained. From these publications the full text was screened, resulting in five more exclusions from the sample for different reasons (i.e. papers presenting a biomedical model [32], a prediction model [33], or that did not include diagnostics or personalized risk estimations [34– 36]. Analyzing the full text of the included articles resulted in an enrichment of the sample with 13 additional publications by cross-referencing. Many candidate cross-reference articles were not included due to the absence of

diagnostics or personalized risk estimations [37–46]. Finally, a total of 31 publications were available for analysis [47–77].

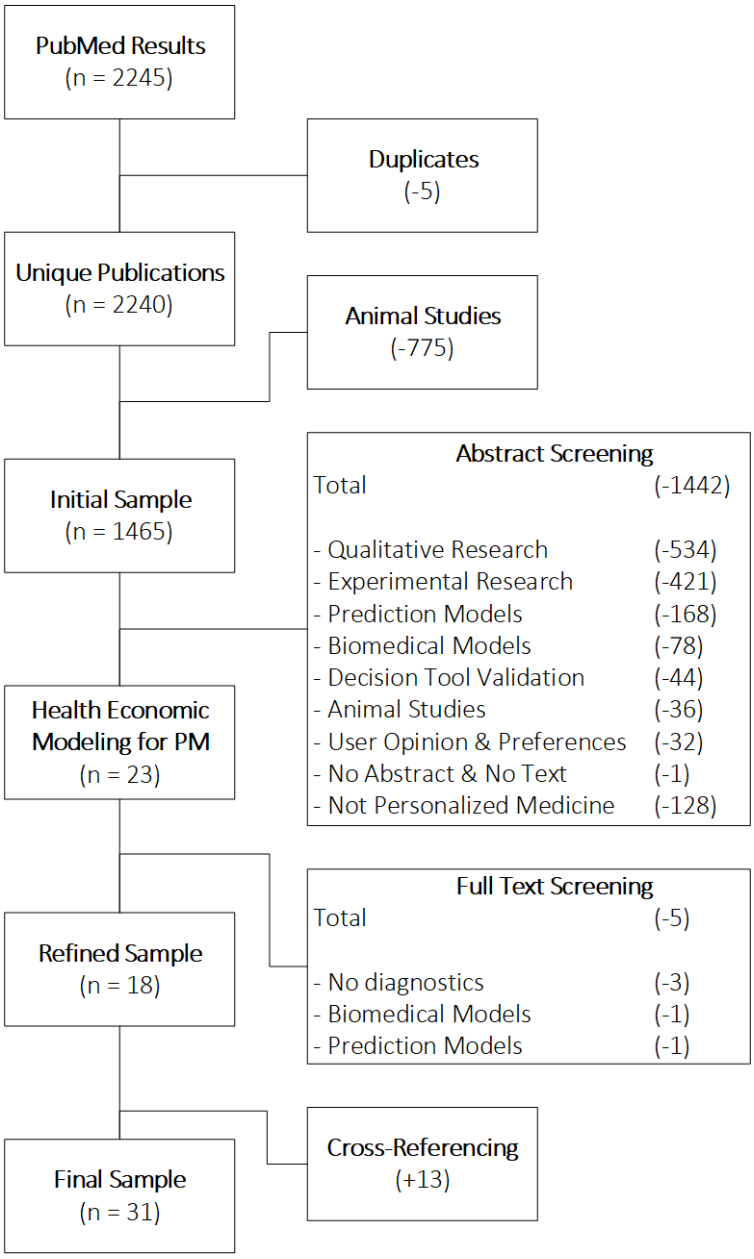


Figure 1. *Graphical representation of the search- and selection process.*

Table 2. *Summary of the study characteristics of the final sample of publications.*

Ref.	Disease	PM Process[2]	Modeling Method[3]
[47][1]	Breast Cancer	Scre. (Prog.)	POMDP
[48][1]	Breast Cancer	Scre. (Prog.)	Statistical Model (Math.)
[49]	Human Immunodeficiency Virus	Targ. (Pred.)	DT
[50]	Alzheimer's Disease	Targ. (Pred.)	MM (12, 30)
[51]	Atrial Fibrillation	Targ. (Pred.)	DT + MM (1, LT)
[52]	Cardiovascular Disease	Targ. (Pred.)	MSM (12, LT)
[53]	Neonatal Disease	Targ. (Pred.)	MM (6, 30)
[54]	Coronary Syndrome	Targ. (Pred.)	DT + MM (1, LT)
[55]	Lung Cancer	Targ. (Pred.)	DT
[56]	Coronary Artery Disease	Targ. (Pred.)	DT
[57][1]	Human Immunodeficiency Virus	Targ. (Pred.)	DDCM
[58][1]	Colorectal Cancer	Scre. (Prog.)	DT + MM (-, LT)
[59]	Prostate Cancer	Scre. (Prog.)	MM (12, 30)[6]
[60]	Rheumatoid Arthritis	Targ. (Pred.), Mon. (Prog.)	MSM (3, 3)
[61]	Colorectal Cancer	Scre. (Prog.), Targ. (Pred.)	DT + MM (12, age 100)
[62][1]	Coronary Artery Disease	Scre. (Prog.)	MSM (12, LT)
[63]	Acute Myeloid Leukemia	Targ. (Pred.), Mon. (Prog.)	DES
[64][1]	Lung Cancer	Targ. (Pred.)	DT
[65][1]	Lung Cancer	Targ. (Pred.)	DT
[66][1]	Hepatitis C Virus	Targ. (Pred.)	DT + MM (6, LT)
[67][1]	Colorectal Cancer	Scre. (Prog.)	DT
[68]	Breast Cancer	Scre. (Prog.)	Cost Equation (Math.)
[69]	Depressive Disorder	Targ. (Pred.), Mon. (Prog.)	MM (3, 3)
[70]	Hepatitis C Virus	Targ. (Pred.), Mon. (Prog.)	MM (12, LT)
[71][1]	Colorectal Cancer	Scre. (Prog.)	DT
[72][1]	Breast Cancer	Scre. (Prog.)	Multiple MSM (-, LT)
[73][1]	Colorectal Cancer	Scre. (Prog.)	DT
[74]	Breast Cancer	Scre. (Prog.)	MM (1, LT)[7]
[75]	Type 2 Diabetes	Targ. (Pred.)	DT + MM (12, 5)
[76][1]	Acute Myeloid Leukemia	Targ. (Pred.)	DT
[77]	Breast Cancer	Scre. (Prog.)	Prob. Function (Math.)

[1]Obtained by cross-referencing.

[2]Scr.: screening; Targ.: targeting; Mon.: monitoring; Pred.: predictive; Prog.: prognostic.

[3]DT: decision tree; MM (. . ., . . .): Markov model (cycle length in months, time horizon in years); MSM (. . ., . . .): Microsimulation model (cycle length in months, time horizon in years); POMDP: partially observable Markov decision process; DDCM: deterministic dynamic compartment model; DES: discrete event simulation; Math.: mathematical; LT: life-time.

[4]HS: health states; PS: process states; BS: behavior states; TR: Test Results.

Model Structure[4]	Analyses[5]
HS, TR	QALY, DRO, DSA
HS	CEA (LE, LY, QALY), DRO, DSA
HS, PS, BS	NNT, CEA (QALY), DSA, PSA
HS	CEA (QALY), CEAC, EVPI, EVPPI, DSA, PSA
HS	CEA (QALY), CEAC, DSA, PSA
HS	TLE, DRO, PSA
HS	CEA (LY, QALY), CEAC, NMB, DSA, PSA
HS	CEA (QALY), DSO, DSA, PSA
TR, PS	CEA (QALY), CEAC, DSA, PSA
TR, PS	Costs, Budget Impact, PSA
HS	CEA (QALY), DRO, DSA
HS, TR	CEA (LY), DSA
HS, TS	CEA (QALY), DSA
HS	CEA (QALY), CEAC, DSA, PSA
HS, TR	CEA (LY), CEAC, DSA, PSA
HS, TR	CEA (QALY), DRO, DSA
PS	DRO, OS, DFS, PSA
HS, TR	CEA (LY), DSA
HS, TR	CEA (QALY), DSA
HS, TR	CEA (QALY) DSA, PSA
TR	CEA (LY, QALY), DSA
-	Costs, PSA
HS, TR, PS	CEA (QALY), DSA
HS	CEA (LY, QALY), CEAC, DSA, PSA
TR, BS	CEA (LY), DSA, PSA
HS, TR	DRO, LY, QALY, DSA
TR	CEA(Case Detected), DSA
HS	CEA (QALY), NNT, DSA, PSA
HS, TR	CEA (QALY), CEAC, DSA, PSA
HS, TR	QALY, DSA
HS	CEA (QALY), LE, DSA

[5]NNT: Number Needed to Treat; CEA(. . .): Cost-effectiveness Analysis (effect. outcome); QALY: quality adjusted life years; LE: lives extended; LY: life years; DSA: deterministic sens. analysis; PSA: probabilistic sens. analysis; CEAC: cost-effectiveness acceptability curve; EVPI: expected value of perfect information; EVPPI: expected value of partial perfect information; TLE: total life expectancy; DRO: disease related outcomes; NMB: net monetary benefit; OS: overall survival; DSF: disease-free survival.

[6]A cycle length of 24 months was also used. Several other time horizons were included: 25, 20, 15 and 10 years.

[7]Cycle lengths of 24 months and 36–48 months were also used.

The final sample concerned studies in various disease areas, including oncology (n = 17), cardiovascular disease (n = 5), human immunodeficiency virus (n = 2), hepatitis C virus (n = 2), Alzheimer's disease (n = 1), neonatal disease (n = 1), rheumatoid arthritis (n = 1), depressive disorder (n = 1), and type 2 diabetes (n = 1) (Table 2). Most studies in the sample stratify patients for targeted therapy (n = 19, 61%), of which four studies combine treatment targeting with response monitoring and one article combines treatment targeting with screening. In total 13 studies stratify patients for screening purposes (42%). No articles used tests for response monitoring only. This indicates that patient stratification is used only for predictive purposes in 45% of the included publications (n = 14), for prognostic purposes in 39% of the included publications (n = 12), and for both predictive and prognostic purposes in 16% of the included publications (n = 5).

The most frequently used modeling methods are decision tree modeling (n = 15, 48%) and Markov modeling (n = 12, 39%), which are often combined (n = 6, 19%). Other frequently used modeling methods are microsimulation modeling (n = 4, 13%), and mathematical modeling by equations (n = 3, 10%). Other observed modeling methods include partially observable Markov decision process (POMDP) modeling, deterministic dynamic compartment modeling (DDCM), and discrete event simulation (DES). Whereas decision trees, Markov models, and DDCM [78] simulate cohorts of patients, POMDP modeling [79], microsimulation modeling [80], and DES [81] are used for patient-level simulations. As Table 3 presents, there is not only an increase in publications over time, but also an increase in the use of modeling methods other than decision tree analysis and Markov modeling. Regarding the motivation of the used modeling methods, one article mentions the straightforward interpretation of a decision tree as reason for its use [49], whereas another article, using a Markov model, suggests that it may not be appropriate to model a screening process as a homogenous process [59].

Table 4 shows that not all methodological challenges as included in the checklist were systematically addressed or identified and reported in the final sample. The challenges regarding physicians' preferences, disease-specific outcome measures, greater uncertainty, and absence of guidelines are addressed or identified and reported in at most five of the included publications. Yet, the challenges regarding the patients' preferences, diagnostic performance, multiple tests, companion diagnostics, and the lack of evidence are more frequently

Table 3. *Summary of the distribution of the publications and the use of cohort modeling or alternative modeling methods over time.*

	Publications		Cohort Modeling[1]		Alternative Methods	
Time Period	#	%[2]	#	%[3]	#	%[3]
≤ 2005	4	13	4	100	0	0
2006- 2010	5	16	4	80	1	20
≥ 2011	22	71	13	59	9	41

[1]*Cohort modeling was defined as decision tree analysis and Markov modeling.*
[2]*As a percentage of all publications included in the final sample.*
[3]*As a percentage of the publications in the corresponding time period.*

addressed or identified and reported (in at least 10 publications). However, the latter result needs to be perceived in relative terms, as the extent to which challenges are addressed varies considerably between studies. For example, several studies modeled multiple tests, but assumed a fixed sequence of these tests. Although in practice, test sequences might be dynamic and thereby influence the diagnostic performance of the process as a whole. Furthermore, of the 19 publications in which a lack of evidence is identified, these data gaps are not purely caused by stratification in 37% of the cases (n = 7).

CONCLUSION

From the articles that were included, it can be concluded that patient stratification, using diagnostics or risk models, is mostly performed for treatment targeting or screening purposes. Only in few of the included publications diagnostics are used for treatment monitoring. Overall, the most frequently observed modeling methods are decision tree analysis and Markov modeling, which is in line with literature [14,18]. However, an increase in the use of more advanced modeling methods is observed. Finally, the results show that the methodological challenges for HEM in the context of PM described in literature are not (yet) frequently addressed or identified and reported.

EXPERT COMMENTARY

The findings from the literature study can have different implications. For instance, they may indicate that the impact of the challenges for HEM in PM is less severe than expected, that the challenges are hard to address and there is a lack of methods to overcome the challenges, or that we are still in an early stage of personalization and that the complexity of PM, therefore, is not yet a major issue. The observation that the most frequently used modeling methods are still cohort models may also indicate that we are still in a premature stage. Currently, it seems to be sufficient to stratify patients into relatively large subgroups, as there is also no regulatory incentive to further personalize these models.

However, treatment decisions are becoming increasingly complex, involving multiple biomarkers or panels of markers from next-generation sequencing to feed a sequence of clinical decisions. In this context, patient-level models are likely to become standard, as cohort models can no longer reflect the dynamic treatment processes in heterogeneous subgroups of individuals and may lead to biased estimates of the impact of new technologies. Representing the variation in patients' clinical pathways is particularly problematic for cohort models, as this would either require numerous separate cohort models, representing all plausible sequences, or one very large model including all these sequences. In both cases, however, models will become substantially complex to manage and models' cognitive ease will decrease dramatically. Conversely, more advanced modeling methods can represent the dynamics of individual pathways in a straightforward and more natural manner.

Policymakers will need to incentivize the use of appropriate modeling approaches to accurately represent clinical practice and accept that this might result in more complex health

Table 4. *Results of the analysis on whether the challenges for HEM in PM are addressed in the final sample of publications.*

	Addressed in model			
Reference	**Patient-level**	**Patients' preferences**	**Physicians' preferences**	**Diagnostic performance**
[46]	+	-	-	+
[47]	-	-	-	-
[48]	-	+[2]	-	-
[49]	-	-	-	+[3]
[50]	-	-	-	+[3]
[51]	+	+[2]	-	-
[52]	-	-	-	+[3]
[53]	-	-	-	-
[54]	-	-	-	+[6]
[55]	-	-	-	+
[56]	-	+[2]	-	-
[57]	-	+[7]	+[2]	+
[58]	-	-	-	-
[59]	+	-	-	-
[60]	-	+[2]	-	+
[61]	+	-	-	+
[62]	+	-	-	-
[63]	-	-	-	+
[64]	-	-	-	+[6]
[65]	-	-	-	+[6]
[66]	-	+	-	+
[67]	-	+[7]	-	-
[68]	-	-	-	+[6]
[69]	-	-	-	-
[70]	-	+[7]	-	+
[71]	+	+[2]	-	+
[72]	-	+[7]	-	+
[73]	-	-	-	+
[74]	-	-	-	-
[75]	-	-	-	+
[76]	-	-	-	+
Total positive score	19%	32%	3%	65%

[1]*Publication obtained by cross-referencing.*
[2]*Only adherence taken into account as patients' preferences.*
[3]*Diagnostic performance taken into account, but 100% sensitivity and specificity assumed.*
[4]*Evidence gap exists in general, not due to PM.*

			Identified and reported		
Multiple tests	Companion diagnostics	Study spec. Outcomes	Data Gaps	Uncertainty	Guidelines
+[5]	-	+	+	-	-
+[5]	+	-	+[4]	-	-
-	+	-	-	-	-
-	+	-	+[4]	+[4]	-
-	+	-	-	-	-
-	+	+	+[4]	-	-
-	+	-	+[4]	-	-
-	+	-	+	-	-
-	+	-	+	-	-
+[5]	+	-	-	-	-
-	+	-	-	-	-
-	-	-	-	-	-
-	-	-	+	-	-
+[5]	+	-	+	-	-
+[5]	+	-	+[4]	-	-
+[5]	+	+	+	-	-
-	+	-	-	-	-
-	+	-	-	-	-
-	+	-	+	-	-
+[5]	+	-	-	-	-
+[5]	-	-	+	-	-
-	-	-	-	-	-
-	+	-	+	-	-
-	+	-	+	-	-
+[5]	+	-	+[4]	-	-
-	-	-	-	-	-
+[5]	-	-	-	-	-
-	-	-	+	-	-
+[5]	+	-	+	-	-
-	+	-	+[4]	-	-
-	-	-	-	-	-
35%	71%	10%	61%	3%	0%

[5]*Multiple tests are taken into account, but fixed sequence of tests is assumed.*
[6]*Chance on positive test result is taken into account as diagnostic performance.*
[7]*Acceptance of test or treatment is taken into account as the only patients' preferences.*

economic models, possibly at the expense of these models' cognitive ease. It is necessary to accept this increase in complexity, as health economic models are likely to become biased and may lose their value in supporting decision-making when they are not matched with the dynamics and complexity of current and future clinical processes.

That the challenges are present, but may be hard to address using current approaches, is illustrated by the fact that many authors do recognize the challenges described in literature, but do not actually address them in the corresponding models. For example, the relevance of shared decision making is highlighted in several publications, as authors argue the need for physicians to provide patients with personalized information on expected treatment outcomes and to involve these patients in decision-making [36,52,70,72,75]. The observation that this interactive and complex decision-making process is not yet integrated into the corresponding models, however, illustrates the challenge to further personalize these models. Another example can be found in breast cancer screening, as individualization of the screening process is considered very important [37,82], whereas all included models are still cohort-based.

Although patient-level simulation seems appropriate, the present review insufficiently can conclude the usefulness of modeling techniques beyond cohort models to adequately represent personalized clinical processes. Yet, it seems obvious that modeling patients on an individual level is desirable and advanced methods, such as microsimulation modeling and discrete event simulation, may be essential. However, to utilize the full potential of these advanced modeling methods, more evidence, for example on subgroup-specific event rates, costs, and quality of life, may be required compared to less advanced methods. Since the results show that evidence gaps are a frequently experienced challenge, the use of advanced methods may therefore not always be feasible to address the challenges associated with PM.

The results of this study provide insights into the level of support in modeling practice for the methodological challenges described in literature. These findings can be used to focus development of a framework for HEM in this context, as they indicate which challenges require additional guidance the most. Additional further research is recommended to investigate the impact of the challenges for HEM in PM when truly personalized clinical processes are being modeled and whether this will affect decisions made by policy makers. Furthermore, whether the use of advanced modeling methods indeed can solve at least some of these challenges, needs further investigation while being weighed against the increased complexity of these models.

Our study has certain limitations. First of all, it is interesting to note that the number of papers that were finally selected from the initial search results is relatively low, whereas PM is a growing field of research and the health economic issues related to various targeted drugs receive global recognition. This might be partly caused by the strict inclusion criteria required for the research objectives and by the fact that we only consulted one database. However, we minimalized the effect of this limitation by cross-referencing and found that the results for the publications obtained by cross-referencing to be in line with the publications in the initial

sample. Furthermore, underreporting may have resulted in the absence of mentioned challenges for HEM in PM and the lack of motivation for the used modeling methods. This also relates to the limited space that is available in peer-reviewed journals for reporting methods and results.

FIVE-YEAR VIEW

The continuing personalization of clinical pathways highlights the need for using more advanced modeling methods to accurately represent the complex context of clinical practice and, therefore, to be meaningful for supporting decision-making. If not already, regulatory agencies will need to critically review the modeling methods that are being used to translate clinical practice into health economic models, initiating the need for using appropriate modeling methods. Furthermore, efforts to illustrate and guide the use of more advanced modeling methods, such as discrete event simulation and system dynamics modeling, will strengthen and spread the knowledge base of these methods, increasing their use for the evaluation of health care interventions, also by clinical experts. Finally, the pharmacoeconomic and clinical communities will realize that using more advanced modeling methods does not necessarily result in more complex models. On the contrary, using more advanced modeling methods will enable models to continue being valuable for what they are needed for, which is supporting decisions in an increasingly complex environment.

KEY ISSUES

- Health economic models in personalized medicine are to a large extend based on cohorts of patients instead of individual patients.
- Some of the specifics of Health Economic Modeling in Personalized Medicine, such as patient-level models and the representation of the dynamics and sequence of preference-sensitive clinical decisions, are still not addressed appropriately.
- This may indicate that the impact of these challenges is less severe than expected, that the challenges are hard to address and there is a lack of methods to overcome the challenges, or that we are still in an early stage of personalization and that the complexity of Personalized Medicine, therefore, is not yet a major issue.
- Further research should focus on identifying the extent of these challenges when truly personalized processes are being modeled and the added value of advanced modeling methods in this context.

DECLARATION OF INTEREST

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

CONTRIBUTION STATEMENT

The research design, including the search strategy and inclusion criteria, was developed as a combined effort of all authors. The search and primary analysis was performed by KD in close cooperation with HK and under the supervision of MIJ. The initial manuscript was drafted by KD and revised by HK and MIJ. The overall guarantor of this study is MIJ.

REFERENCES

1. Schleidgen S, et al. What is personalized medicine: sharpening a vague term based on a systematic literature review. BMC Med Ethics. 2013;14(1):1–12.

2. Van Til JA, IJzerman MJ. Why should regulators consider using patient preferences in benefit-risk assessment? Pharmacoeconomics. 2014;32(1):1–4.

3. Gheita TA, et al. The potential of genetically guided treatment in Behçet's disease. Pharmacogenomics. 2016;17(10):1165–1174.

4. Liu Y, et al. Prostate cancer – a biomarker perspective. Front Endocrinol (Lausanne). 2012;3:72.

5. Towse A, Garrison LP. Economic incentives for evidence generation: promoting an efficient path to personalized medicine. Value Health. 2013;16(S6):S39–43.

6. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. Br J Cancer. 2014;110(3):551–555.

7. Ferrante Di Ruffano L, et al. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. J Clin Epidemiol. 2012;65(3):282–287.

8. Horgan D, et al. Getting Personal: accelerating Personalised and Precision Medicine Integration into Clinical Cancer Research and Care in Clinical Trials. Public Health Genomics. 2015;18(6):325–328.

9. Lawler M, Sullivan R. Personalised and Precision Medicine in Cancer Clinical Trials: panacea for Progress or Pandora's Box? Public Health Genomics. 2015;18(6):329–337.

10. Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. Value Health. 2015;18(2):137–140.

11. IJzerman MJ, et al. Implementation of comparative effectiveness research in personalized medicine applications in oncology: current and future perspectives. Comp Effectiveness Res. 2015;5:65.

12. Marshall DA, et al. Transforming Healthcare Delivery: integrating Dynamic Simulation Modelling and Big Data in Health Economics and Outcomes Research. Pharmacoeconomics. 2015;34(2):115–126.

13. Annemans L, et al. Current methodological issues in the economic assessment of personalized medicine. Value Health. 2013;16(S6):S20–6.

14. Miller JD, et al. Current Challenges in Health Economic Modeling of Cancer Therapies: A Research Inquiry. Am Health Drug Benefits. 2014;7(3):153–162.

15. Phillips KA, et al. To the translation of genomic information into clinical practice and health policy: utilization, preferences, and economic value. Curr Opin Mol Ther. 2008;10(3):260–266.

16. Ginsburg GS, Kuderer NM. Comparative effectiveness research, genomics-enabled personalized medicine, and rapid learning health care: a common bond. J Clin Oncol. 2012;30(34):4233–4242.

17. Postma MJ, et al. Health technology assessments in personalized medicine: illustrations for cost–effectiveness analysis. Expert Rev Pharmacoecon Outcomes Res. 2011;11(4):367–369.

18. Sailer AM, et al. Cost-effectiveness modelling in diagnostic imaging: a stepwise approach. Eur Radiol. 2015;25(12):3629–3637.

19. Ferrusi IL, et al. Do economic evaluations of targeted therapy provide support for decision makers? . J Oncol Practice. 2011;7(S3):36s–45s.

20. Golubnitschaja O, et al. Risk assessment, disease prevention and personalised treatments in breast cancer: is clinically qualified integrative approach in the horizon?. Epma J. 2013;4(1):6.

21. Hulme C, et al. Determining cost-effectiveness of advanced cancer care: a systematic review of economic models. BMJ Support Palliat Care. 2011;1(S1):A21–A2.

22. Mullins CD, et al. Uncertainty in assessing value of oncology treatments. Oncologist. 2010;15(S1):58–64.

23. Lee KA, et al. A systems science approach to fatigue management in research and health care. Nurs Outlook. 2014;6(5)2:313–321.

24. Marshall DA, et al. Applying dynamic simulation modeling methods in health care delivery research— the SIMULATE checklist: report of the ISPOR simulation modeling emerging good practices task force. Value Health. 2015;18(1):5–16.

25. Marshall DA, et al. Selecting a dynamic simulation modeling method for health care delivery research-part 2: report of the ISPOR dynamic simulation modeling emerging good practices task force. Value Health. 2015;18(2):147–160.

26. Beaulieu M, et al. Systematic review of pharmacoeconomic studies of pharmacogenomic tests. Pharmacogenomics. 2010;11(11):1573–1590.

27. Phillips KA, Van Bebber SL. A systematic review of cost-effectiveness analyses of pharmacogenomic interventions. Pharmacogenomics. 2004;5(8):1139–1149.

28. Plumpton CO, et al. A systematic review of economic evaluations of pharmacogenetic testing for prevention of adverse drug reactions. Pharmacoeconomics. 2016;34(8):1–23.

29. Verhoef TI, et al. A systematic review of cost-effectiveness analyses of pharmacogenetic-guided dosing in treatment with coumarin derivatives. Pharmacogenomics. 2010;11(7):989– 1002.

30. Wolfswinkel JF, et al. Using grounded theory as a method for rigorously reviewing literature. Eur J Inf Syst. 2013;22(1):45–55.

31. Nalejska E, et al. Prognostic and predictive biomarkers: tools in personalized oncology. Mol Diagn Ther. 2014;18(3):273–284.

32. Peer X, An G. Agent-based model of fecal microbial transplant effect on bile acid metabolism on suppressing Clostridium difficile infection: an example of agent-based modeling of intestinal bacterial infection. J Pharmacokinet Pharmacodyn. 2014;41(5):493–507.

33. Schubert KO, et al. The use of clinical and biological characteristics to predict outcome following First Episode Psychosis. Aust N Z J Psychiatry. 2015;49(1):24–35.

34. Guerriero C, et al. The cost-effectiveness of smoking cessation support delivered by mobile phone text messaging: txt2stop. Eur J Health Econ. 2013;14(5):789–797.

35. Lee T, et al. Personalized prophylactic anticoagulation decision analysis in patients with membranous nephropathy. Kidney Int. 2014;85(6):1412–1420.

36. Soeteman DI, et al. Modeling the effectiveness of initial management strategies for ductal carcinoma in situ. J Natl Cancer Inst. 2013;105(11):774–781.

37. Al-Badriyeh D, et al. Pharmacoeconomic evaluation of voriconazole versus posaconazole for antifungal prophylaxis in acute myeloid leukaemia. J Antimicrob Chemother. 2010;65(5):1052–1061.

38. Cutler CS, et al. A decision analysis of allogeneic bone marrow transplantation for the myelodysplastic syndromes: delayed transplantation for low-risk myelodysplasia is associated with improved outcome. Blood. 2004;104(2):579–585.

39. Desai K, et al. Modeling the impact of HIV chemoprophylaxis strategies among men who have sex with men in the United States: HIV infections prevented and cost-effectiveness. Aids. 2008;22(14):1829–1839.

40. Herman WH, et al. The cost-effectiveness of lifestyle modification or metformin in preventing type 2 diabetes in adults with impaired glucose tolerance. Ann Intern Med. 2005;142(5):323–332.

41. Hertenstein B, et al. Allogenic bone marrow transplantation of chemotherapy for patients with acute myeloid leukemia in first complete remission: a decision analysis approach. Ann Hematol. 1996;72(4):223–230.

42. Kasuya M, Meguro K. Health economic effect of donepezil treatment for CDR 0.5 converters to Alzheimer's disease as shown by the Markov model. Arch Gerontol Geriatr. 2010;50(3):295–299.

43. Kurosawa S, et al. A Markov decision analysis of allogeneic hematopoietic cell transplantation versus chemotherapy in patients with acute myeloid leukemia in first remission. Blood. 2011;117(2):2113–2120.

44. Lee SJ, et al. Unrelated donor bone marrow transplantation for chronic myelogenous leukemia: a decision analysis. Ann Intern Med. 1997;127(12):1080–1088.

45. Paltiel AD, et al. HIV preexposure prophylaxis in the United States: impact on lifetime infection risk, clinical outcomes, and cost-effectiveness. Clin Infect Dis. 2009;48(6):806–815.

46. Rein DB, et al. The cost-effectiveness of birth-cohort screening for hepatitis C antibody in U.S. primary care settings. Ann Intern Med. 2012;156(4):263–270.

47. Ayer T, et al. OR Forum—A POMDP approach to personalize mammography screening decisions. Oper Res. 2012;60(5):1019–1034.

48. Carles M, et al. Cost-effectiveness of early detection of breast cancer in Catalonia (Spain). BMC Cancer. 2011;11:192.

49. Chen A, et al. Clinical effectiveness and cost-effectiveness of HIV pre-exposure prophylaxis in men who have sex with men: risk calculators for real-world decision-making. Plos One. 2014;9(10):e108742.

50. Djalalov S, et al. Genetic testing in combination with preventive donepezil treatment for patients with amnestic mild cognitive impairment: an exploratory economic evaluation of personalized medicine. Mol Diagn Ther. 2012;16:389–399.

51. Eckman MH, et al. Cost-effectiveness of using pharmacogenetic information in warfarin dosing for patients with nonvalvular atrial fibrillation. Ann Intern Med. 2009;150(2):73–83.

52. Ferket BS, et al. Personalized prediction of lifetime benefits with statin therapy for asymptomatic individuals: a modeling study. Plos Med. 2012;9(12):e1001361.

53. Greeley SAW, et al. The cost-effectiveness of personalized genetic medicine: the case of genetic testing in neonatal diabetes. Diabetes Care. 2011;34(3):622–627.

54. Guzauskas GF, et al. A risk-benefit assessment of prasugrel, clopidogrel, and genotype-guided therapy in patients undergoing percutaneous coronary intervention. Clin Pharmacol Ther. 2012;91(5):829–837.

55. Handorf EA, et al. Cost effectiveness of personalized therapy for first-line treatment of stage IV and recurrent incurable adenocarcinoma of the lung. J Oncol Pract. 2012;8(5):267–274.

56. Hochheiser LI, et al. Economic utility of a blood-based genomic test for the assessment of patients with symptoms suggestive of obstructive coronary artery disease. Popul Health Manag. 2014;17(5):287–296.

57. Juusola JL, et al. The cost-effectiveness of preexposure prophylaxis for HIV prevention in the United States in men who have sex with men. Ann Intern Med. 2012;156(8):541–550.

58. Kievit W, et al. Cost effectiveness of a new strategy to identify HNPCC patients. Gut. 2005;54(1):97–102.

59. Kobayashi T, et al. Prostate cancer screening strategies with re-screening interval determined by individual baseline prostate-specific antigen values are cost-effective. Eur J Surg Oncol. 2007;33(6):783–789.

60. Krieckaert CL, et al. Personalised treatment using serum drug levels of adalimumab in patients with rheumatoid arthritis: an evaluation of costs and effects. Ann Rheum Dis. 2015;74(2):361–368.

61. Ladabaum U, et al. Strategies to identify the Lynch syndrome among patients with colorectal cancer: a cost-effectiveness analysis. Ann Intern Med. 2011;155(2):69–79.

62. Ladapo JA, et al. Clinical outcomes and cost-effectiveness of coronary computed tomography angiography in the evaluation of patients with chest pain. J Am Coll Cardiol. 2009;54(25):2409–2422.

63. Leunis A, et al. The development and validation of a decision-analytic model representing the full disease course of acute myeloid leukemia. Pharmacoeconomics. 2013;31(7):605–621.

64. Lieberthal RD, et al. An economic model to value companion diagnostics in non-small-cell lung cancer. Per Med. 2013;10(4):139–147.

65. De Lima Lopes G, et al. Cost-effectiveness of epidermal growth factor receptor mutation testing and first-line treatment with gefitinib for patients with advanced adenocarcinoma of the lung. Cancer. 2012;118:1032–1039.

66. Liu S, et al. Cost effectiveness of fibrosis assessment prior to treatment for chronic hepatitis C patients. Plos One. 2011;6(12):e26783.

67. Mvundura M, et al. The cost-effectiveness of genetic testing strategies for Lynch syndrome among newly diagnosed patients with colorectal cancer. Genet Med. 2010;12(2):93–104.

68. O'Donoghue C, et al. Aggregate cost of mammography screening in the United States: comparison of current practice and advocated guidelines. Ann Intern Med. 2014;160(3):145.

69. Perlis RH, et al. When is pharmacogenetic testing for antidepressant response ready for the clinic? A cost-effectiveness analysis based on data from the STAR*D study. Neuropsychopharmacology. 2009;34(10):2227–2236.

70. Petta S, et al. Group WEFs. Personalized cost-effectiveness of boceprevir-based triple therapy for untreated patients with genotype 1 chronic hepatitis C. Dig Liver Dis. 2014;46(10):936–942.

71. Ramsey SD, et al. An economic viewpoint on alternative strategies for identifying persons with hereditary nonpolyposis colorectal cancer. Genet Med. 2003;5(5):353–363.

72. Van Ravesteyn NT, et al. What level of risk tips the balance of benefits and harms to favor screening mammography starting at age 40?. Ann Intern Med. 2012;156(9):609–617.

73. Reyes CM, et al. Comparison of selection strategies for genetic testing of patients with hereditary nonpolyposis colorectal carcinoma: effectiveness and cost-effectiveness. Cancer. 2002;95(9):1848–1856.

74. Schousboe JT, et al. Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness. Ann Intern Med. 2011;155(1):10–20.

75. Sullivan SD, et al. Cost-effectiveness of risk stratification for preventing type 2 diabetes using a multi-marker diabetes risk score. J Med Econ. 2011;14(5):609–616.

76. Sung L, et al. Treatment options for patients with acute myeloid leukemia with a matched sibling donor: a decision analysis. Cancer. 2003;97(3):592–600.

77. Vilaprinyo E, et al. Cost-effectiveness and harm-benefit analyses of risk-based screening strategies for breast cancer. Plos One. 2014;9(2):e86858.

78. Kermack WO, McKendrick AGA. Contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences; 1927;115(772):700–721.

79. Kaelbling LP, et al. Planning and acting in partially observable stochastic domains. Artif Intell. 1998;101(1-2):99–134.

80. Zucchelli E, Jones AM, The evaluation of health policies through microsimulation methods. Health, Econometrics and Data Group (HEDG) Working Papers. 2010.

81. Karnon J, et al. Modeling using discrete event simulation: a report of the ISPOR-SMDM modeling good research practices task force–4. Value Health. 2012;15(6):821–827.

82. Gail M, Rimer B. Risk-based recommendations for mammographic screening for women in their forties. J Clin Oncol. 1998;16(9):3105–3114.

APPENDIX 1: search strategy

We performed a broad search on PubMed using the search strategy specified below.

Primary search terms in the title or abstract:

- "Personalized" referring to Personalized Medicine
- "Personalised" referring to Personalised Medicine
- "Precision Medic*" referring to Precision Medicine, Precision Medication, etc.

Secondary search terms in the title or abstract:

- Model* referring to Model, Models, Modeling, etc.
- Simula* referring to Simulation, Simulations, Simulating, etc.

Additional health economic search terms:

- "*effect*" referring to effect, effects, effectiveness, cost-effectiveness, etc.
- "*util*" referring to utility, cost-utility, risk-utility, etc.
- "*benefit*" referring to benefit, benefits, risk-benefit, etc.
- "*risk*" referring to risk, risks, benefit-risk, etc.
- "*conseq*" referring to consequence, consequences, etc.
- "cost"
- "costs"
- "decision-anal*" referring to decision-analytic, decision-analytical, etc.

Combined the exact used search algorithm was: *(personalised[tiab] OR personalized[tiab] OR precision medic*[tiab]) AND (model*[tiab] OR simula*[tiab]) AND (*effect*[tiab] OR *util*[-tiab] OR *benefit*[tiab] OR *risk*[tiab] OR *conseq*[tiab] OR cost[tiab] OR costs[tiab] OR decision-anal*[tiab]).*

APPENDIX 2: exclusion categories

Categories	Description
Qualitative Research	Reviews, Summaries, Discussions and Proposals of Qualitative Frameworks.
Experimental Research	Clinical Trials, Prospective Studies, Retrospective Studies, and Proof of Concept Studies.
Prediction Models	Decision Tools, Mathematical Models, and Statistical Models.
Biomedical Models	Biomechanical and Biological Models: Models or simulations of the human body or (combination of) organs.
Animal Studies	Studies on animals and human derived tissue transferred to animals.
Decision Tool Validation	Validation of Decision Tools or Proposed Frameworks.
User Opinion and Preferences	Surveys, Patient Preference Studies, and Focus Groups.
No Abstract and Text	Publications for which no abstract and no full text was available.
Not Personalized Medicine	ICT, personal learning, personalized alert notifications, personalized evacuation routes, personal surveillance, labor market predictions, etc.

CHAPTER 3

Matching the model with the evidence: comparing discrete event simulation and state-transition modeling for time-to-event predictions in a cost-effectiveness analysis of treatment in metastatic colorectal cancer patients

Koen Degeling
Mira D. Franken
Anne M. May
Martijn G.H. van Oijen
Miriam Koopman
Cornelis J.A. Punt
Maarten J. IJzerman
Hendrik Koffijberg

This chapter has been published as:
Degeling K, Franken MD, May AM, van Oijen MGH, Koopman M, Punt CJA, et al. Matching the model with the evidence: comparing discrete event simulation and state-transition modeling for time-to-event predictions in a cost-effectiveness analysis of treatment in metastatic colorectal cancer patients. Cancer Epidemiology. 2018;57:60-7.

ABSTRACT

Background: Individual patient data, e.g. from clinical trials, often need to be extrapolated or combined with additional evidence when assessing long-term impact in cost-effectiveness modeling studies. Different modeling methods can be used to represent the complex dynamics of clinical practice; the choice of which may impact cost-effectiveness outcomes. We compare the use of a previously designed cohort discrete-time state-transition model (DT-STM) with a discrete event simulation (DES) model.

Methods: The original DT-STM was replicated and a DES model developed using AnyLogic software. Models were populated using individual patient data of a phase III study in metastatic colorectal cancer patients, and compared based on their evidence structure, internal validity, and cost-effectiveness outcomes. The DT-STM used time-dependent transition probabilities, whereas the DES model was populated using parametric distributions.

Results: The estimated time-dependent transition probabilities for the DT-STM were irregular and more sensitive to single events due to the required small cycle length and limited number of event observations, whereas parametric distributions resulted in smooth time-to-event curves for the DES model. Although the DT-STM and DES model both yielded similar time-to-event curves, the DES model represented the trial data more accurately in terms of mean health-state durations. The incremental cost-effectiveness ratio (ICER) was €172,443 and €168,383 per Quality Adjusted Life Year gained for the DT-STM and DES model, respectively.

Conclusion: DES represents time-to-event data from clinical trials more naturally and accurately than DT-STM when few events are observed per time cycle. As a consequence, DES is expected to yield a more accurate ICER.

INTRODUCTION

Healthcare expenditures have increased importantly over the last decades, especially in oncology due to expensive novel targeted agents and personalized treatments based on molecular markers in order to provide patients with the best possible care [1,2]. Cost-effectiveness analysis of such novel medical technologies is becoming increasingly relevant, as it may inform treatment, resource allocation, and research prioritization decisions. This is illustrated by the standardized approaches to value cancer treatment options in terms of efficacy and costs for clinicians [3,4] and guidance for performing cost-effectiveness analysis alongside clinical trials [5].

High quality individual patient data (IPD) on health outcomes, resource use, and care procedures, e.g. obtained from randomized controlled trials (RCTs), are the preferred source of evidence for cost-effectiveness analysis. However, single individual patient datasets do not always provide all (or the only) evidence required for estimating the (long-term) cost-effectiveness of medical technologies [6,7], indicating the need for cost-effectiveness models to synthesize evidence from additional sources or to extrapolate beyond the time horizon of e.g. RCTs [5,8]. Such cost-effectiveness models should adequately represent clinical practice and, therefore, reflect the true nature of the evidence used to define them, including evidence obtained from RCTs and other sources of IPD. In other words, the model should match the evidence.

The primary outcome of many clinical oncology studies is the time until an event of interest occurs, e.g. the patients' overall survival or progression-free survival from the moment of randomization, which are typically recorded continuously over time. However, the most frequently applied cost-effectiveness modeling method, i.e. discrete-time state-transition modeling (DT-STM) [9], uses transition probabilities over discrete time cycles with a fixed length to represent the progression of time. For example, in an DT-STM with time cycles of three weeks patients can only progress to another health state after this predefined and rigid time length, even though in daily practice patients may progress at any time instead of only at a multiple of three weeks. The length of these time cycles needs to be chosen so that the complex dynamics of clinical practice are appropriately represented [9]. For DTSTM to represent clinical practice better, shorter cycle lengths would be preferable [10]. Although half-cycle corrections may be applied to avoid bias and to better approximate clinical practice [11], this still insufficiently allows complex clinical dynamics if the cycle length is too long [12].

Using shorter cycles lengths can be disadvantageous, mainly because of increase in number of cycles that needs to be simulated. Besides increasing the computational burden of the simulation [9,12], the larger number of cycles makes it more challenging to represent the uncertainty in the transition probabilities, as the uncertainty in the numerous cycle-specific probabilities needs to be reflected while also maintaining the correlation between them. Furthermore, because the expected number of observations within a cycle decreases with decreasing cycle length, the likelihood of substantial irregularities in transition probabilities between successive cycles is expected to increase. These irregularities are likely to impact the simulation outcomes and do not correspond to clinical practice, as the probability of an event

is commonly expected to be similar between successive moments, i.e. the transition-curves follow a smooth pattern over time.

Discrete event simulation (DES) is an alternative modeling technique to which the challenges associated with discrete time cycles do not apply. Events can occur at any time in a DES model, because the time to these events are typically modeled using smooth time-to-event distributions, e.g. Gamma or Weibull distributions [13]. In DES, the behavior of a system is translated into an ordered sequence of well-defined events, which comprise specific changes in the system's state at a specific point in time [13]. DES is well suitable for modeling clinical processes, as it is able to incorporate patient-level characteristics and clinical histories, competing resources, and interactions between different actors, e.g. physicians and patients [14]. Although originating from the operations research field, DES is increasingly being used for cost-effectiveness modeling [15].

Several studies have compared the use of DT-STM and DES for cost-effectiveness analyses of medical technologies. Using the same model structure and evidence, quantitative outcomes such as the incremental cost-effectiveness ratio (ICER), are unlikely to be substantially different between these modeling methods [16,17]. However, substantial differences in outcomes may occur, if the use of DES results in a more appropriate representation of clinical practice compared to DT- STM, for example by including patient characteristics or considering resource constraints [18]. Especially in the scenario in which insufficient observations are available for the chosen cycle length, and irregularities in the cycle-specific transition probabilities are substantial when using DT-STM, the use of DES might be preferable.

The objective of this study is to compare the evidence structure and outcomes of a recently published cost-effectiveness DT-STM [19] with those of a newly developed DES model. The comparison will be performed based on the dataset of the randomized clinical phase III CAIRO3 study, in which maintenance treatment with capecitabine and bevacizumab (CAP-B) or observation in metastatic colorectal cancer patients after six induction cycles of capecitabine, oxaliplatin, and bevacizumab (CAPOX-B) was evaluated [20]. The results of this study should facilitate a better understanding of the potential impact of selecting a modeling method for cost-effectiveness modeling studies informed by IPD.

METHODS

Maintenance treatment in metastatic colorectal cancer

The CAIRO3 study (NCT00442637) is a randomized clinical phase III study, which was carried out by the Dutch Colorectal Cancer Group (DCCG) in 64 hospitals in the Netherlands. A total of 558 metastatic colorectal patients with stable disease or better after six cycles of CAPOX-B induction therapy were randomized to either receive CAP-B maintenance treatment or observation until progression, which is referred to as the post-induction stage. CAPOX-B treatment was to be reintroduced upon progression on either maintenance or observation, and continued until second progression (primary end-point), which is referred to as the re-introduction stage. Although second progression was the primary end-point of the CAIRO3

study, the cost-effectiveness analysis of the CAIRO3 study also considered additional lines of treatment after second progression [19], which is referred to as the salvage therapy stage. Study results have been previously published [20].

State-transition model

A cohort DT-STM, i.e. Markov model, was originally developed for the cost-effectiveness analysis of the CAIRO3 study and included four health states: post-induction, reintroduction, salvage therapy, and death (Figure 1a). The model was defined using cohort level cycle-specific transition probabilities, which were estimated from the CAIRO3 trial using Life Tables in IBM SPSS Statistics software, version 23, IBM Corp. (Armonk, NY, USA). This indicates that the probability of moving from one state to another depended only on the time passed since the start of the simulation, e.g. time from randomization until first progression. Half-cycle correction was applied and 100 cycles of three weeks were simulated in total. The DT-STM was built using TreeAge Pro Healthcare v.2014, TreeAge Software (Williamstown, MA, USA), and is described in detail elsewhere [19].

To facilitate an adequate comparison between the two modeling methods, the DT-STM was first replicated in AnyLogic multi-method simulation software, v.7.3, The AnyLogic Company (Chicago, IL, USA), the environment also used for developing the DES model. This replicated DT-STM was then compared to the original DT-STM to assess potential variation in outcomes due to the use of different software environments. In total, 100 events were generated at intervals of three weeks, corresponding to the setup in the original DT-STM. Following each event, the occupation of the health states was recorded and used to calculate health and economic outcomes at the corresponding point in time. The model was validated by structured "walk-throughs", comparing (intermediate) results with calculations by hand, extreme value analysis, trace analysis, and cross validation with the original DT-STM during model development, and sensitivity analysis using the final model [21,22].

Discrete event simulation model

The DES model was defined on patient-level using AnyLogic software and according to the ISPOR-SMDM Modeling Good Research Practice Task Force guidelines [14]. The model was defined to have the same health states as the DT-STM (Figure 1b). Although DES allows for constrained resources to be accounted for, resource use was not considered in the DT-STM and, consequently, also not in the DES model.

Both Weibull and Gamma parametric distributions [13] were estimated from the CAIRO3 trial data with maximum likelihood estimation (MLE), or methods of moments estimation (MME) where MLE was not successful, using the *fitdist* function of the *fitdistrplus* [23] package in R Statistical Software [24]. Estimated parametric distributions were compared graphically based on density plots, Q-Q plots, and P-P plots, and numerically based on the Akaike information criterion and Bayesian information criterion. Since performance was similar without meaningful differences, and all Weibull distributions could be estimated with MLE, whereas MME was required for some Gamma distributions, Weibull distributions were assumed for all health state-specific time-to-event parameters in the DES model. Transitions between

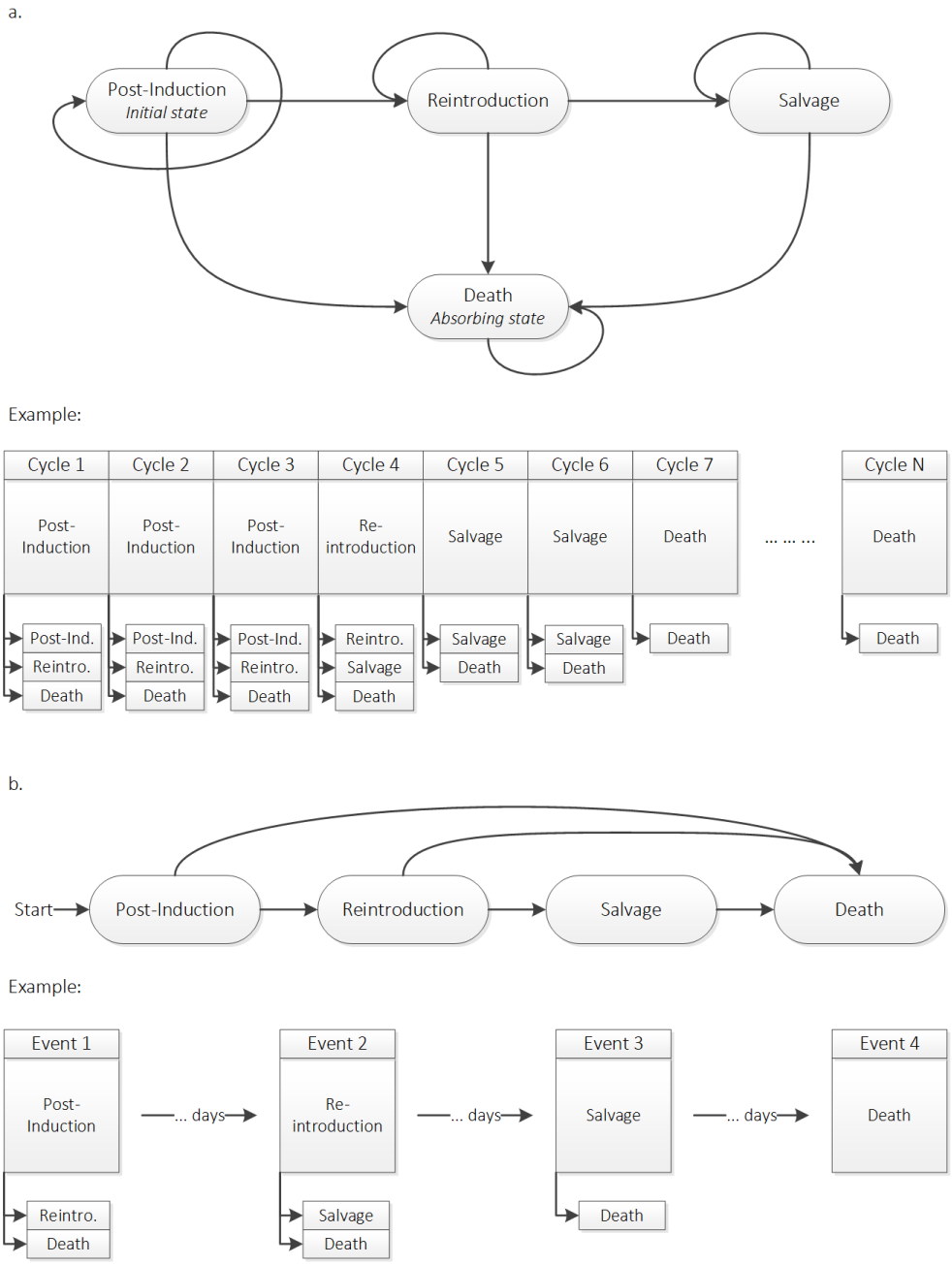


Figure 1. *Graphical representation of the health states defined in the Discrete-Time State-Transition (a) and the Discrete Event Simulation (b) model.*

health states, i.e. events, were based on patient-level processing times, which were randomly drawn from Weibull distributions. Competing risks were handled by stratifying state-specific time-to-event distributions according to the two competing events that were considered, i.e. progression and death, and selecting the event to occur based on the respective observed probabilities of progression and survival [25]. To illustrate this, for a patient entering the reintroduction state a randomly drawn value compared to the chance of progression determined whether the patient would survive and progress to the salvage therapy state. Next, the time to the selected event, i.e. progression or death, was randomly drawn from the corresponding Weibull distribution.

A total of 10,000 patients were simulated per treatment strategy in the DES model, resulting in relative standard errors for the mean costs and effects of approximately 0.5%. No fixed runtime was assumed, so the simulation terminated when all patients had left the model, i.e. reached the death state. Patient-level outcomes were calculated using the time spent in each health state and summarized to enable comparison of the two treatment strategies. The DES model was validated by structured "walk-throughs", comparing (intermediate) results with calculations by hand, extreme value analysis, trace analysis, and cross validation with both DT-STMs during model development, and by sensitivity analysis [21,22].

Model comparison

First, the original DT-STM and the replicated DT-STM were compared based on the cost-effectiveness outcomes of the CAIRO3 case study, to assess potential variation in outcomes due to differences in software environments. For this analysis, the incremental cost-effectiveness ratio (ICER) expressed in incremental costs per Quality Adjusted Life Year (QALY) gained served as the primary outcome. Costs and effects were discounted at discount rates of 4% and 1.5% per year, respectively, according to Dutch pharmacoeconomic guidelines [26]. Probabilistic sensitivity analysis (PSA) was performed using Monte Carlo simulation with 10,000 samples to assess the effect of the uncertainty surrounding the input parameters on the primary outcome measure [27]. Since the original cost-effectiveness analysis did not account for uncertainty in the correlated cycle-specific transition probabilities [19], uncertainty in the correlated distribution parameters used to represented the time-to-event evidence in the DES model, was also not considered to maintain comparability between both models. Parameter values used to populate both models, including their distributions, are listed in the publication of the original CAIRO3 cost-effectiveness analysis [19], as well as in Supplementary Materials 1.

Subsequently, the replicated DT-STM and the DES model were qualitatively and quantitatively compared based on the case study, to assess potential differences between the two modeling methods. The models were qualitatively compared based on the evidence structure. Thereafter, modeling methods were quantitatively compared based on cost-effectiveness outcomes and simulation outcomes, i.e. the simulated health-state durations. All results were graphically represented, using Kaplan-Meier curves for the simulation outcomes and incremental cost-effectiveness planes and cost-effectiveness acceptability curves (CEAC) for the cost-effectiveness outcomes.

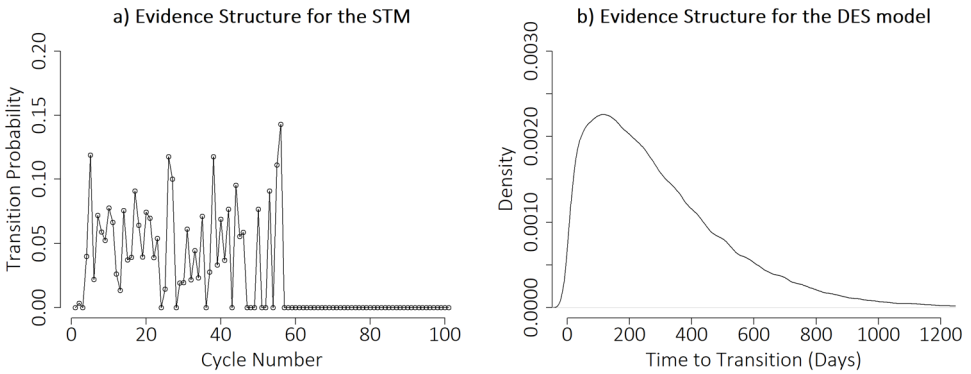


Figure 2. *a) probability curve for the time to transition of the post-introduction to the reintroduction health state for the maintenance strategy per cycle (with a 3-week duration) for the DT-STM, and b) probability density curve for the time to transition of the post-introduction to the reintroduction health state for the maintenance strategy in the DES model.*

RESULTS

The replicated DT-STM developed for this study yielded comparable cost-effectiveness outcomes as the original DT-STM developed in a different software environment. The results for the original DT-STM have been previously published elsewhere and are not presented here for the sake of readability [19]. The replicated DT-STM will be referred to just as "DT-STM" in the subsequent part of this manuscript.

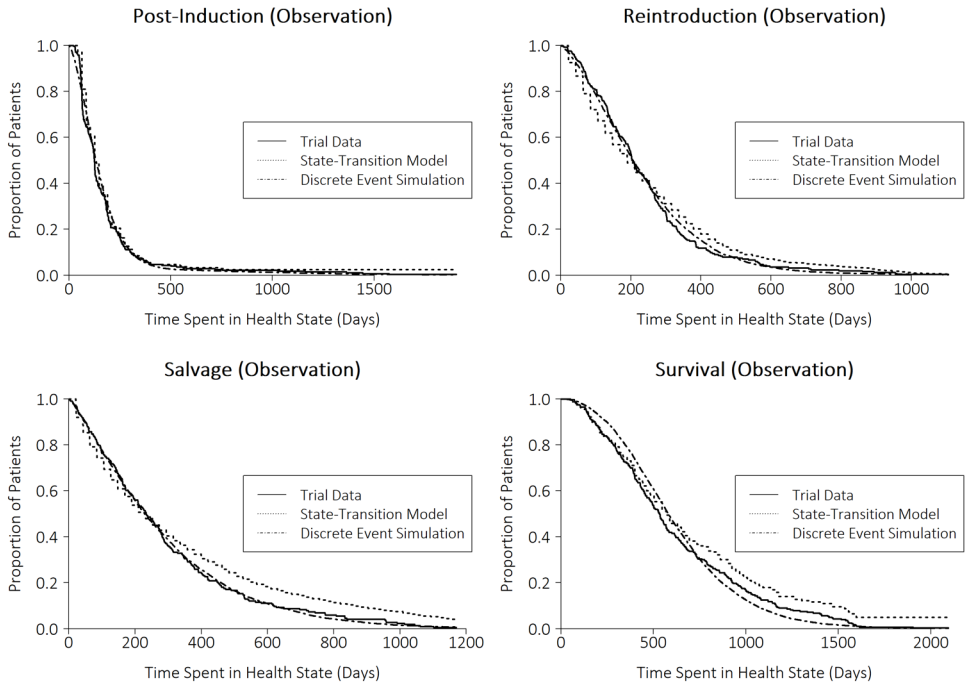
Simulation of health state-transitions

Health state-transitions in the DES model yield smooth time-to-event curves defined using Weibull distributions estimated based on the CAIRO3 data. In contrast, the time-dependent probabilities used for health state-transitions in the DT-STM become irregular (non-smooth) when only few events are observed for some transitions. The irregularities in these transition probabilities are caused by a decreasing number of patients retained in a health state over time, causing large variations in the observed subsequent probability of a health state-transition. An example of this is presented in Figure 2, which depicts the difference between the DT-STM and DES model in health state-transitions from the post-induction state to the reintroduction state for the maintenance treatment strategy.

The Kaplan-Meier curves for the health state-durations simulated in the DT-STM and DES model, compared to the CAIRO3 data, demonstrate that both modeling methods represent the clinical data well (Figure 3). However, when the mean time-to-transition presented in the descriptive statistics below the figure are considered, the DES model seems to represent the trial data more accurately. In example, the mean health-state duration of the post-induction state for the observation strategy was 175.7 days, 207.5 days and 173.4 days for the trial data, DT-STM and the DES model, respectively.

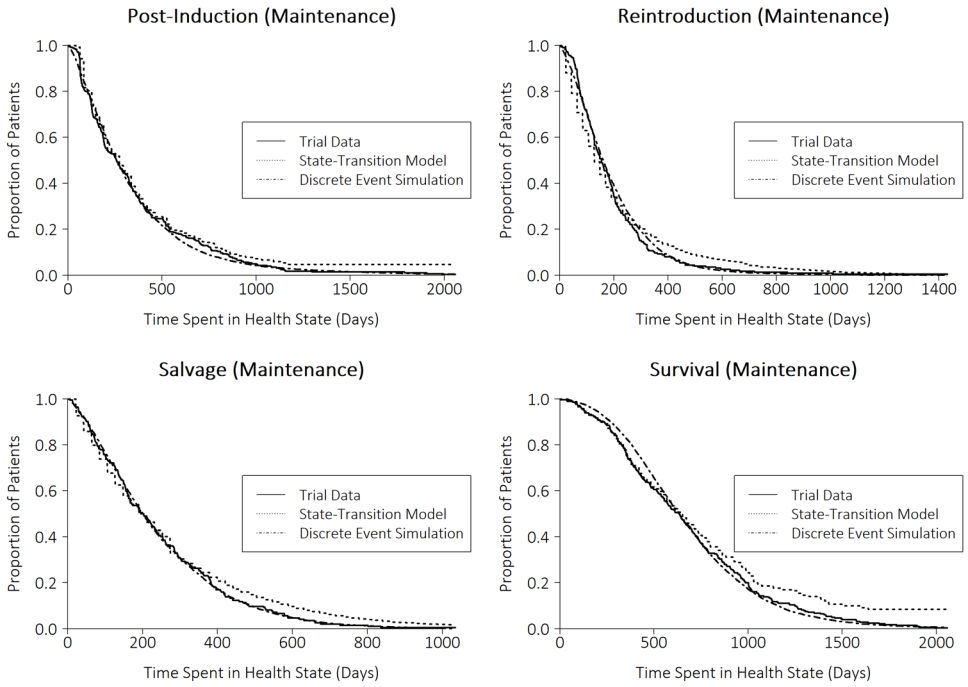
Cost-effectiveness analysis

The cost-effectiveness outcomes obtained from the DT-STM and the DES model are presented in the incremental cost-effectiveness planes of Figure 4. The incremental effectiveness estimates, including their 95% confidence intervals (CI), for CAP-B maintenance therapy compared to the observation strategy are 0.21 (CI: 0.015; 0.430) and 0.18 (CI: 0.006; 0.374) QALYs, and the incremental costs are €35,536 (CI: 19,945; 54,629) and €30,053, (CI: 17,047; 46,132) for the DT-STM and DES model, respectively. The mean ICERs are €172,443 and €168,383 per QALY gained for the DT-STM and DES model, respectively.



	Post-Induction			Reintroduction			Salvage Therapy			Overall Survival		
	Trial	STM	DES	Trial	STM	DES	Trial	STM	DES	Trial	STM	DES
2.5%	31.9	42.0	14.5	29.7	21.0	18.6	17.1	21.0	15.4	101.4	84.0	124.2
Median	124.0	126.0	132.3	203.0	189.0	205.8	233.0	231.0	228.5	532.0	567.0	580.0
Mean	175.7	207.5	173.4	238.8	253.4	238.8	292.3	349.9	291.8	619.4	707.0	626.6
97.5%	717.9	756.0	529.8	711.0	903.0	643.2	978.6	1323.0	913.5	1556.9	2100.0	1384.2
n	279	10000	10000	254	9196	9698	216	7616	7599	279	10000	10000

Figure 3a. *Kaplan-Meier curves for the time-spent in the health states for the observation strategy.*



	Post-Induction			Reintroduction			Salvage Therapy			Overall Survival		
	Trial	STM	DES	Trial	STM	DES	Trial	STM	DES	Trial	STM	DES
2.5%	48.9	63.0	20.5	26	21.0	13.0	17.5	21.0	16.2	92.9	84.0	133.8
Median	257.0	273.0	259.1	153.0	126.0	159.7	198.0	210.0	200.2	635.0	651.0	633.1
Mean	354.4	410.9	349.4	192.0	207.2	194.2	242.6	266.6	242.5	680.2	766.2	688.0
97.5%	1131.9	2100.0	1246.0	608	890.4	569.5	662.4	924.0	692.6	1639.9	2100.0	1568.3
n	279	10000	10000	225	8385	8971	194	6810	6776	279	10000	10000

Figure 3b. *Kaplan-Meier curves for the time-spent in the health states for the maintenance strategy.*

The PSA for both models only demonstrated a small difference in the amount of uncertainty surrounding the mean ICER point-estimates (Figure 4). This is illustrated by the magnitude of the 95%-confidence ellipses surrounding these estimates, being slightly smaller for the DES model. However, as both mean ICER point-estimates and corresponding confidence ellipses are located rather similarly compared to the willingness-to-pay (WTP) threshold, the CEACs for both models are similar (Figure 5).

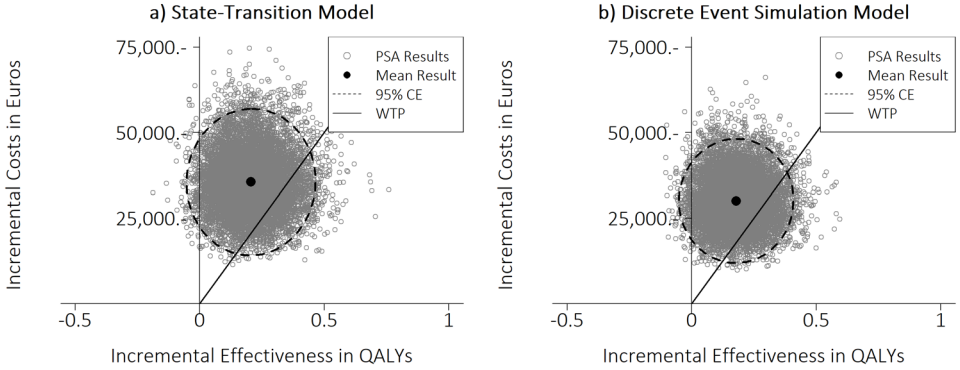


Figure 4. *Incremental Cost-Effectiveness Planes comparing the maintenance treatment strategy with the observation strategy at a Willingness to Pay (WTP) of €100,000.- per Quality Adjusted Life-Year (QALY) gained, for a) the discrete-time state-transition model, and b) the discrete event simulation model. CE, Confidence Ellipse.*

DISCUSSION

Smooth health state-transition curves served as input for the DES model, presenting the data in an informative manner. Conversely, it is more complicated to interpret the time-dependent health state-transitions probabilities used as input for the DT-STM. We have shown that these probabilities are irregular over time, due to scarce observations in many of the time cycles. Therefore, the health state-transition curves used in the DES model were much more representative of a "natural" patient flow through health states over time. Additionally, the Kaplan-Meier curves per health state simulated from the DES model matched the original study Kaplan-Meier curves slightly better, especially with regard to the mean time to transition from one health state to another, e.g. from randomization to the start of therapy reintroduction. The increasing difference between the trial data and STM over time, suggests a cumulative effect over successive health states, which may be amplified by a combination of irregularities in transition probabilities and their time-dependency.

Cost-effectiveness outcomes were comparable for the DT-STM and the DES model (ICER €172,443 and €168,383, respectively). The rather small differences observed, can be explained by the disparities in simulated mean time to transitions between both models. Furthermore, the magnitude of uncertainty surrounding the mean ICER point-estimate was smaller for the DES model. The observed difference in the uncertainty might be caused by the irregularities in the health state-transition probabilities in the DT-STM, consequently causing more extreme effects compared to the smooth health-state transition curves of the DES model. Results of this study did not alter the previously published conclusion that CAP-B maintenance may not be regarded as cost-effective [19].

These results confirm that cost-effectiveness outcomes are not expected to be substantially different between DT-STM and DES models, if both models are based on the same evidence [16,17]. It is, however, imaginable that ICER outcomes closer to a country's willingness to pay

threshold might incur different conclusions on cost-effectiveness depending on the choice of modeling method. This was previously demonstrated by Jahn et al comparing a DES model and a DT-STM evaluating decision tools for adjuvant chemotherapy treatment in breast cancer [28].

Even though the DES methodology may initially seem more complex for novices, its model structure and evidence structure more closely match transitions and events as observed in clinical trials, compared to that of DT-STM. Once familiar with the DES methodology, the parametric distributions used to describe time-to-event data are straightforward to estimate and interpret. Furthermore, these parametric distributions enable uncertainty in their parameter estimates to be included in the PSA more easily than the (correlated) uncertainty that is present in every individual time-dependent transition probability [29]. However, by discretizing parametric time-to-event distributions into transition probabilities that can be used to populate a DT-STM, uncertainty in these transition probabilities can be represented. Additionally, by discretizing a parametric distribution rather than directly estimating transition probabilities from individual patient data, issues with regard to irregularities in these time-dependent transition probabilities may also be addressed. Furthermore, extrapolation beyond the time horizon of RCTs, although challenging, can be performed by fitting these parametric distributions [30]. Although parametric distributions can be used to address these general and DT-STM related challenges, doing so can be considered suboptimal due to the required discretization, whereas these parametric distributions can be incorporated directly in DES. In this respect, issues regarding appropriately reflecting uncertainty surrounding health economic outcomes, scarce events, and extrapolation may more easily be addressed using DES methodology. Regardless of these advantages to DES, DT-STM typically is computation-

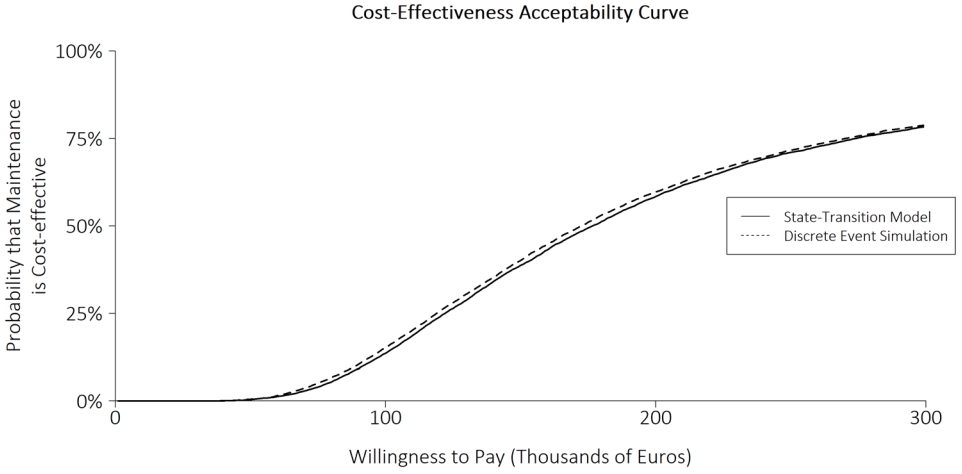


Figure 5. *Cost-Effectiveness Acceptability Curves representing the probability that the maintenance treatment strategy is cost-effective compared to the observation strategy for a range of Willingness to Pay threshold values.*

ally simpler, can be implemented using spreadsheets, and requires limited (programming) skills to do so, whereas implementation of DES is mainly limited to specialized simulation and statistical software [15,17,31]. Hence, regarding external review of models, DT-STM currently has an advantage, while experience with DES in health economics is developing [15,31].

DES seems the preferable modeling method compared to DT-STM for the evaluation of individual patient time-to-event data, which is also supported by the health economic modeling literature [15,31]. In particular when time cycle size needs to be very small to adequately reflect dynamic treatment and monitoring processes, leading to irregularities in the estimated time-dependent transition probabilities due to a lack of observed events. However, DT-STM is still the most commonly used modeling method in cost-effectiveness modeling, for which different reasons can be identified. Firstly, as mentioned before, DES might initially be thought of as a more complex methodology requiring more evidence. This study demonstrated that DES models do not necessarily require more evidence or are more complex. Secondly, comprehensive guidance is available on how to use a (cohort) DT-STM for the evaluation of healthcare interventions [9], whereas the available guidance on the use of DES is less specific [14]. Researchers and clinicians with interest in health economics alike, however, should be aware of the potential advantages of DES compared to DT-STM, especially with regard to cost-effectiveness analyses informed by patient-level time-to-event data obtained from e.g. obtained from clinical trials.

This study compared a cohort state-transition model qualitatively and quantitatively based on an extensive health economic evaluation informed by patient-level time-to-event data obtained from the CAIRO3 study. Both models were developed in the same software environment and analyzed according to health economic good practices guidelines, optimizing the validity of our results. However, the generalizability of these single case study results is limited, though the results found are in line with literature [16,17,28]. Furthermore, the full potential of DES was not utilized, since no patient-level characteristics were incorporated and, deliberately, parameter uncertainty in the time-to-event distributions parameter estimates themselves were not considered. The inclusion of patient-level characteristics in DES models undoubtedly allows for even better representation of clinical practice. Finally, Weibull distributions were assumed for representing health state durations in the DES model, which may potentially influence health economic outcomes. To assess the impact of this design choice, simulations were additionally performed with Gamma distributions instead of Weibull distributions, which did not result in meaningfully different results.

In conclusion, the results show that the DT-STM and DES model did not yield substantially different outcomes if they are developed based on the same health states and evidence. Which modeling method should be applied, depends on the complexity of the clinical process to be modeled, the available evidence, and the modelers' experience. In our opinion, DES is the preferable modeling method in the scenario that patient-level time-to-event data is available, e.g. from clinical studies, as its model structure and evidence structure represent the dynamics of daily clinical practice more naturally.

AUTHOR CONTRIBUTIONS

The research design was developed as combined effort of KD, MF, AM, MK, and HK, and was revised by MO, CP, and MIJ. Models were developed, and analyses were performed by KD and MF, under the supervision of MK and HK. All authors contributed to the interpretation and discussion of the results. The initial manuscript was drafted by KD, MF, and HK, and critically revised by AM, MO, MK, CP, and MIJ. The overall guarantor of this study is HK.

FUNDING SOURCES AND CONFLICTS OF INTEREST

No funding was received for performing this study. The authors declare no conflict of interest.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The CAIRO3 study protocol, which includes secondary data analyses such as performed in this study, was approved by the Committee on Human-Related Research Arnhem-Nijmegen in the Netherlands. Written informed consent was not required for this study separately, as written informed consent was obtained from all participants in the CAIRO3 study, which includes secondary data analyses such as performed in this study.

SUPPLEMENTARY MATERIALS

Supplementary material related to this article can be found in the online version at: https://doi.org/10.1016/j.canep.2018.09.008

REFERENCES

1. Meropol NJ, et al. American Society of Clinical Oncology Guidance Statement: The Cost of Cancer Care. J Clin Oncol. 2009;27(23):3868-74.

2. Luengo-Fernandez R, et al. Economic burden of cancer across the European Union: a population-based cost analysis. Lancet Oncol. 2013;14(12):1165-74.

3. Cherny NI, et al. A standardised, generic, validated approach to stratify the magnitude of clinical benefit that can be anticipated from anti-cancer therapies: the European Society for Medical Oncology Magnitude of Clinical Benefit Scale (ESMO-MCBS). Ann Oncol. 2015;26(8):1547-73.

4. Schnipper LE, et al. American Society of Clinical Oncology Statement: A Conceptual Framework to Assess the Value of Cancer Treatment Options. J Clin Oncol. 2015;33(23):2563-77.

5. Ramsey S, et al. Good Research Practices for Cost-Effectiveness Analysis Alongside Clinical Trials: The ISPOR RCT-CEA Task Force Report. Value Health. 2005;8(5):521-33.

6. Sculpher MJ, et al. Whither trial-based economic evaluation for health care decision making? Health Econ. 2006;15(7):677-87.

7. Saramago P, et al. Deriving Input Parameters for Cost-Effectiveness Modeling: Taxonomy of Data Types and Approaches to Their Statistical Synthesis. Value Health. 2012;15(5):639-49.

8. Buxton MJ, et al. Modelling in Ecomomic Evaluation: An Unavoidable Fact of Life. Health Econ. 1997;6(3):217-27.

9. Siebert U, et al. State-Transition Modeling: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. Value Health. 2012;15(6):812-20.

10. Sonnenberg FA, Beck JR. Markov Models in Medical Decision Making. Med Decis Making. 1993;13(4):322-38.

11. Elbasha EH, Chhatwal J. Theoretical Foundations and Practical Applications of Within-Cycle Correction Methods. Med Decis Making. 2016;36(1):115-31.

12. Chhatwal J, et al. Changing Cycle Lengths in State-Transition Models. Med Decis Making. 2016;36(8):952-64.

13. Law AM. Simulation Modeling and Analysis. 4th ed. Singapore: McGraw-Hill Higher Education; 2007.

14. Karnon J, et al. Modeling using Discrete Event Simulation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. Value Health. 2012;15(6):821-7.

15. Karnon J, Haji Ali Afzali H. When to use discrete event simulation (DES) for the economic evaluation of health technologies? A review and critique of the costs and benefits of DES. Pharmacoeconomics. 2014;32(6):547-58.

16. Karnon J. Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation. Health Econ. 2003;12(10):837-48.

17. Simpson KN, et al. Comparison of Markov Model and Discrete-Event Simulation Techniques for HIV. Pharmacoeconomics. 2009;27(2):159-65.

18. Standfield LB, et al. An empirical comparison of Markov cohort modeling and discrete event simulation in a capacity-constrained health care setting. Eur J Health Econ. 2017;18(1):33-47.

19. Franken MD, et al. Cost-effectiveness of capecitabine and bevacizumab maintenance treatment after first-line induction treatment in metastatic colorectal cancer. Eur J Cancer. 2017;75:204-12.

20. Simkens LHJ, et al. Maintenance treatment with capecitabine and bevacizumab in metastatic colorectal cancer (CAIRO3): a phase 3 randomised controlled trial of the Dutch Colorectal Cancer Group. Lancet. 2015;385(9980):1843-52.

21. Eddy DM, et al. Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Value Health. 2012;15(6):843-50.

22. Vemer P, et al. AdViSHE: A Validation-Assessment Tool of Health-Economic Models for Decision Makers and Model Users. Pharmacoeconomics. 2016;34(4):349-61.

23. Delignette-Muller ML, Dutang C. fitdistrplus: An R Package for Fitting Distributions. J Stat Softw. 2015;64(4):1-34.

24. R Core Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016. Available from: https://www.r-project.org/

25. Barton P, et al. The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis. Health Technol Assess. 2004;8(11):104.

26. National Health Care Institute. Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg. 2016. Available from: https://www.zorginstituutnederland.nl/overons/publicaties/publicatie/2016/02/29/richtlijn-voor-het-uitvoeren-van-economische-evaluaties-in-de-gezondheidszorg. Accessed 17 July 2017.

27. Briggs AH, et al. Model Parameter Estimation and Uncertainty Analysis: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group–6. Med Decis Making. 2012;32(5):722-32.

28. Jahn B, et al. Lessons Learned from a Cross-Model Validation between a Discrete Event Simulation Model and a Cohort State-Transition Model for Personalized Breast Cancer Treatment. Med Decis Making. 2016;36(3):375-90.

29. Degeling K, et al. Accounting for parameter uncertainty in the definition of parametric distributions used to describe individual patient variation in health economic models. BMC Med Res Methodol. 2017;17(1):170.

30. Williams C, et al. Estimation of Survival Probabilities for Use in Cost-effectiveness Analyses: A Comparison of a Multi-state Modeling Survival Analysis Approach with Partitioned Survival and Markov Decision-Analytic Modeling. Med Decis Making. 2017;37(4):427-39.

31. Caro JJ, Moller J. Advantages and disadvantages of discrete-event simulation for health economic analyses. Expert Rev Pharmacoecon Outcomes Res. 2016;16(3):327-9.

CHAPTER 4

Comparison of timed automata with discrete event simulation for modeling of biomarker-based treatment decisions: an illustration for metastatic castration-resistant prostate cancer

Koen Degeling
Stefano Schivo
Niven Mehra
Hendrik Koffijberg
Rom Langerak
Johann S. de Bono
Maarten J. IJzerman

This chapter has been published as:
Degeling K, Schivo S, Mehra N, Koffijberg H, Langerak R, de Bono JS, et al. Comparison of timed automata with discrete event simulation for modeling of biomarker-based treatment decisions: an illustration for metastatic castration-resistant prostate cancer. Value in Health. 2017;20(10):1411-9.

ABSTRACT

Background: With the advent of personalized medicine, the field of health economic modeling is being challenged and the use of patient-level dynamic modeling techniques might be required.

Objectives: To illustrate the usability of two such techniques, timed automata (TA) and discrete event simulation (DES), for modeling personalized treatment decisions.

Methods: An early health technology assessment on the use of circulating tumor cells, compared with prostate-specific antigen and bone scintigraphy, to inform treatment decisions in metastatic castration-resistant prostate cancer was performed. Both modeling techniques were assessed quantitatively, in terms of intermediate outcomes (e.g., overtreatment) and health economic outcomes (e.g., net monetary benefit). Qualitatively, among others, model structure, agent interactions, data management (i.e., importing and exporting data), and model transparency were assessed.

Results: Both models yielded realistic and similar intermediate and health economic outcomes. Overtreatment was reduced by 6.99 and 7.02 weeks by applying circulating tumor cell as a response marker at a net monetary benefit of €1033 and €1104 for the TA model and the DES model, respectively. Software-specific differences were observed regarding data management features and the support for statistical distributions, which were considered better for the DES software. Regarding method-specific differences, interactions were modeled more straightforward using TA, benefiting from its compositional model structure.

Conclusions: Both techniques prove suitable for modeling personalized treatment decisions, although DES would be preferred given the current software-specific limitations of TA. When these limitations are resolved, TA would be an interesting modeling alternative if interactions are key or its compositional structure is useful to manage multi-agent complex problems.

INTRODUCTION

There is a long history of modeling in health economics, and its use has had a major impact on resource allocation decisions in various jurisdictions [1,2]. Nevertheless, in the era of personalized or precision medicine (PM), new challenges to the field of health economic modeling (HEM) arise [3]. In particular, modeling of personalized treatment pathways, the use of companion diagnostics, and the likely switches in drug prescriptions are challenging [4,5]. As a consequence, there is a need for models to take into account dynamic behavior and anticipate for patients' and physicians' preferences [6]. This leads to the observation that the most commonly used modeling strategy, that is, (cohort) health state transition modeling [7], with limited flexibility, may not be optimal for handling such complexity [4,8,9], and there is a need for more advanced modeling technologies [10].

One very specific concern in biomarker-guided treatment decisions is the lack of clinical evidence about the utility of diagnostic testing such as genome sequencing and/or the relative effectiveness of sequences of drug treatment [11]. Such issues particularly arise in clinical conditions in which therapy switches are frequent and when such switches are to be informed by precise information about somatic mutations and molecular expressions. One such condition is metastatic castration-resistant prostate cancer (mCRPC), which is the stage in prostate cancer at which metastases are present and medical or surgical castration no longer controls progression of the disease [12]. A recommended treatment strategy from the European Society for Medical Oncology suggests first-line chemotherapy with docetaxel, followed by second-line chemotherapy with cabazitaxel [13,14], although a number of other second- or third-line treatments exist, such as abiraterone, enzalutamide, radium-223, and sipuleucel-T [13,14], which are now increasingly being used before chemotherapy. Because of low treatment response rates [15,16], early recognition of response allows switching from ineffective to possibly effective treatment, which could prevent overtreatment with ineffective and expensive drugs, also reducing treatment-induced morbidity, which may all potentially reduce the health burden associated with overtreatment [17].

Although current decisions to switch treatment are largely based on clinical evaluation, bone scans (BSs), CT scans, and prostate-specific antigen (PSA) tests [18], recent studies indicate that more specific information based on liquid biopsies, for example, circulating tumor cells (CTCs), may be valuable in evaluating treatment response in CRPC [19,20]. A decline in CTCs after treatment has superior performance to PSA, and a composite biomarker panel comprising CTCs is an effective surrogate end point for overall survival [21,22]. The use of this novel molecular information, however, poses a challenge for both physicians and modelers, because currently there are no guidelines for molecular biomarkers and the actual clinical decisions are increasingly based on a combination of imaging biomarkers and laboratory assessments. In addition, both the chance of therapy switches and the transition to other health states are time dependent and largely determined by physician interpretation and patient characteristics.

To assess whether advanced modeling technologies indeed empower modelers to address these challenges, we used two advanced modeling techniques, discrete event simulation (DES) and timed automata (TA), for modeling the personalized treatment process of mCRPC.

The TA modeling paradigm has been developed originally in computer science for modeling distributed systems, such as networks of computers [23]. A typical TA model consists of several automata, which can model agents (e.g., patients or physicians) or other process components (e.g., tests or guidelines). Distinctive of these models is the automata's potential to communicate and interrupt each other, for example, a patient and physician taking a joint decision, while being capable of acting independently. Each automaton can transition between states if prespecified conditions are met (e.g., a treatment ends) or by incentives from other automata (e.g., according to the guideline, a patient should be tested on disease status). The compositional structure of TA has contributed to successful applications of modeling and analyzing complex systems and multi-agent models [24], although to our knowledge, TA has not yet been used for the health economic evaluation of health care interventions. In addition to TA's patient-level discrete handling of time and state transition structure, which naturally represents clinical pathways, its compositional structure and straightforward modeling of interactions are expected to be particularly useful for modeling patient-level and preference-sensitive decisions, such as those present in the dynamic treatment process of mCRPC.

TA's compositional structure and straightforward representation of interactions distinguish it from DES models, which are generally structured around one system or process. In DES the behavior of a system (e.g., clinical pathway) is translated into an ordered sequence of well-defined events [25]. These events correspond to specific changes in the system's state at a specific point in time (e.g., when a test is performed). DES is known from operations research and has been applied in various industries, including health care [26,27], showing to be well suited for modeling clinical processes, including their resources, constraints, and interactions, while making efficient use of computer resources. Although DES is not new to the field of HEM, its value might increase now that models need to take into account the preference-sensitive, patient-level, interactive, and dynamic clinical processes associated with PM [4,5,28]. Particularly its ability to represent dynamic processes and include patient-level histories makes DES suitable for modeling dynamic application of biomarkers, such as the use of CTCs for response monitoring in mCRPC treatment.

The objective of this study was to illustrate the utility of TA and DES for modeling personalized treatment decisions. Both modeling paradigms were assessed qualitatively, on the basis of the population of both models in terms of complexity and data requirements, and quantitatively regarding the simulated intermediate and health economic outcomes. The specific case study on mCRPC treatment was chosen to achieve the aforementioned objective because early response prediction is an unmet clinical need and because incorporation of novel promising biomarkers, such as CTCs, to guide decision processes at the patient level must be evaluated in HEM.

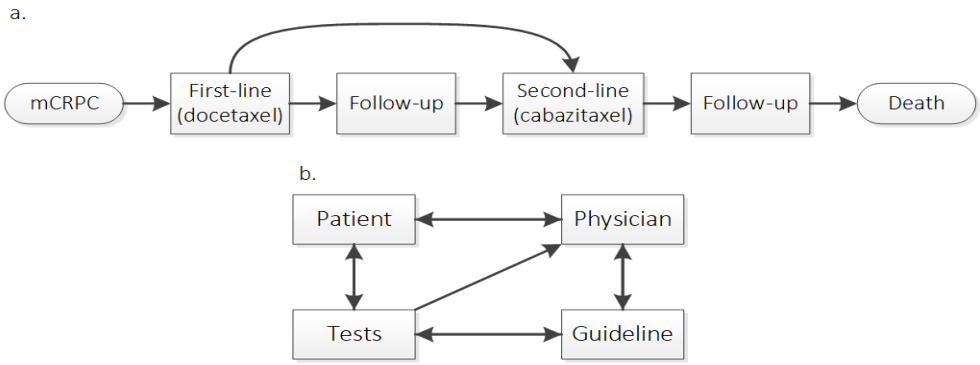


Figure 1. *Model structures: a) schematic overview of the mCRPC treatment process on which the DES model is based and b) schematic overview of the interactions on which the TA model is based. DES, discrete event simulation; mCRPC, metastatic castration-resistant prostate cancer; TA, timed automata.*

METHODS

Case Study

Current guidelines for monitoring treatment response in patients with mCRPC who only have bone metastases involve a combination of PSA and BSs [13,29,30]. Although approved for monitoring disease progression by the US Food and Drug Administration in 1986 [31], the most recent opinion is that PSA alone is not a reliable biomarker of treatment effectiveness [22,32]. Therefore, clinical decisions to switch treatment are commonly made on grounds of PSA, BSs, and clinical progression, although there is a high likelihood of overtreatment or treatment with ineffective drugs.

Liquid biopsies, such as CTCs, have the potential to be used as alternative response markers. CTCs can be detected from blood and have been shown to be prognostic for survival in different tumors [33–35]. CTCs are expected to be useful as a response marker, allowing early detection of drug response or to enable earlier treatment switching from an ineffective treatment, compared with current practice [19,36]. Although the ability of CTCs to associate with survival has been illustrated [19,21,22,37] and CTCs have been approved by the Food and Drug Administration for disease monitoring in advanced disease [38], a clinical trial to evaluate the utility of CTCs as a response marker in mCRPC is yet to be conducted. In advance of the CTC-STOP trial [39], which has recently started and should fill this evidence gap, an early simulation study, that is, in a context in which evidence is limited, was performed to provide an estimate of the potential health economic impact of applying CTCs as a response biomarker in mCRPC treatment.

DES Model

A DES model was developed using AnyLogic multimethod simulation software (AnyLogic, St. Petersburg, Russia) [40]. The model was structured by the four mCRPC process states as shown in Figure 1a, that is, first-line treatment with docetaxel, post-docetaxel follow-up, second-line treatment with cabazitaxel, and post-cabazitaxel follow-up.

Simulated patients' characteristics (e.g., treatment response based on the treatment effectiveness) were assigned at model start-up. Test results were generated on the basis of the patient's response to current treatment and the diagnostic performance of the applicable test. These results were recorded to inform future treatment decisions. The moment at which tests were performed and treatment decisions (i.e., whether to continue or stop current treatment on the basis of progression of the disease) were made was tailored according to the patient's treatment strategy. Treatment decisions were based on test results, guidelines, and physicians' adherence to these guidelines. The simulation engine generated events to model nonprogression-related treatment interruptions and survival.

To obtain health outcomes, the simulated patients' utility was determined on the basis of the time spent in the process states and the patients' response to the treatments. Total costs were obtained by adding costs to events in the patients' pathways, such as initiating a new treatment cycle or performing a test. The simulation terminated when all patients had died, which was after approximately 3 years. For the analysis of each scenario (see Cost-Effectiveness Analysis section), results were averaged over 100 runs of 100,000 patients, with different random number seeds for each run.

The DES model was validated by ensuring no unnecessary detail was present, structured "walkthroughs," expert consultation, comparing results with calculations by hand, extreme value analysis, trace analysis, sensitivity analysis, and cross validation with the TA model [41,42]. Additional external validation, that is, comparing simulation outcomes with those observed in clinical practice, is valuable when performing modeling studies. The required evidence to do so was, however, not available for the case study, because the CTC-STOP trial is the first to study the impact of early treatment switches on clinical outcomes in mCRPC treatment.

Detailed descriptions of several of the DES model's key components are available in the Supplemental Materials.

The TA Model

The TA model was developed and analyzed using the UPPAAL, Uppsala, Sweden software [43]. The model is based on interactions between the automata representing the patient, physician, test(s), and test guideline (Figure 1b). The basic structure of the automaton representing a patient follows the mCRPC process as shown in Figure 1a. This patient automaton was enriched with additional states and transitions to represent the probabilistic choices for treatment effectiveness and patient survival (see Figure S1 in Appendix 1).

The patient automaton interacts with both the physician and the test automata (see Figures S2 and S3 in Appendix 1). Whenever a test is performed, the patient automaton synchronizes with the test automaton, reporting the effectiveness of the current treatment. The test guideline automaton defines the periodicity of the tests, activating the test automaton each time a new test needs to be performed (see Figure S4 in Appendix 1). The physician automaton is notified of the result by a message from the test automaton. If a decision needs to

be taken according to the test guideline automaton, the physician automaton performs the needed computations, which leads to a treatment decision (i.e., whether to stop or continue current treatment). If the current treatment is stopped, the physician automaton sends a message to the patient automaton, making the patient automaton exit from the current treatment state (going from first-line treatment to second-line treatment or from second-line treatment to second follow-up).

Variables are used to trace costs and treatment periods. The amount of time a patient spends in a given state is measured with particular "clock" variables that automatically increment only when an automaton is in the intended states. The simulation ended when the patient automaton reached the dead state, or when more than 500 weeks were spent in the model. Final health and economic outcomes were computed on the basis of the simulation of 1 million individual patients.

The TA model was validated by ensuring no unnecessary detail was present, structured "walkthroughs," comparing results with calculations by hand, extreme value analysis, trace analysis, model checking, sensitivity analysis, and cross validation with the DES model [41,42].

Because the model contains probabilistic events, we analyzed it with UPPAAL's statistical model-checking engine [44], which computes a (large) number of simulation runs on the model to answer queries with a certain prespecified confidence. In other words, this model-checking engine determines the probability of user-defined scenarios.

Detailed descriptions of several of the TA model's key components are available in the Supplemental Materials.

Data Requirements in Both Models

To model the applicability of CTCs in mCRPC treatment monitoring, data needed to be acquired from various sources. First- and second-line response rates were derived from the original randomized controlled trials of docetaxel and cabazitaxel [15,16]. Nonprogression-related stop rates (i.e., clinical progression and toxicity) and survival rates were estimated on the basis of the results of these randomized controlled trials. Because no evidence was available on the effect of the timing of treatment switches on survival, survival was assumed to be equal for both treatment strategies. The diagnostic performance of the BS was derived from literature [45]. The diagnostic performance of more than 50% PSA decline, with regard to appropriately diagnosing progression, was estimated on the basis of data of the FIRSTANA (NCT01308567) and PROSELICA (NCT01308580) studies.

The diagnostic performance of CTC enumeration was estimated on the basis of data of the IMMC38 (NCT00133900) and COU-AA301 (NCT00638690) studies [21]. Physicians' behavior, that is, the likelihood that physicians would follow guideline-suggested treatment switches, was based on discussion with clinical experts [46]. Decision moments and decision rules (to define progression by response marker) were obtained from the protocol of the CTC-STOP trial. The patients' quality of life during treatment with docetaxel and cabazitaxel was esti-

mated on the basis of a phase II trial on docetaxel [47]. Treatment costs were derived from a Canadian study on the costs of mCRPC treatment [48] and were converted from Canadian dollars in 2013 to euros in 2015 using the inflation rate of 1.5% in 2013 and 0.5% in 2014 [49] and the currency rate of €0.7 per Canadian dollar in 2015 [50]. The costs of a PSA test, BS, and physician consultation were derived from the price list of the Dutch Healthcare Authority for 2015 [51]. The cost of CTC enumeration was estimated on the basis of experts' opinion. An overview of all input data is presented in Table 1.

Table 1. *Evidence overview, including low and high values for sensitivity analysis.*

Input Parameter	Base value	Low value	High value	Source
Treatments				
Docetaxel effectiveness	45%	36%	54%	[15]
Stop-chance docetaxel (non-progression) per week[1]	0.005686	0.004549	0.006823	[15]
Median time to progression after docetaxel in weeks[1]	3	2[4]	4[4]	[15]
Cabazitaxel effectiveness	39.20%	31%	47%	[16]
Stop-chance cabazitaxel (non-progression) per week[1]	0.013010	0.010408	0.015612	[16]
Diagnostic Performance				
PSA test[6]				
Sensitivity	51%	40.8%[3]	61.2%[3]	-
Specificity	55%	44%[3]	66%[3]	-
Bone scan				
Sensitivity	69%	55%[3]	83%[3]	[45]
Specificity	64%	77%[3]	51%[3]	[45]
CTC analysis[6]				
Sensitivity	77.1%	61.7%[3]	92.5%[3]	-
Specificity	62.8%	50.2%[3]	75.4%[3]	-
Decision Making				
Physicians' obedience to guidelines (PSA + BS)[2]	85%	68%	100%[5]	-
Physicians' obedience to guidelines (CTC)[2]	65%	52%	78%	-
Health Outcomes				
Probability of death during docetaxel per week[1]	0.009909	0.007927	0.011891	[15]
Probability of death during cabazitaxel (<= 25 months) per week[1]	0.012013	0.009610	0.014416	[16]
Probability of death during cabazitaxel (> 25 months) per week[1]	0.233323	0.186658	0.279988	[16]

Table 1 (continued). *Evidence overview, including low and high values for sensitivity analysis.*

Input Parameter	Base value	Low value	High value	Source
QoL during docetaxel and cabazitaxel (responders)[1]	0.705	0.564	0.846	[47]
QoL during docetaxel and cabazitaxel (non-responders)[1]	0.529	0.423	0.635	[47]
QoL during follow-up (responders)[1]	0.881	0.705	1.000[5]	[47]
QoL during follow-up (non-responders)[1]	0.705	0.564	0.846	[47]
Costs				
Docetaxel per cycle	€ 428.00	€ 342.40	€ 513.60	[48]
Cabazitaxel per cycle	€ 4,170.00	€ 3,336.00	€ 5,004.00	[48]
PSA test	€ 9.11	€ 7.29	€ 10.93	[51]
Bone scan	€ 285.36	€ 228.29	€ 342.43	[51]
CTC analysis[2]	€ 500.00	€ 400.00	€ 600.00	-
Physician consult	€ 65.17	€ 52.14	€ 78.20	[51]
Treatment Decisions				
Docetaxel and cabazitaxel cycle duration in weeks	3	-	-	[15, 16]
Maximum amount of docetaxel treatment in cycles	10	-	-	[15, 16]
First decision moment: after … cycles (control arm)	4	-	-	-[6]
First decision moment: after … cycles (exp. arm)	3	-	-	-[6]
Number of PSA tests taken into account per decision	3	-	-	-[6]
Number of PSA tests that need to indicate progression	2	-	-	-[6]
Number of Bone scans taken into account per decision	1	-	-	-[6]
Number of Bone scans that need to indicate progression	1	-	-	-[6]
Number of CTC analyses taken into account per decision	2	-	-	-[6]
Number of CTC analyses that need to indicate progression	2	-	-	-[6]

[1]*Estimated from literature.*
[2]*Assumption.*
[3]*Assumed that when sensitivity decreases with 20%, specificity increases with 20% and vice versa.*
[4]*Manually rounded so that alternative scenarios arise.*
[5]*Maximum of 100% or 1.*
[6]*Estimated, see Methods section.*

Cost-Effectiveness Analysis

Patient-level pathways and health and economic outcomes were simulated and average outcomes were calculated for two strategies, that is, response monitoring by PSA and BS (control) versus CTCs (experimental), using both models. The primary outcome measure was the cost-utility expressed in incremental costs per quality-adjusted life-year (QALY) gained. The incremental cost-effectiveness ratio, net health benefit (NHB), and net monetary benefit (NMB) were used to compare the treatment strategies. Secondary outcome measures included the amount of treatment, the amount of overtreatment, and the proportion of patients receiving second-line treatment with cabazitaxel. Because of the short survival in both models, no discounted costs and effects were included in the analyses. A willingness to pay of €50,000/QALY gained was applied for calculating the NHB and NMB, on the basis of the Dutch guidelines for cost-effectiveness analyses [52].

Univariate sensitivity analysis was performed to assess the effect of uncertainty surrounding the input parameters on the primary outcome measure. No probabilistic sensitivity analysis was performed, because the available evidence was insufficient to express the uncertainty surrounding the accumulated, estimated, and assumed parameter values. The treatment decision– related parameters, such as the decision moments and the number of tests, were excluded from the sensitivity analysis. For each considered parameter, low and high values were calculated as 20% of the base-case value (Table 1). To define a low and high value scenario for the diagnostic performance of the tests, the sensitivity of a test was used as reference parameter and the specificity was assumed to change inversely with sensitivity (i.e., when sensitivity decreased, specificity increased and vice versa).

RESULTS

Implementation of the therapy switches is possible using both modeling paradigms. Although both models require and use the same evidence (Table 1), adapting the evidence to the model and recording the output from the model are considered more efficient for the DES model. The reason for this is that AnyLogic, the software environment used for implementing the DES model, supports data management by importing and exporting data from and to Microsoft Excel and text files, whereas this is not supported by the UPPAAL software. Nevertheless, the parameter variation experiments in the TA model are performed by a script communicating the different scenarios to the UPPAAL software and recording the outcomes from the simulations.

Both models yield comparable results for the base-case scenario (Figure 2 and Table 2). The largest difference is observed in the cost-utility ratio, which is €921/QALY gained for the DES model (so costs are lower for the experimental strategy) and €15,233/QALY gained for the TA model. Conversely, differences are substantially smaller for the NHB and NMB, which are 0.022 QALY and €1104 and 0.021 QALY and €1033 for the DES model and the TA model, respectively. In the control strategy, the average treatment costs per patient are €22,870 and €23,393 for the DES model and the TA model, respectively. In the experimental strategy, the average treatment costs per patient are €22,850 and €23,846 for the DES model and the TA

model, respectively. On average, patients' survival is approximately 69 weeks for both modeling methods. The average utilities are 0.664 and 0.681 for the DES model and 0.692 and 0.714 for the TA model in the control and the experimental strategy, respectively.

The secondary outcome measures explain the clinical improvements that occur when CTCs are used as a response marker. In the experimental arm, patients were approximately 10 (DES) and 11 (TA) percentage points more likely to start second-line treatment with cabazitaxel (Table 2). Furthermore, using CTCs as a response marker reduces the average amount of first- and second-line overtreatment. Expected first-line overtreatment with docetaxel per patient decreases by approximately 5 weeks (i.e., almost two treatment cycles) in both models, and overtreatment with cabazitaxel is reduced by approximately 2 weeks in both models. These reductions in overtreatment do not only reduce the health burden associated with overtreatment but also compensate the higher costs of the CTC analysis and the increase in patients receiving second-line treatment with cabazitaxel.

The univariate sensitivity analysis shows comparable results for the NMB for both modeling methods (Figure 3). Tornado plots for the NHB, incremental costs, and the incremental effectiveness are shown in Figures S5 to S7 in Appendix 1. Although the effect of varying the parameter values is consistent between both methods, the magnitude of the effect differs between the modeling methods. The characteristics of the CTC analysis, that is, the diagnostic performance and the costs, have the largest impact on the cost-effectiveness of this technology's application in the treatment process of mCRPC. Other parameters with a substantial influence on the cost-effectiveness are the physicians' behavior and certain treatment characteristics. In Table 3 it can be seen that if physicians' behavior is not considered, approximately 0.03 QALYs can be gained at lower costs by applying CTC as a response marker (NHB 0.061 QALY [DES] and 0.056 QALY [TA]).

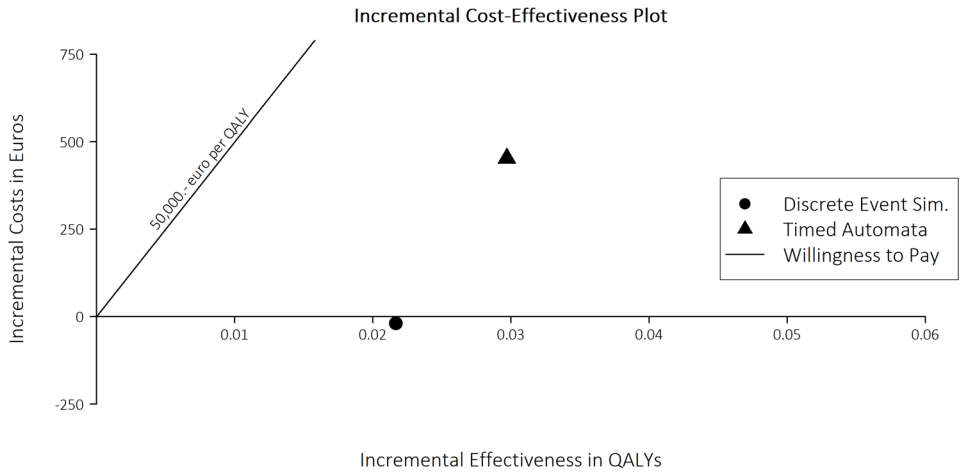


Figure 2. *Incremental cost-effectiveness plot. The cost-effectiveness of the experimental strategy was compared with that of the control strategy at a willingness to pay of €50,000/QALY for both modeling methods. QALY, quality-adjusted life-year.*

Table 2. *Primary and secondary outcomes for the base-case values of the input parameters described in Table 1. ICER, incremental cost-effectiveness ratio; NHB, net health benefit; NMB, net monetary benefit; QALY, quality-adjusted life-year.*

Outcome Measure	Strategy	Discrete Event Simulation	Timed Automata
ICER (€/QALY)	-	€-921.-	€ 15,233
NMB (€)	-	€-1,104.-	€-1,033.-
NHB (QALYs)	-	0.022	0.021
Total Costs (€)	Control	€ 22,870.-	€ 23,393.-
	Exp.	€ 22,850.-	€ 23846.-
	Δ	€-20.-	€ 453.-
Effectiveness (QALYs)	Control	0.89	0.92
	Exp.	0.91	0.95
	Δ	0.02	0.03
Survival (wk)	Control	69.4	69.5
	Exp.	-	-
Utility (0-1)	Control	0.664	0.692
	Exp.	0.681	0.714
Docetaxel Treatment (wk)	Control	22.6	22.3
	Exp.	16.0	16.0
Docetaxel Overtreatment (wk)	Control	12.1	11.8
	Exp.	7.0	7.0
Cabazitaxel Initiated (%)	Control	65.1	64.8
	Exp.	75.3	75.4
Cabazitaxel Treatment (wk)	Control	13.4	12.8
	Exp.	10.8	10.7
Cabazitaxel Overtreatment (wk)	Control	6.7	7.6
	Exp.	4.8	5.4

DISCUSSION

We have shown that preference-sensitive and biomarker-based treatment decisions, including multi-agent interactions and different levels of abstraction according to the available evidence, can be modeled using either DES or TA. Both models yielded realistic intermediate and health economic outcomes, although differences were observed regarding both the quantitative and qualitative aspects of the comparison.

Regarding the quantitative outcomes, the large difference in the incremental cost-effectiveness ratio may not reflect a relevant overall difference between the models, because this ratio is very sensitive to small incremental effects (denominator). The NMB and NHB, there-

Table 3. *Primary and secondary outcomes for the scenario in which physicians' behavior is not considered. ICER, incremental cost-effectiveness ratio; NHB, net health benefit; NMB, net monetary benefit; QALY, quality-adjusted life-year.*

Outcome Measure	Strategy	Discrete Event Simulation	Timed Automata
ICER (€/QALY)	-	€-59,217.-	€-26,204
NMB (€)	-	€-3,046.-	€-2,819.-
NHB (QALYs)	-	0.061	0.056
Total Costs (€)	Control	€ 22,813.-	€ 23,294.-
	Exp.	€ 21,161.-	€ 22,325.-
	Δ	€-1,651.-	€-970.-
Effectiveness (QALYs)	Control	0.89	0.92
	Exp.	0.91	0.96
	Δ	0.02	0.04
Survival (wk)	Control	69.3	69.2
	Exp.	-	-
Utility (0-1)	Control	0.665	0.693
	Exp.	0.686	0.721
Docetaxel Treatment (wk)	Control	22.3	21.8
	Exp.	14.3	14.3
Docetaxel Overtreatment (wk)	Control	11.8	11.4
	Exp.	5.9	5.9
Cabazitaxel Initiated (%)	Control	65.5	65.6
	Exp.	77.9	78.2
Cabazitaxel Treatment (wk)	Control	13.4	12.7
	Exp.	10.0	10.0
Cabazitaxel Overtreatment (wk)	Control	6.7	7.4
	Exp.	4.3	4.9

fore, give a more realistic representation of the differences in health economic outcomes and these are, in fact, similar for both methods. With respect to the intermediate simulation outcomes (e.g., the amount of overtreatment), both modeling methods also yielded similar results. Although evidence is limited, especially regarding clinical benefits of early treatment switches, and no conclusions on the cost-effectiveness can be drawn, the promising reduction in overtreatment warrants further research on applying CTC as a response marker in mCRPC treatment. Interestingly, the results additionally show that physicians' adherence to guidelines, regarding the use and interpretation of biomarkers, influences health economic outcomes substantially, highlighting the need to carefully check any assumptions implemented for taking into account these kinds of behavior.

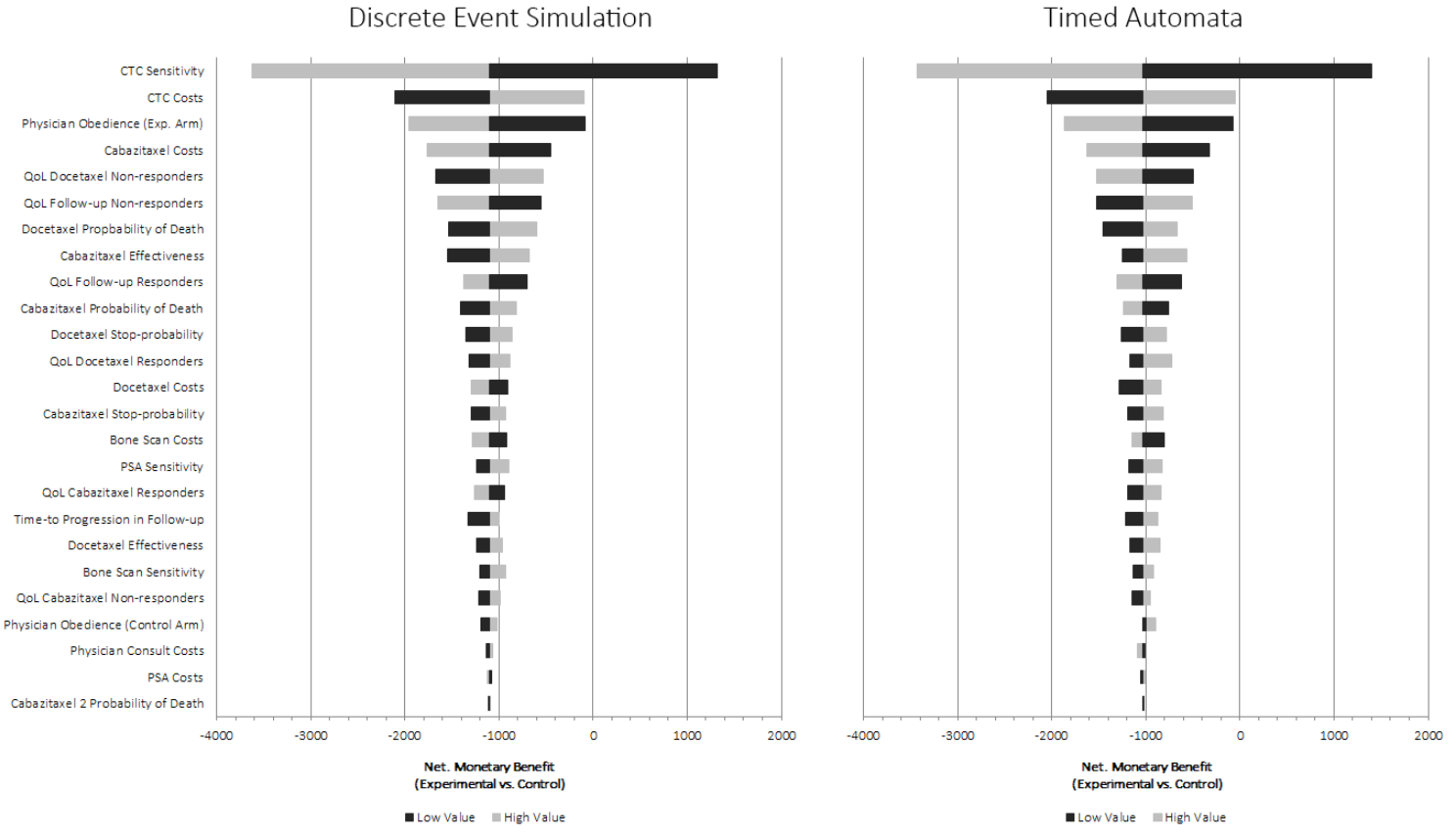


Figure 3. *Tornado plot showing the impact of the input parameters on the cost-effectiveness expressed as the net monetary benefit. The parameter values used for this analysis can be found in Table 1. CTC, circulating tumor cell; PSA, prostate-specific antigen; QOL, quality of life.*

Although an exhaustive quantitative comparison goes beyond the objective to illustrate the utility of TA and DES for modeling personalized treatment decisions, there are several software- and method-specific characteristics that may explain the quantitative differences observed. A potential cause regards the selection of patients with respect to competing risks (e.g., death and stopping treatment). Although AnyLogic provides modelers with complete control over the selection of events on the basis of conditions, UPPAAL does not allow the selection of competing risks to be ordered in a branch (i.e., probabilistic transition). Another possible software-specific cause would be the generation of random numbers. Nevertheless, an additional simulation experiment showed that there are no meaningful differences in random numbers generated in both software environments. The interaction between agents may be a plausible method-specific cause of the minor observed differences, because an ongoing event in one automaton can be interrupted by other automata, whereas this type of interaction is not typically included in DES models. Further research based on other case studies is required to further clarify the origin of the quantitative differences between both modeling paradigms.

Despite the fact that no meaningful differences were observed with regard to the generation of random numbers, other software-specific particularities were observed. UPPAAL's model-checking engine supports formal validation of the model structure, for example, by automated testing for the occurrence of (unrealistic) clinical pathways and simulation outcomes. Such formal validation checks are not provided by other software environments, in which model consistency has to be checked manually, and may therefore not be exhaustive. Conversely, UPPAAL currently supports only uniform and exponential distributions, requiring complex and error-prone workarounds to use other types of distributions (e.g., preprocessing based on phase-type distributions [53]). Another limitation of UPPAAL is the inefficient management of experiments and data, that is, importing data and exporting results, because both need to be performed manually or by using scripts, which may introduce errors. Because UPPAAL currently is the only software environment available for implementing TA, these limitations heavily impact the usability of TA as a method. Nevertheless, UPPAAL was not developed and has never been used as a dedicated tool for HEM. Current issues might be resolved to increase its usefulness for HEM. Other software environments (e.g., R Statistical Software [54] or AnyLogic) may also support the implementation of TA in the future, which may solve current software-specific issues.

A major advantage of TA compared with DES is that it allows separation of the main process or problem into multiple smaller problems (i.e., automata), thereby substantially increasing the model's cognitive ease, as was learned from discussing the models with clinical experts of the CTC-STOP trial and at conferences. This compositional nature of TA also benefits modelers, because it facilitates the separation of complex concerns among different actors or entities (e.g., tests or guidelines) in the model and makes modeling interactions between agents and these entities more straightforward compared with DES. Although both methods represent time discretely and cognitive ease strongly relates to end users' preferences, the compositional nature and ease of modeling interactions clearly distinguish TA from DES.

Given the aforementioned current software-specific limitations to implementing TA, DES would be the preferred method. Nevertheless, if these software-specific limitations are resolved, TA might be preferred over DES, in particular if interactions between agents are key or its compositional multi-agent structure is useful to manage complex problems.

Our study has certain limitations. The existing evidence gaps regarding the case study are a limitation not only to the health economic evaluation of CTC as a response marker in mCRPC, but to some extent also to the model comparison, because models based on more evidence are likely to become more complex. Furthermore, the generalizability of the results is unknown, given the application of a single case study, and the qualitative aspects of the comparison may be subjective.

In addition to in-depth case study–based quantitative comparisons of TA and DES, further research might be directed toward the extent to which the challenges associated with PM can be addressed for more complex clinical processes, while maintaining model transparency, using either DES or TA. Furthermore, future studies might focus on modeling the physicians' and patients' interactions and preferences in detail. Finally, the potential value of UPPAAL's optimization features, which have been applied in other environments [55], could be explored in the health economic context.

CONCLUSIONS

DES and TA have both been shown to be suitable for modeling preference-sensitive, dynamic, and biomarker-based treatment decisions. Although intermediate and health economic outcomes for the case study were similar for both models, differences were observed regarding several method- and software-specific aspects. Regarding the method-specific aspects, the compositional multi-agent structure of TA may be preferred if interactions between agents are key or complex (multi-agent) problems need to be disaggregated. Nevertheless, although UPPAAL's model-checking engine allows automated and formal model validation, two software-specific limitations strongly impact the usability of TA for HEM: the limited number of supported parametric distributions and the inefficient management of data and experiments. Given the current software-specific limitations to implementing TA, DES is likely to remain the preferred method for modeling personalized treatment decisions in the near future.

ACKNOWLEDGMENTS

We thank Leon Terstappen for his technical expertise and David Lorente for providing valuable information on the diagnostic performance of CTC enumeration.

FUNDING SOURCES AND CONFLICTS OF INTEREST

This research was supported by the CTCTrap project, which was funded by the FP7 health.2012.1.2-1 (grant no. 305341). In addition, J. S. de Bono acknowledges support from Movember and Prostate Cancer UK to the Movember London Prostate Cancer Centre of Ex-

cellence, from a Cancer Research UK Centre grant, and from a Biomedical Research Centre (BRC) grant to The Royal Marsden and the Institute of Cancer Research. The authors declare no conflict of interest.

SUPPLEMENTARY MATERIALS

Supplementary material related to this article can be found in the online version at: http://dx.doi.org/10.1016/j.jval.2017.05.024

REFERENCES

1. Ramsey S, et al. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force Report. Value Health 2005;8(5):521–33.

2. National Institute for Health and Care Excellence. The Guidelines Manual. London: National Institute for Health and Care Excellence, 2012.

3. Degeling K, et al. A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models. Exp Rev Pharmacoecon Outcomes Res 2017;17(1):17–25.

4. Annemans L, et al. Current methodological issues in the economic assessment of personalized medicine. Value Health 2013;16(S6): S20–6.

5. Phillips KA, et al. Challenges to the translation of genomic information into clinical practice and health policy: utilization, preferences, and economic value. Curr Opin Mol Ther 2008;10(3):260–6.

6. Marshall DA, et al. Applying dynamic simulation modeling methods in health care delivery research—the SIMULATE checklist: report of the ISPOR Simulation Modeling Emerging Good Practices Task Force. Value Health 2015;18(1):5–16.

7. Siebert U, et al. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. Value Health 2012;15(6):812–20.

8. Ferrusi IL, et al. Do economic evaluations of targeted therapy provide support for decision makers? J Oncol Pract 2011;7(S3):36s–45.

9. Miller JD, et al. Current challenges in health economic modeling of cancer therapies: a research inquiry. Am Health Drug Benefits 2014;7(3):153–62.

10. IJzerman M, et al. Implementation of comparative effectiveness research in personalized medicine applications in oncology: current and future perspectives. Comp Eff Res 2015;5:65–72.

11. Messner DA. Evaluating the comparative effectiveness of treatment sequences in oncology: a new approach. J Comp Eff Res 2015;4(6):537–40.

12. Kirby BJ, et al. Functional characterization of circulating tumor cells with a prostate-cancer-specific microfluidic device. PLoS One 2012;7(4):e35976.

13. Heidenreich A, et al. EAU guidelines on prostate cancer, part II: treatment of advanced, relapsing, and castration-resistant prostate cancer. Eur Urol 2014;65(2):467–79.

14. Horwich A, et al. Prostate cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2013;24(S6):vi106–14.

15. Tannock IF, et al. Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. N Engl J Med 2004;153(15):1502–12.

16. de Bono JS, et al. Prednisone plus cabazitaxel or mitoxantrone for metastatic castration-resistant prostate cancer progressing after docetaxel treatment: a randomised open-label trial. Lancet 2010;376:1147–54.

17. Silberstein JL, et al. Current clinical challenges in prostate cancer. Transl Androl Urol 2013;2(3):122–36.

18. Fitzpatrick JM, et al. Optimal management of metastatic castration-resistant prostate cancer: highlights from a European expert consensus panel. Eur J Cancer 2014;50(9):1617–27.

19. de Bono JS, et al. Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer. Clin Cancer Res 2008;14(19):6302–9.

20. Miyamoto DT, et al. Circulating tumour cells-monitoring treatment response in prostate cancer. Nat Rev Clin Oncol 2014;11(7):401–12.

21. Lorente D, et al. Decline in circulating tumor cell count and treatment outcome in advanced prostate cancer. Eur Urol 2016;70(6):985–92.

22. Scher HI, et al. Circulating tumor cell biomarker panel as an individual-level surrogate for survival in metastatic castration-resistant prostate cancer. J Clin Oncol 2015;33(12):1348–55.

23. Kaynar DK, et al. The Theory of Timed I/O Automata (2nd ed). San Rafael: Morgan & Claypool Publishers, 2010.

24. UPPAAL. 2016. Available from: http://www.uppaal.org/.

25. Law AM. Simulation Modeling and Analysis (4th ed). Singapore: McGraw-Hill Higher Education, 2007.

26. Karnon J. Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation. Health Econ 2003;12(10):837–48.

27. Brennan A, et al. A taxonomy of model structures for economic evaluation of health technologies. Health Econ 2006;15(12):1295–310.

28. Brailsford S, Schmidt B. Towards incorporating human behaviour in models of health care systems: an approach using discrete event simulation. Eur J Oper Res 2003;150(1):19–31.

29. Tait C, et al. Quantification of skeletal metastases in castrate-resistant prostate cancer predicts progression-free and overall survival. BJU Int 2014;114(6b):E70–3.

30. Dennis ER, et al. Bone scan index: a quantitative treatment response biomarker for castration-resistant metastatic prostate cancer. J Clin Oncol 2012;30(5):519–24.

31. Kuriyama M, et al. Use of human prostate-specific antigen in monitoring prostate cancer. Cancer Res 1981;41(10):3874–6.

32. Karzai FH, et al. Beyond PSA: managing modern therapeutic options in metastatic castration-resistant prostate cancer. South Med J 2015;108(4):224–8.

33. Rack B, et al. Circulating tumor cells predict survival in early average-to-high risk breast cancer patients. J Natl Cancer Inst 2014;106(5):1–11.

34. Thalgott M, et al. Circulating tumor cells versus objective response assessment predicting survival in metastatic castration-resistant prostate cancer patients treated with docetaxel chemotherapy. J Cancer Res Clin Oncol 2015;141(8):1457–64.

35. Romiti A, et al. Circulating tumor cells count predicts survival in colorectal cancer patients. J Gastrointestin Liver Dis 2014;23(3):279–84.

36. Shaffer DR, et al. Circulating tumor cell analysis in patients with progressive castration-resistant prostate cancer. Clin Cancer Res 2007;13(7):2023–9.

37. Goldkorn A, et al. Circulating tumor cell counts are prognostic of overall survival in SWOG S0421: a phase III trial of docetaxel with or without atrasentan for metastatic castration-resistant prostate cancer. J Clin Oncol 2014;32(11):1136–42.

38. Hu B, et al. Circulating tumor cells in prostate cancer. Cancer 2013;5(4):1676–90.

39. NHS Health Research Authority. CTC-STOP: utilising circulating tumour cell (CTC) counts to optimize systemic therapy of metastatic prostate cancer. Available from: http://www.hra.nhs.uk/news/researchsummaries/ctc-stop/. [Accessed March 6, 2017].

40. The AnyLogic Company. AnyLogic multi-method simulation software. Available from: http://www.anylogic.com/.

41. Eddy DM, et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Value Health 2012;15(6):843–50.

42. Vemer P, et al. AdViSHE: a validation-assessment tool of health-economic models for decision makers and model users. Pharmacoeconomics. 2016;34(4):349–61.

43. Behrmann G, et al. A tutorial on UPPAAL. In: Bernardo M, Corradini F, eds. Formal Methods for the Design of Real-Time Systems. Berlin, Germany: Springer-Verlag, 2004.

44. David A, et al. UPPAAL SMC tutorial. Int J Softw Tools Technol Transf 2015;17(4):397–415.

45. Even-Sapir E, et al. The detection of bone metastases in patients with high-risk prostate cancer: 99mTc-MDP planar bone scintigraphy, single- and multi-field-of-view SPECT, 18F-Fluoride PET, and 18F-Fluoride PET/CT. J Nucl Med 2006;47(2):287–97.

46. Lorente D, et al. Interrogating metastatic prostate cancer treatment switch decisions: a multi-institutional study. Eur Urol Focus. 2018;4(2):235-244.

47. Caffo O, et al. Impact of docetaxel-based chemotherapy on quality of life of patients with castration-resistant prostate cancer: results from a prospective phase II randomized trial. BJU Int 2011;108(11):1825–32.

48. Dragomir A, et al. Drug costs in the management of metastatic castration-resistant prostate cancer in Canada. BMC Health Serv Res 2014;14(252):1–14.

49. Trading Economics. Euro area inflation rate. Available from: http:// www.tradingeconomics.com/euro-area/inflation-cpi. [Accessed May 6, 2015].

50. XE. XE currency charts: CAD to EUR. Available from: http://www.xe.com/currencycharts/?from=CAD&to=EUR&view=1Y. [Accessed May 6, 2015].

51. The Dutch Healthcare Authority. Tarieventabel DBC-zorgproducten en overige producten. Available from: https://www.nza.nl/98174/149057/Tarieventabel-DBC-zorgproducten-en-overige-producten-per-1-januari_2015.xls. [Accessed May 11, 2015].

52. Dutch National Health Care Institute. Kosten-effectiviteit in de praktijk. Available from: https://www.zorginstituutnederland.nl/publicaties/rapport/2015/06/26/kosteneffectiviteit-in-de-praktijk. [Accessed February 10, 2017].

53. Altiok T. On the phase-type approximations of general distributions. IIE Trans 1985;17(2):110–6.

54. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2016.

55. Ahmad W, et al. Green computing: power optimisation of VFI-based real-time multiprocessor dataflow applications. Amersfoort: Presented at: ICT OPEN, 2015.

APPENDIX 1: supplementary figures

High-quality images of the supplementary figures presented in this appendix are available online at: http://dx.doi.org/10.1016/j.jval.2017.05.024

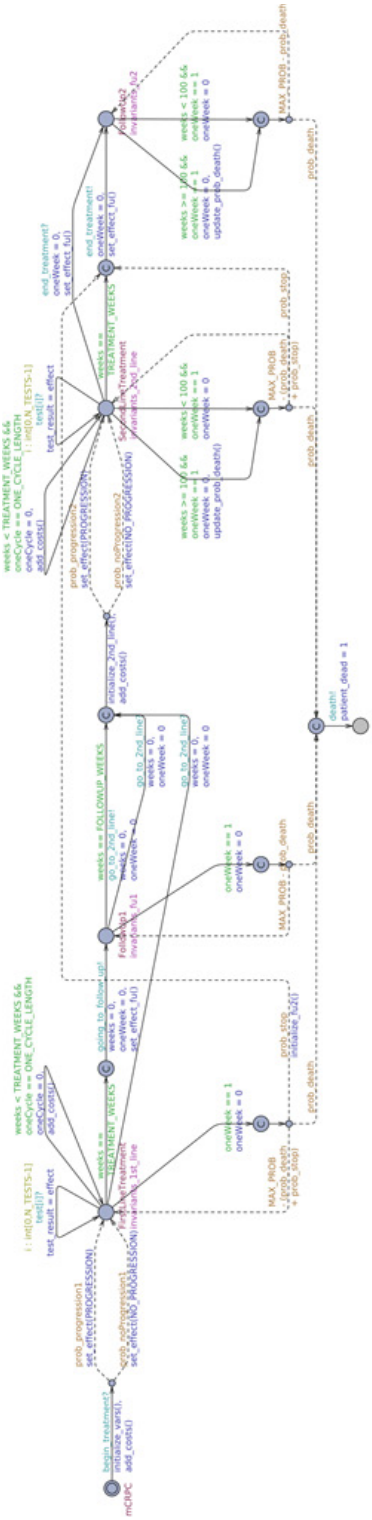


Figure S1. *Graphical representation of the Patient automaton.*

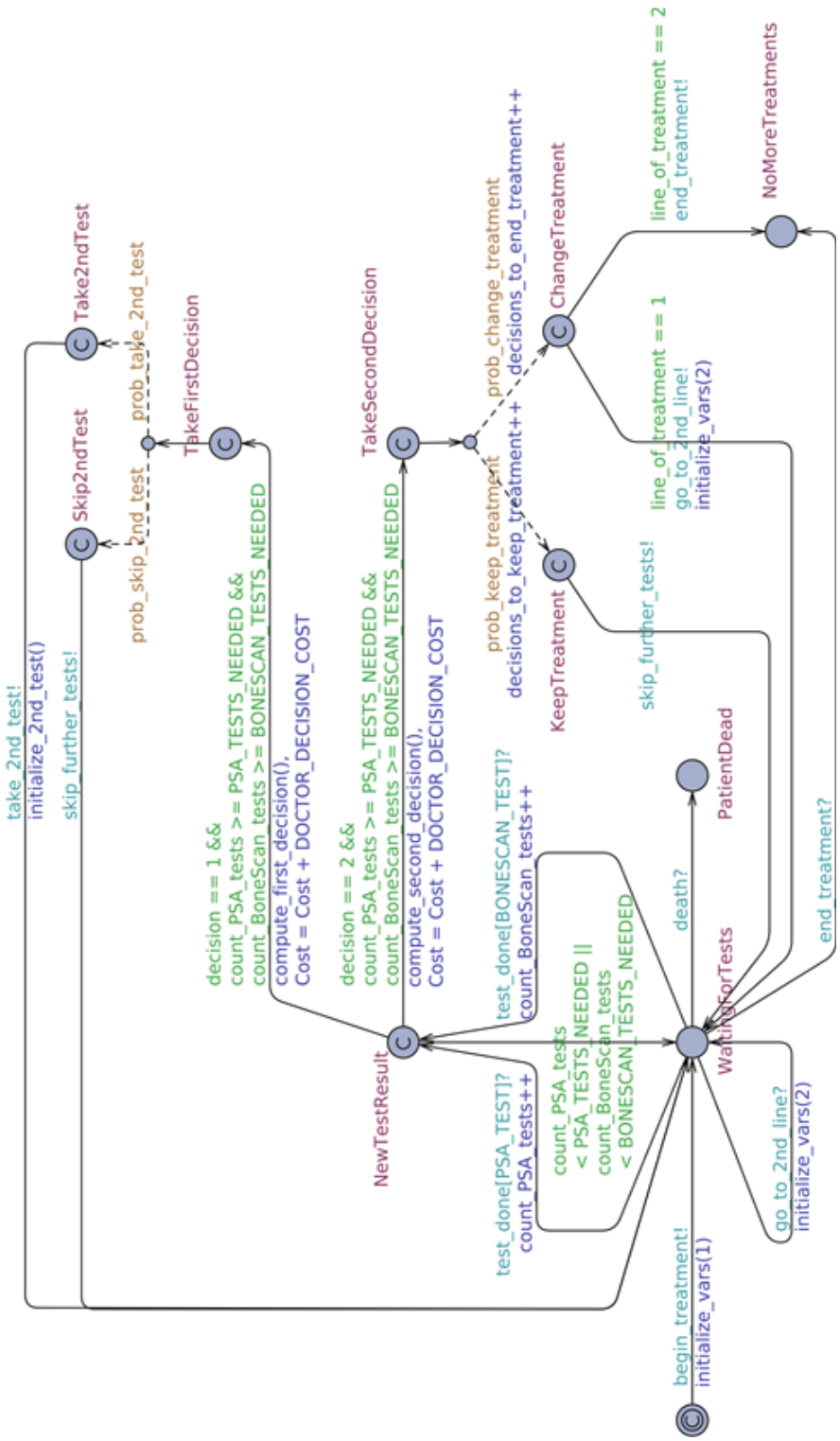


Figure S2a. *Graphical representation of the Physician automaton for the control strategy.*

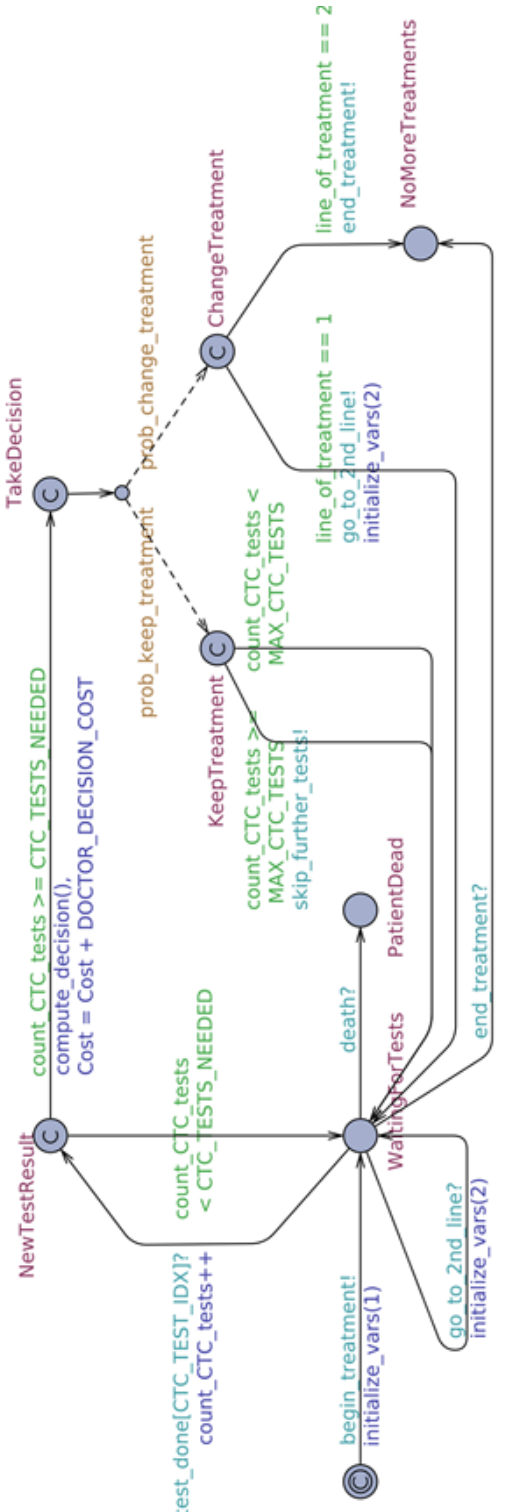


Figure S2b. *Graphical representation of the Physician automaton for the experimental strategy.*

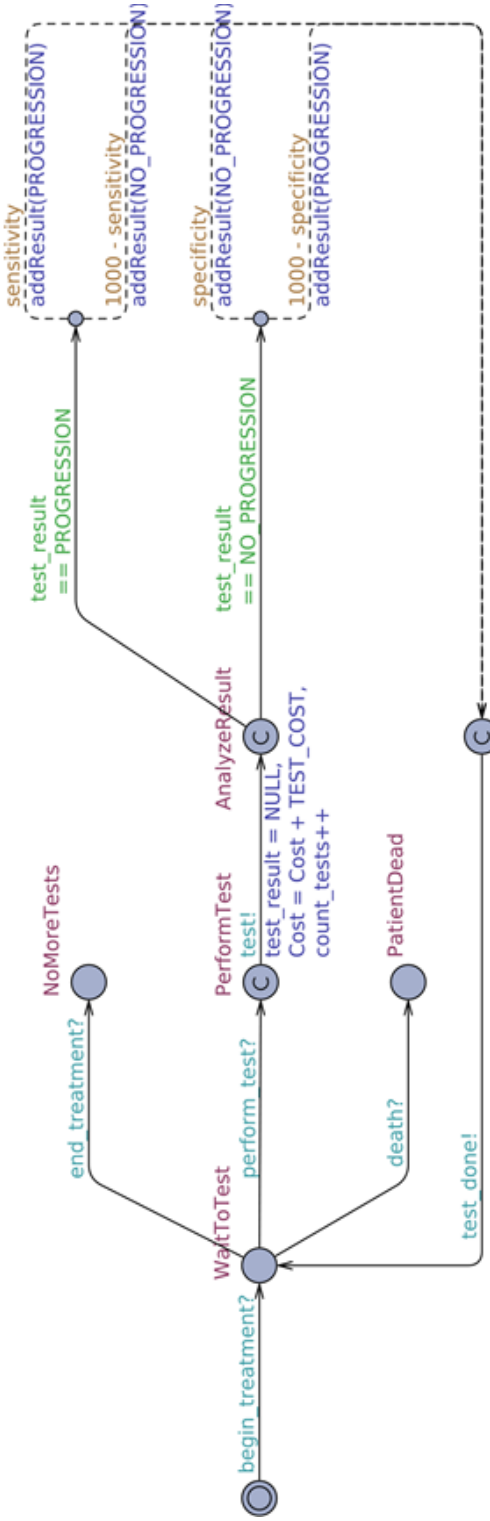


Figure S3. *Graphical representation of the Test automaton.*

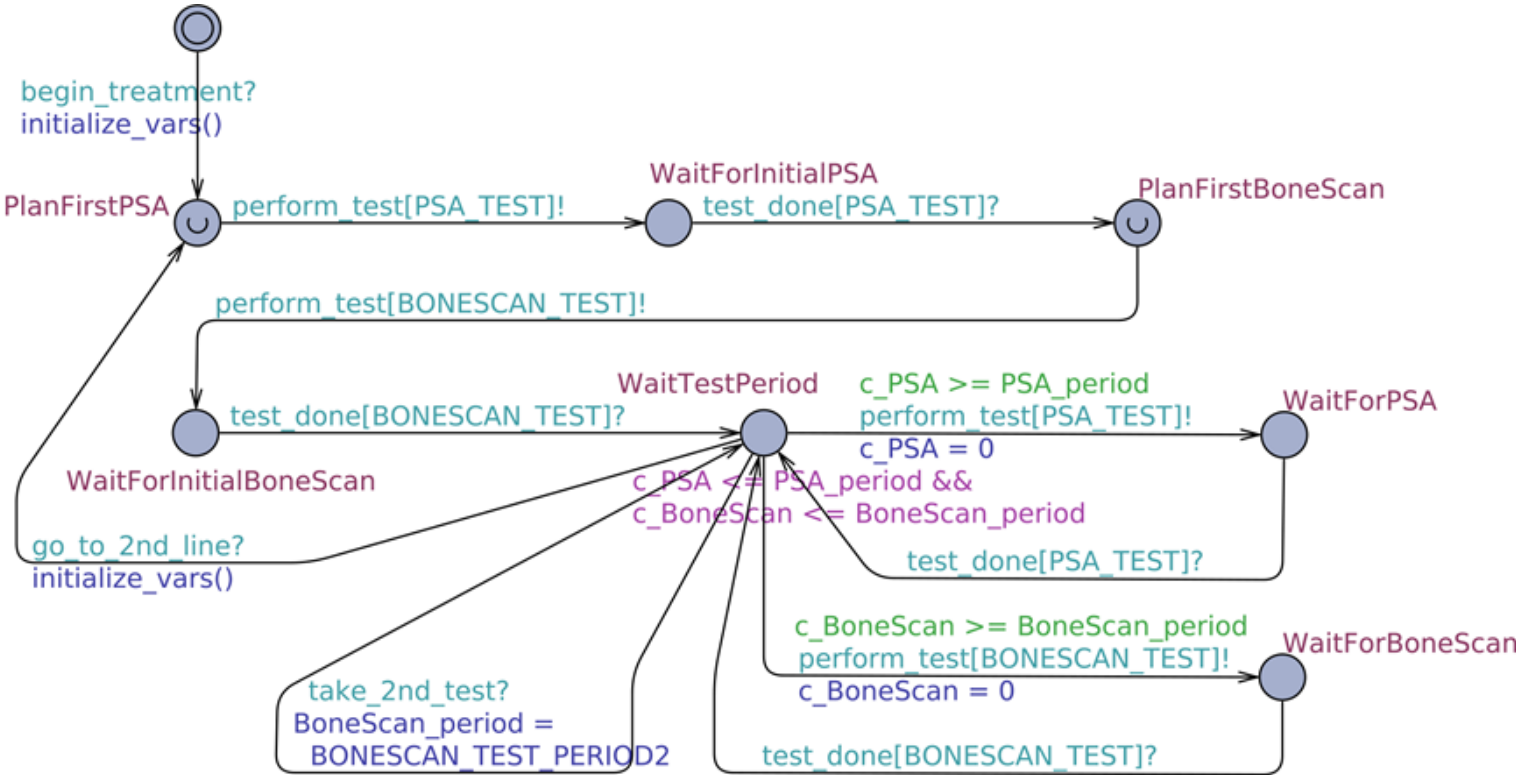


Figure S4a. *Graphical representation of the Test Guideline automaton for the control strategy.*

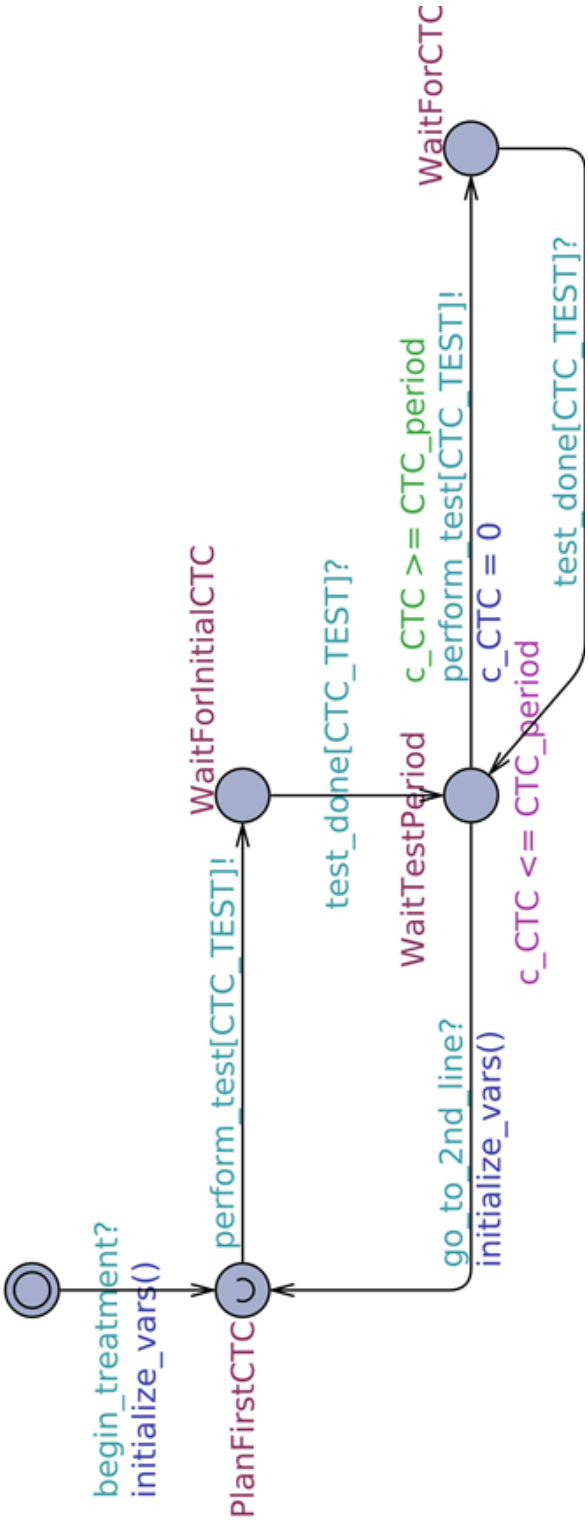


Figure S4b. *Graphical representation of the Test Guideline automaton for the experimental strategy.*

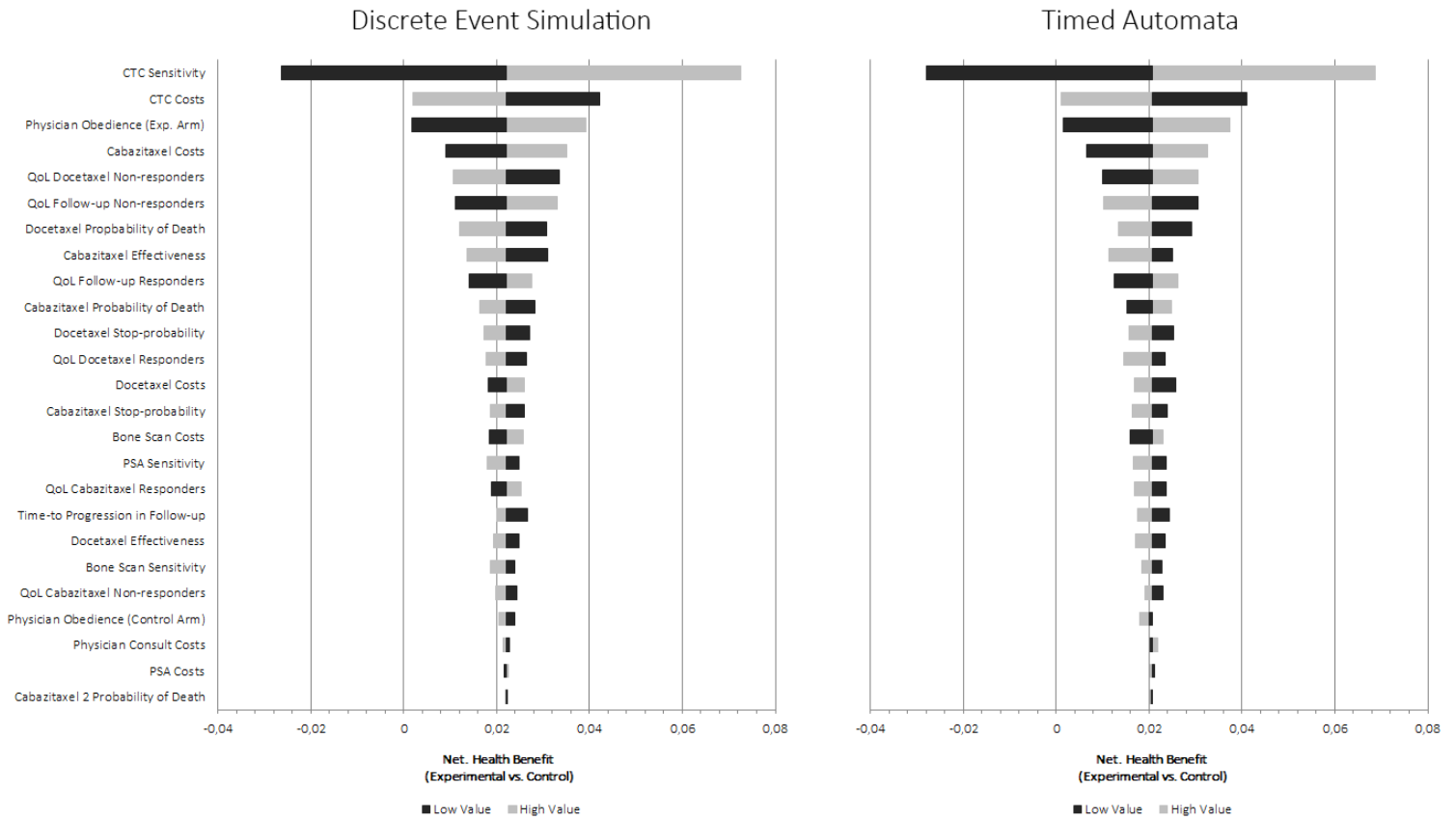


Figure S5. *Tornado plot showing the effect of the uncertainty surrounding the input parameters on the Net. Health Benefit. The parameter values used for this analysis can be found in Table 1.*

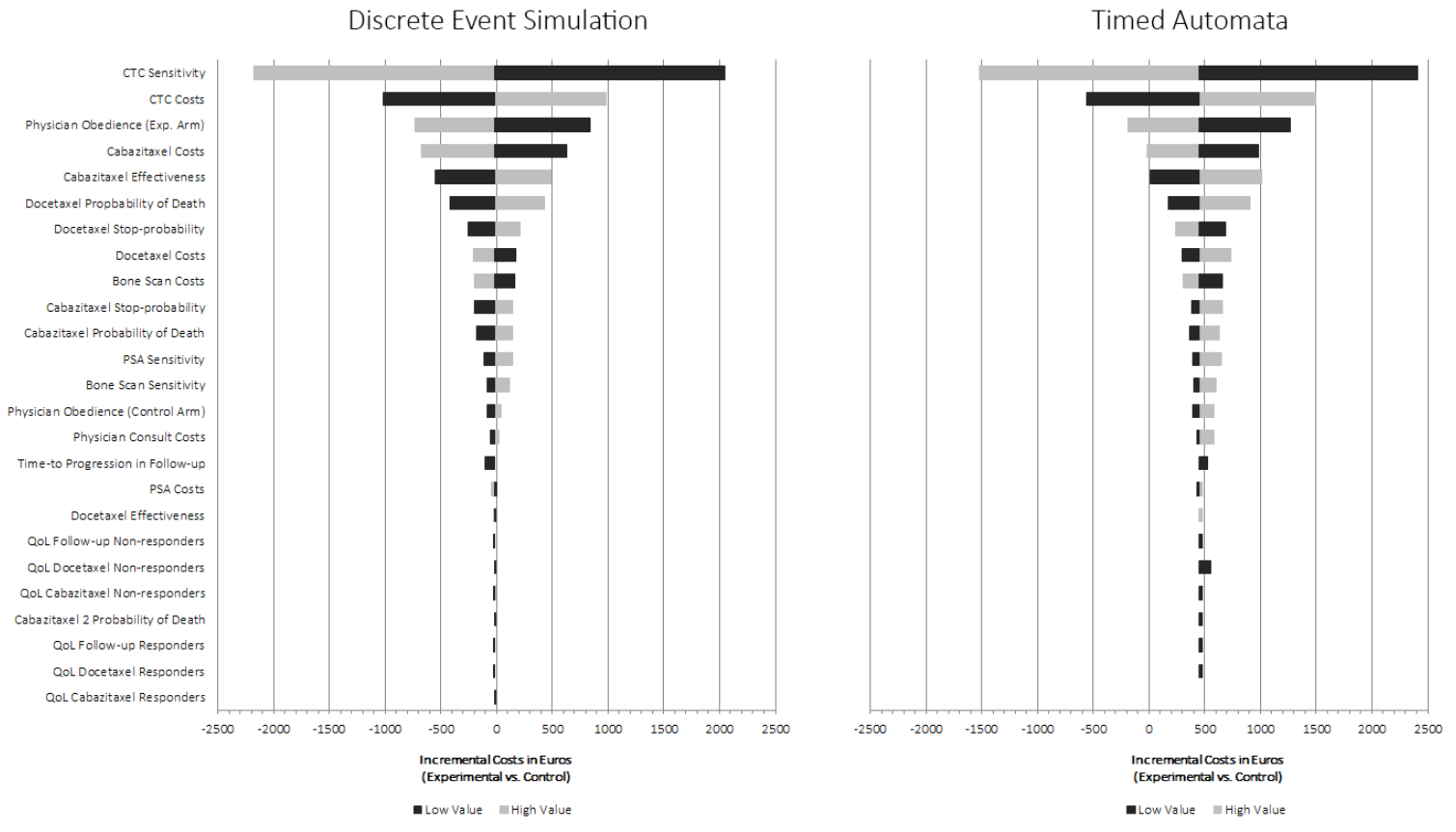


Figure S6. *Tornado plot showing the effect of the uncertainty surrounding the input parameters on the incremental costs. The parameter values used for this analysis can be found in Table 1.*

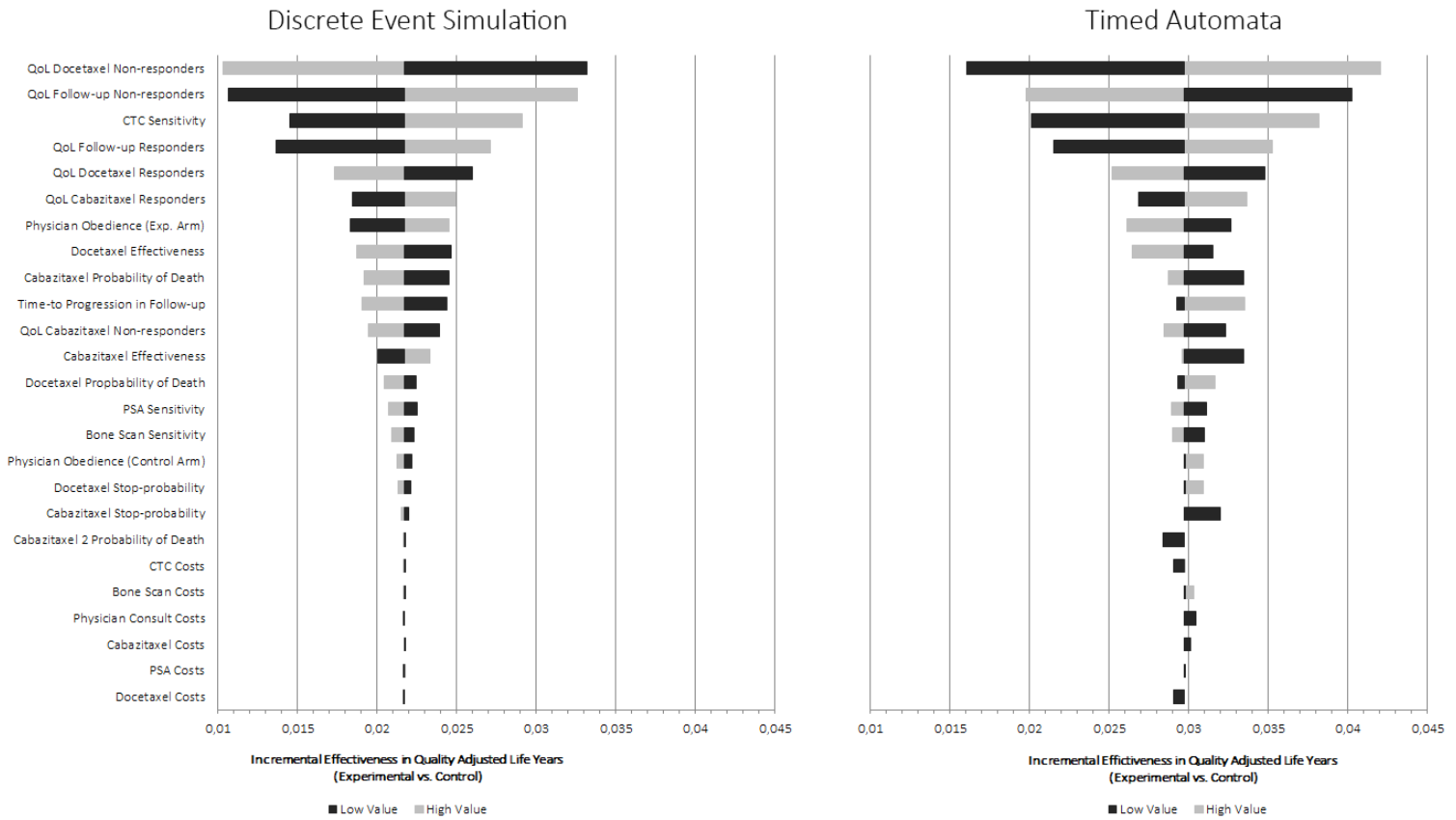


Figure S7. *Tornado plot showing the effect of the uncertainty surrounding the input parameters on the incremental effectiveness. The parameter values used for this analysis can be found in Table 1.*

PART 2

Methodological guidance for developing patient-level simulation models

CHAPTER 5

Accounting for parameter uncertainty in the definition of parametric distributions used to describe individual patient variation in health economic models

Koen Degeling
Maarten J. IJzerman
Miriam Koopman
Hendrik Koffijberg

This chapter has been published as:
Degeling K, IJzerman MJ, Koopman M, Koffijberg H. Accounting for parameter uncertainty in the definition of parametric distributions used to describe individual patient variation in health economic models. BMC medical research methodology. 2017;17(1):170.

ABSTRACT

Background: Parametric distributions based on individual patient data can be used to represent both stochastic and parameter uncertainty. Although general guidance is available on how parameter uncertainty should be accounted for in probabilistic sensitivity analysis, there is no comprehensive guidance on reflecting parameter uncertainty in the (correlated) parameters of distributions used to represent stochastic uncertainty in patient-level models. This study aims to provide this guidance by proposing appropriate methods and illustrating the impact of this uncertainty on modeling outcomes.

Methods: Two approaches, 1) using non-parametric bootstrapping and 2) using multivariate Normal distributions, were applied in a simulation and case study. The approaches were compared based on point-estimates and distributions of time-to-event and health economic outcomes. To assess sample size impact on the uncertainty in these outcomes, sample size was varied in the simulation study and subgroup analyses were performed for the case-study.

Results: Accounting for parameter uncertainty in distributions that reflect stochastic uncertainty substantially increased the uncertainty surrounding health economic outcomes, illustrated by larger confidence ellipses surrounding the cost-effectiveness point-estimates and different cost-effectiveness acceptability curves. Although both approaches performed similar for larger sample sizes (i.e. n = 500), the second approach was more sensitive to extreme values for small sample sizes (i.e. n = 25), yielding infeasible modeling outcomes.

Conclusions: Modelers should be aware that parameter uncertainty in distributions used to describe stochastic uncertainty needs to be reflected in probabilistic sensitivity analysis, as it could substantially impact the total amount of uncertainty surrounding health economic outcomes. If feasible, the bootstrap approach is recommended to account for this uncertainty.

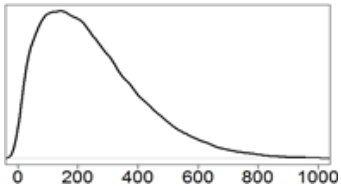
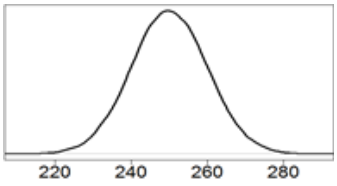
BACKGROUND

Clinical decision-making is aiming towards patient-specific and preference-sensitive treatment, based on multiple biomarkers for treatment targeting and monitoring patients' response to treatment [1]. Consequently, there is an increasing need for corresponding patient-level models to accurately represent clinical practice when estimating the health economic impact of novel healthcare interventions [2, 3]. To facilitate decision making, such models should adequately reflect all types of uncertainty in the synthesized evidence used for analysis [4]. This is particularly relevant in patient-level modeling studies in which reflecting patient heterogeneity may effectively increase uncertainty, for example by relatively low sample sizes in defined subgroups or by an increasing number of parameters that need to be estimated to account for patient characteristics in individualized predictions.

Uncertainty in evidence can be disaggregated into stochastic uncertainty (i.e. patient-level variation or first-order uncertainty) and parameter uncertainty (i.e. second-order uncertainty) [4]. This can be illustrated using a time-to-event parameter, e.g. the time-to-progression after surgery. If, for a certain patient group, this parameter is defined by a mean estimate (Table 1 – Box A), the parameter uncertainty in this mean estimate needs to be accounted for, which can be done using a parametric distribution in the probabilistic sensitivity analysis (PSA) (Table 1 – Box B) [4]. For example, this parameter uncertainty could be reflected by defining a Normal distribution for the mean time-to-progression based on the estimated mean and standard error, by applying the Central Limit Theorem [5]. Additionally, it is possible to account for stochastic uncertainty by using parametric distributions to describe individual patient variation, such as typically used in patient-level state-transition models [6] and discrete event simulation (DES) models [7]. For example, a Weibull distribution [8] can be used in a DES model to derive and assign patient-specific time-to-progression values (Table 1 – Box C), rather than assigning the estimated mean time-to-progression to all patients (Table 1 – Box A). As for the parameter defined by the mean estimate in Box A of Table 1, the uncertainty in the parameters defining this Weibull distributions needs to be accounted for [4]. However, there is currently no clear guidance on how to reflect parameter uncertainty (i.e. second- order uncertainty) in these parametric distributions used to describe stochastic uncertainty (i.e. patient-level variation or fist-order uncertainty) in PSA (Table 1 – Box D).

Reflecting parameter uncertainty in parametric distributions is challenging because these distributions are typically defined by multiple parameters, e.g. the shape and scale parameter of the Weibull distribution, with their values preferably estimated from individual patient data (IPD). When estimated from IPD, these distributions' parameters are correlated for most of the parametric distributions commonly used in health economic analyses, such as Gamma, Beta, Log-Normal, and Weibull distributions [8]. Consequently, it is incorrect to estimate the values of, for example, the shape and scale parameters independently, and to define independent (i.e. separate) distributions for each of these parameters' uncertainty in PSA. Therefore, guidance is needed towards approaches that maintain the correlation in the parameters of distributions reflecting stochastic uncertainty when accounting for the parameter uncertainty these parameters introduce.

Table 1. *Illustration of the example that is discussed in the introduction. SEE, Standard Error of the Estimate.*

Type of Uncertainty	Time-to-Event *(Stochastic uncertainty not reflected)*	Time-to-Event *(Stochastic uncertainty reflected)*
Stochastic uncertainty *(Patient-level variation)* *(First-order uncertainty)*	**Box A** Patient-level variation is not described. Consequently, an estimated mean time-to-progression is assigned to all patients: • Mean est. = 250 days 	**Box C** Patient-level variation in time-to-progression is described by a Weibull distribution: • Shape parameter est. = 1.5 • Scale parameter est. = 277 • μ ≈ 250 days 
Parameter uncertainty *(Second-order uncertainty)*	**Box B** Parameter uncertainty in the estimated mean time-to-progression is described by a Normal distribution: • Mean parameter est. = 250 • SEE parameter est. = 10 and included in the PSA. 	**Box D** Parameter uncertainty in the estimated Shape and Scale parameter of the Weibull distribution needs to be described and included in the PSA. Patient-level variation in time-to-progression is reflected by drawing from a different distribution in each PSA sample: PSA | Shape | Scale | Mean 1 | 1.50 | 242 | 219 2 | 1.62 | 253 | 227 3 | 1.67 | 292 | 260 n | … | … | … *Methods to generate these parameter values are illustrated in this paper.*

The lack of specific guidance is illustrated by recently published modeling studies that do not fully utilize IPD while performing PSA, for example, by adjusting predicted values (stochastic uncertainty) with a random percentage [9], by excluding correlation between parametric distributions' parameters [10], or altogether not reporting (including) parameter uncertainty in parametric distributions [11]. As a consequence of inadequately reflecting all types of uncertainty in the synthesized evidence used for analysis, suboptimal resource allocation and

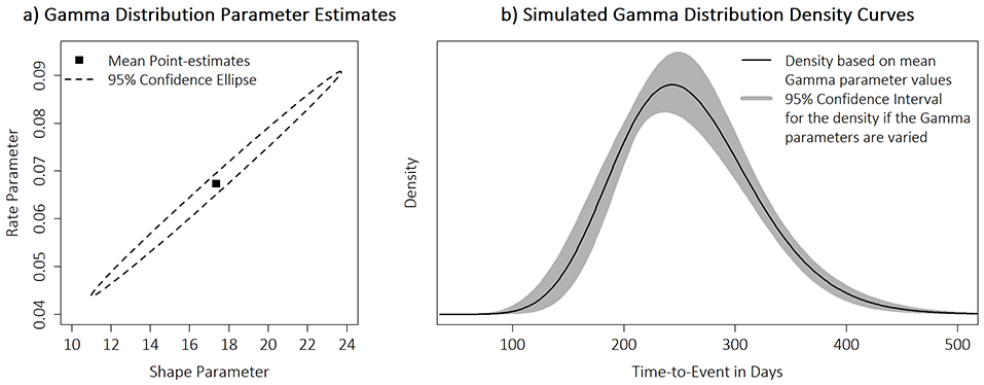


Figure 1. *a) Exemplary plot of a Gamma distribution's shape and scale parameters, showing their point-estimates and the 95% confidence ellipse representing the uncertainty and correlation in these estimates. b) Density plot of the Gamma distribution as defined in Figure 1a. The black line represents the density curve from which individual values are drawn in each run of the PSA if parameter uncertainty in the parameter estimates is ignored. The grey surface represents the 95% confidence interval of density curves from which individual values are drawn if different (correlated) sets of parameter values are used in each run of the PSA, i.e. if parameter uncertainty in the parameter estimates is accounted for.*

research prioritization decisions may be made due to biased outcomes of PSA, overestimated confidence in the corresponding expected values of the PSA, and ensuing biased estimation of the value of collecting additional evidence to better inform decision making [12].

The objective of this study is to provide explicit guidance for health economic modelers on how parameter uncertainty in parametric distributions used to describe stochastic uncertainty (i.e. patient-level variation) can be considered in PSA. In order to do so, two alternative solutions are illustrated and compared: 1) a non-parametric bootstrapping approach and 2) an approach using multivariate Normal distributions. Both approaches are used in a DES simulation study and in a DES case study in which stochastic uncertainty in time-to-event data is described by parametric distributions. Additionally, the potential impact of increased parameter uncertainty due to subgroup stratification on health economic outcomes is illustrated.

METHODS

Consider the scenario in which time-to-event observations from a clinical study are used to describe stochastic uncertainty by fitting a Gamma distribution [8], i.e. the distributions' (correlated) shape and rate parameter values are estimated from the IPD. Figure 1a shows an example of the estimated values of these parameters, including the 95% confidence ellipse, representing the sets (i.e. combinations) of parameter values. If uncertainty in the parameter estimates (i.e. parameter uncertainty) is ignored by using the parameters' point-estimates in each run of the PSA, individual time-to-event values would be repeatedly drawn based on the same mean density curve of Figure 1b in all runs of the PSA. However, if this parameter uncertainty is reflected by using different sets of correlated parameter estimates for each run of the PSA, a variety of distributions is simulated, which is illustrated by the 95%

confidence interval (surface) in Figure 1b. As this variety of distributions reflects uncertainty in the individual time-to-event values as well as overall (mean) time-to-event values, this uncertainty can be seen as parameter uncertainty (in this case of the 'hyperparameters, i.e. the parameters defining the time-to-event' distribution). Both proposed approaches, which are introduced subsequently, can be used to generate these correlated sets of distributions' parameter values based on IPD. A different set of generated parameters values can be used in the PSA, incorporating one set of correlated values in each Monte Carlo sample.

Non-parametric bootstrapping (Bootstrap approach)

Non-parametric bootstrapping is a statistical technique that can be used to construct an approximate sampling distribution of a statistic of interest, without the need for assumptions regarding the distribution of this statistic [13]. It has been applied in health economics, for example, to construct confidence intervals for the incremental net benefit in economic evaluations alongside clinical trials [5, 14]. In the Bootstrap approach, the distributions' parameters are repeatedly estimated based on different bootstrap samples of the original dataset, which are obtained by resampling the original dataset with replacement, such that the size of the bootstrap sample equals the size of the original dataset [13]. A detailed discussion of the Bootstrap approach is provided in Appendix 1. Briefly, reflecting parameter uncertainty with this approach consists of the following four steps:

1. Generate a feasible* bootstrap sample of the original dataset, by resampling this dataset with replacement, such that the sample size of the bootstrap sample equals that of the original dataset.
2. Fit the pre-specified distribution(s)** to the bootstrap sample and record the estimated parameter values.
3. Repeat Step 1 and Step 2 r times, where r equals the required number of PSA runs.
4. Perform the PSA, using a different set of parameter values to define the distribution(s) for each PSA run.

** The definition of feasible bootstrap samples may vary between studies. Please see Appendix 1 for a more in-depth discussion.*

*** Note that if multiple distributions are fitted in Step 2, all distributions need to be fitted on the same bootstrap sample to preserve correlation among all distributions and other parameters used to describe variables in the dataset.*

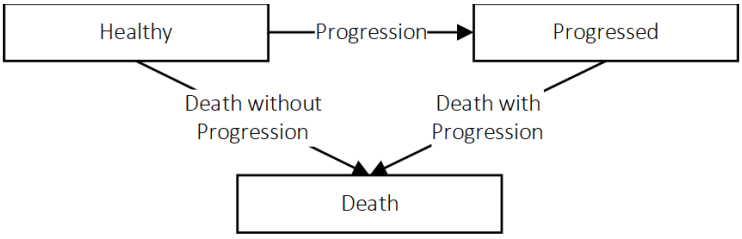


Figure 2. *Graphical representation of the DES model that was used for the simulation study as described in the Methods section.*

Multivariate normal distributions (MVNorm approach)

The MVNorm approach assumes the distributions' parameters to be Normal distributed, which is valid for sufficiently large sample sizes according to the Central Limit Theorem [5], as suggested by Briggs et al. for regression models in general [15]. Given a distribution's parameter estimates and their variance-covariance matrix, a multivariate Normal distribution can be defined and used to draw correlated sets of parameter values. A detailed discussion of the MVNorm approach is provided in Appendix 2. Briefly, reflecting parameter uncertainty with this approach consists of the following four steps:

1. Fit the pre-specified distribution to the original dataset and record the estimated parameter values and (calculate) the variance-covariance matrix.
2. Define a multivariate Normal distribution from the parameters' estimates and their variance-covariance matrix according to Step 1.
3. Draw r feasible* sets of parameter values from the defined distribution (Step 2), where r equals the required number of PSA runs.
4. Perform the PSA, using a different set of parameter values to define the distribution(s) for each PSA run.

** The drawn sets of parameter values need to be assessed for their feasibility, i.e. whether the parameter values are appropriate for the prespecified distributions. Please see Appendix 1 for a more in-depth discussion.*

Simulation study

A simulation study was performed to assess potential differences in the performance of both approaches and compare them to the scenario in which parameter uncertainty in the time-to-event distributions was not considered. This simulation study was performed in R Statistical Software version 3.3.2 [16] and used a basic health economic DES model to compare two treatment strategies in terms of health economic outcomes. This health economic DES model was structured according to a basic three state disease progression model, i.e. healthy, progressed, and death, and included two competing risks for patients in the healthy state, i.e. progression and death (Figure 2). Time-to-event data was simulated using Weibull distributions, separately for the intervention and control patient populations, which differed in terms of survival and treatment costs in the progressed state. The exact definitions of the time-to-event distributions and cost and effect parameters are provided in Appendix 3.

The simulation study was carried out for several sample sizes, i.e. 500, 100, 50, and 25 patients, performing 2500 simulation runs (i.e. hypothetical trials) of 5000 PSA runs, each including 5000 patients per treatment strategy, for each sample size (Figure 3). Weibull distributions were used to describe stochastic uncertainty in time-to-event data to avoid potential bias due to mismatching distributions, as Weibull distributions were also used to simulate the hypothetical populations. Distributions were fitted using the *fitdist* function of the *fitdistrplus* package [17] in R Statistical Software [16]. Stochastic uncertainty in the model's cost and effect parameters was deliberately not considered to represent the common scenario in which IPD is not available for all model parameters, and PSA samples for these parameters were generated according to the distributions as defined in Appendix 3. Random draws from

a multivariate Normal distribution were performed using the *mvrnorm* function of the *MASS* package [18].

Estimations of the "true" mean value and distribution of the parameters and health economic outcomes were obtained by analyzing the model for 2500 different samples from the simulated populations, with corresponding sample sizes. This resembles the scenario in which the population values, i.e. values based on observing the total population, would be approximated in practice by performing 2500 clinical studies. These "true" results were used to compare the generated sets of distributions' parameter values and the health and economic outcomes using either of both approaches. The distributions of the time-to-event distributions' parameter values and health economic outcomes were compared based on their relative entropy, i.e. the Kullback-Leibler divergence [19], using the *KLdiv* function of the *flexmix* package [20–22]. The relative entropy is a measure of the difference between two probability distributions, for which lower values indicate a better match of distributions. Additionally, the impact of considering the parameter uncertainty in the time-to-event distributions' parameters was illustrated in an incremental cost-effectiveness plane for one hypothetical trial, i.e. one random run out of 2500 simulation runs, and in mean cost-effectiveness acceptability curves (CEACs), including a 95% confidence interval, based on all simulation runs.

Case study

To illustrate how parameter uncertainty in time-to-event distributions' parameter estimates could impact health economic outcomes in practice, a case study was performed based on the randomized phase 3 CAIRO3 study (NCT00442637) that was carried out by the Dutch Colorectal Cancer Group [23]. A total of 558 metastatic colorectal patients with stable disease or better after six cycles of capecitabine, oxaliplatin, and bevacizumab (CAPOX-B) induc-

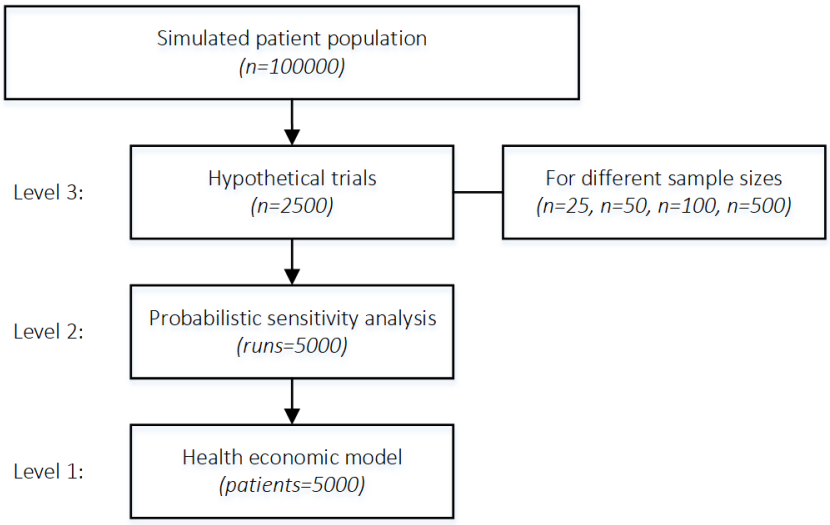


Figure 3. *Graphical representation of the simulation study as described in the Methods section.*

tion therapy were randomized to either receive capecitabine and bevacizumab maintenance treatment or observation until progression of disease. CAPOX-B treatment was to be re-introduced upon progression on either maintenance or observation, and continued until second progression (PFS2) the primary endpoint of the study.

A previous developed DES model was adapted to use the sets of distributions' parameter values generated by the Bootstrap and MVNorm approach in the PSA (Appendix 4). The model was developed and validated in AnyLogic multimethod simulation software [24] according to good research practices guidelines [7, 25, 26], and structured according to the same health states as the state-transition model used for the original evaluation of the CAIRO3 study: post-induction, re-induction, salvage, and death [27] (Figure 4). Event-specific probabilities and Weibull distributions were used to describe time-to-event data and handle the competing risks of disease progression and death in the post-induction and re-induction state [28]. Parameters of the distributions used to reflect parameter uncertainty in non-time-to-event parameters, e.g. costs and utilities, were deliberately defined exactly as in the original evaluation of the CAIRO3 study for all subgroups [27], so that observed differences in outcomes could be designated specifically to the uncertainty in time-to-event distributions' parameter estimates. The health economic outcomes for all case study analyses were presented in incremental cost-effectiveness planes and CEACs based on 10,000 PSA runs of 10,000 patients per treatment strategy.

Clinical relevant subgroups were defined to resemble the personalized context described in the introduction, illustrating the impact of parameter uncertainty in time-to-event distributions' parameter estimates on health economic outcomes for different sample sizes. Patients were stratified according to their treatment response (stable disease (SD) versus complete or partial response (CR/PR)) and stage of disease (synchronous versus metachronous), which resulted in a total of 8 subgroups with sample sizes ranging from 50 to 410 (Table 2). As illustrated in Table 2, this stratification created subgroups in which events were observed only once, or not all, which prohibits fitting of a Weibull distribution, and the probability of that event occurring was therefore set to zero.

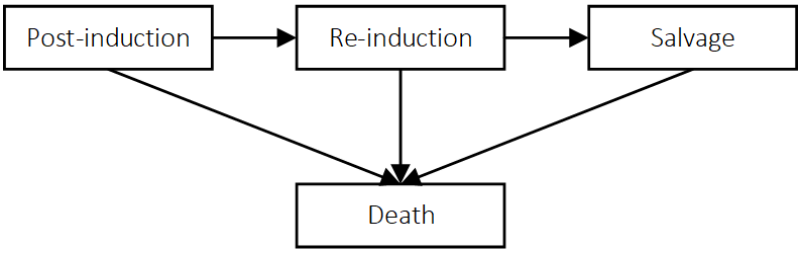


Figure 4. *Graphical representation of the DES model that was used for the case study as described in the Methods section.*

Table 2. *Overall and treatment strategy-specific subgroup definitions, sample sizes, and number of events for the case study. SD, stable disease; CR/PR, complete or partial response.*

		Event Observations				
Subgroup	**Sample Size**	**Progression *(Post-induction)***	**Death *(Post-induction)***	**Progression *(Re-induction)***	**Death *(Re-induction)***	**Death *(salvage)***
Subgroup 0: No Subgroups	558	519	39	370	109	410
Observation	279	270	9	200	54	216
Maintenance	279	249	30	170	55	194
Subgroup 1: SD	191	185	6	128	43	142
Observation	95	95	0	68	22	73
Maintenance	96	90	6	60	21	69
Subgroup 2: CR/PR	367	334	33	242	66	268
Observation	184	175	9	132	32	143
Maintenance	183	159	24	110	34	125
Subgroup 3: Synchronous	410	382	28	275	82	300
Observation	191	186	5	143	33	153
Maintenance	219	196	23	132	49	147
Subgroup 4: Metachronous	147	137	10	95	27	110
Observation	88	84	4	57	21	63
Maintenance	59	53	6	38	6	47
Subgroup 5: SD & Synchronous	141	136	5	94	34	102
Observation	67	67	0	49	15	52
Maintenance	74	69	5	45	19	50
Subgroup 6: SD & Metachronous	50	49	1	34	9	40
Observation	28	28	0	19	7	21
Maintenance	22	21	1	15	2	19
Subgroup 7: CR/PR & Synchronous	269	246	23	181	48	198
Observation	124	119	5	94	18	101
Maintenance	145	127	18	87	30	97
Subgroup 8: CR/PR & Metachronous	97	88	9	61	18	70
Observation	60	56	4	38	14	42
Maintenance	37	32	5	23	4	28

RESULTS

Simulation study

The potential impact of considering parameter uncertainty in time-to-event distributions' parameter estimates in PSA is illustrated for several sample sizes in Figure 5, which shows results of one single run of the simulation study. Ignoring parameter uncertainty in the time-to-event distributions' parameter estimates leads to an underestimation of the uncertainty surrounding cost-effectiveness outcomes. This is illustrated by the smaller confidence ellipse for this scenario (long-dashed black ellipse) compared to other scenarios in which uncertainty in the time-to-event distributions' parameter estimates is accounted for using one of the proposed approaches (dashed gray and dotted light-gray ellipses). The indicated effect is already observed for rather large sample sizes (i.e. n = 500) and increases as sample size decreases, illustrated by the larger distance between the long-dashed black confidence ellipse and the dashed gray and dotted light-gray confidence ellipses for smaller sample sizes. Although the Bootstrap (dashed gray ellipse) and MVNorm approach (dotted light-gray ellipse) yield incremental cost-effectiveness point-estimates similar to the "real" value, both approaches slightly overestimate the magnitude of the uncertainty for sample sizes of n = 100 and smaller, demonstrated by smaller confidence ellipses for the "real" uncertainty (solid black ellipse). For very small sample sizes (i.e. n = 25), the MVNorm approach generates unrealistic parameter values, e.g. indicating a mean survival far beyond life-expectancy, leading to extreme and unrealistic health economic outcomes, which results in an unrealistic large confidence ellipse (not presented). However, since Figure 5 represents only one run of the simulation study, these results may not be representative for the approaches in general.

To assess the performance of the approaches in general, 2500 of these comparisons were performed in the simulation study. Results show that both approaches yield comparable mean parameter estimates and standard errors (Appendix 5). Although the MVNorm approach seems to perform slightly better for small sample sizes (n = 25) in terms of mean parameter estimates, this approach too often yields extreme and unrealistic outcomes in the health economic simulation. Considering the Kullback-Leibler divergence, both approaches perform similar, though on average the MVNorm approach seems to represent the "estimated true" distributions slightly better for very small sample sizes (i.e. n = 25) (Appendix 6). Also the CEACs presented in Figure 6 show that the results for both approaches are similar (light gray and gray lines). However, compared to the results of the strategy in which the uncertainty in time-to-event distributions' parameter estimates is not considered (black line), both approaches yield different mean CEACs (solid lines) with different confidence intervals (dashed lines), illustrating the potential health economic impact of ignoring this uncertainty.

Case study

Incremental cost-effectiveness planes for the cohort analysis and selected subgroup analyses of the case study are presented in Figure 7. CEACs and incremental cost-effectiveness planes for the cohort analysis and all subgroups analyses are available in Appendices 7 and 8. The cost-effectiveness point-estimate for the cohort (Subgroup 0) is not affected by considering the parameter uncertainty in the time-to-event distributions' parameter estimates in PSA, as the points in the corresponding incremental cost-effectiveness plane overlap. However, there

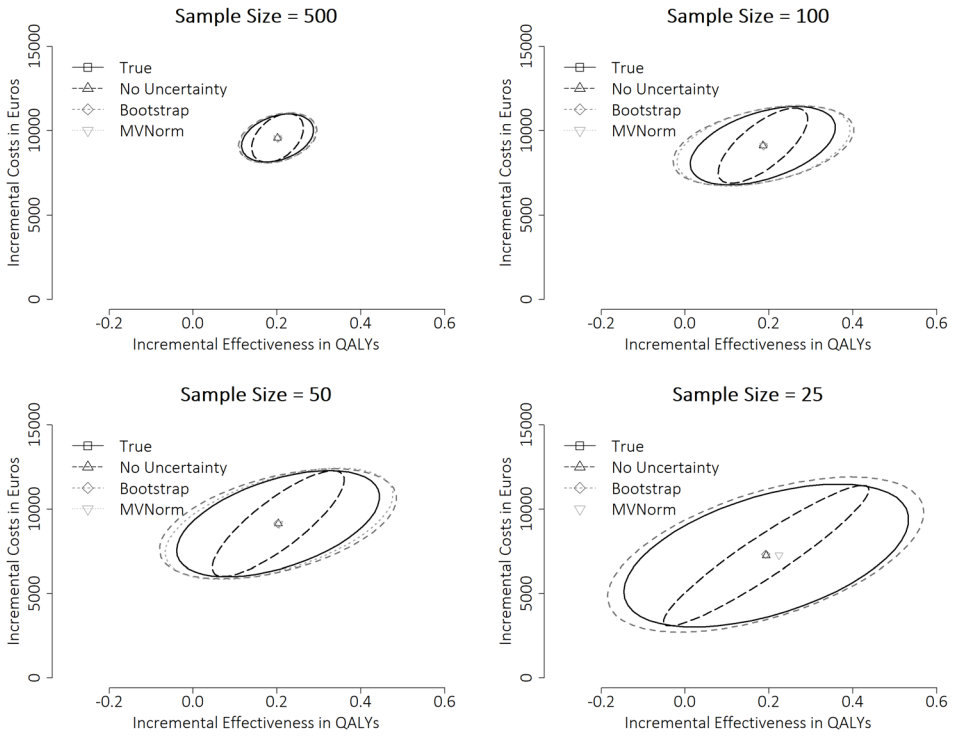


Figure 5. *Incremental cost-effectiveness planes for one run of the simulation study showing the mean point-estimates and corresponding 95% confidence ellipses for the different sample sizes. The confidence ellipse for the MVNorm approach with sample size $n = 25$ was not plotted, because it was unrealistic large.*

is a substantial increase in the amount of uncertainty surrounding this point-estimate when the uncertainty in the distributions' parameter estimates is accounted for, illustrated by the distance between the long-dashed black confidence ellipse and the dashed gray and dotted light- gray confidence ellipses. The potential impact of this increase in uncertainty is illustrated by the results for Subgroup 3. Without considering the parameter uncertainty in the distributions' parameter estimates, the results indicate that health loss due to maintenance treatment (experimental strategy) is unlikely in this subgroup, as the corresponding (dashed black) confidence ellipse is entirely located right of the vertical axis. However, the dashed gray and dotted light-gray confidence ellipses, including this uncertainty, show there actually is a non-zero probability of health loss, as these are partly located left of the vertical axis.

As the cohort is stratified, and sample size decreases, the impact of considering parameter uncertainty in the time-to-event distributions' parameter estimates in PSA increases substantially, which is illustrated by the results for Subgroup 4 and Subgroup 6. These results show that the increase in uncertainty due to stratification may become so large that (further) subgroup stratification might not be informative. Additionally, the results for Subgroup 6 illustrate that for small sample sizes (i.e. $n = 50$) the point-estimates of the cost-effectiveness

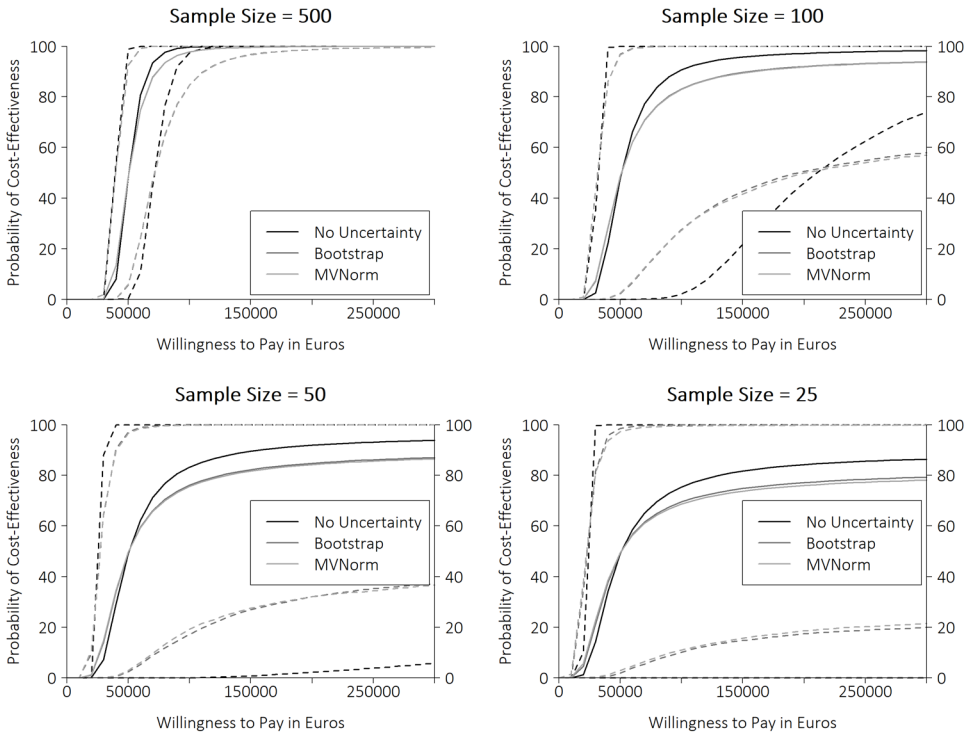


Figure 6. *CEACs showing the mean probability of cost-effectiveness (solid lines) and corresponding 95% confidence interval (dashed lines) for all runs of the simulation study.*

outcomes themselves may also be affected by the uncertainty in the distributions' parameters, which is illustrated by the non-overlapping points in the corresponding incremental cost-effectiveness plot.

No meaningful differences in the point estimates and the magnitude of uncertainty surrounding these estimates between the Bootstrap and MVNorm approach are observed, as their point-estimates and confidence ellipse overlap to a great extent. However, as also observed in the simulation study, the use of the MVNorm approach occasionally results in extreme and unlikely parameter values, which were to be excluded from the simulation.

DISCUSSION

As demonstrated in this paper, parameter uncertainty in parametric distributions used to describe stochastic uncertainty (i.e. patient-level variation) should be explicitly accounted for in PSA by modelers, as it does impact incremental cost-effectiveness point-estimates and CEACs. If this type of uncertainty is ignored, suboptimal resource allocations or research prioritization decisions may be made, due to an underestimation of the total uncertainty sur-

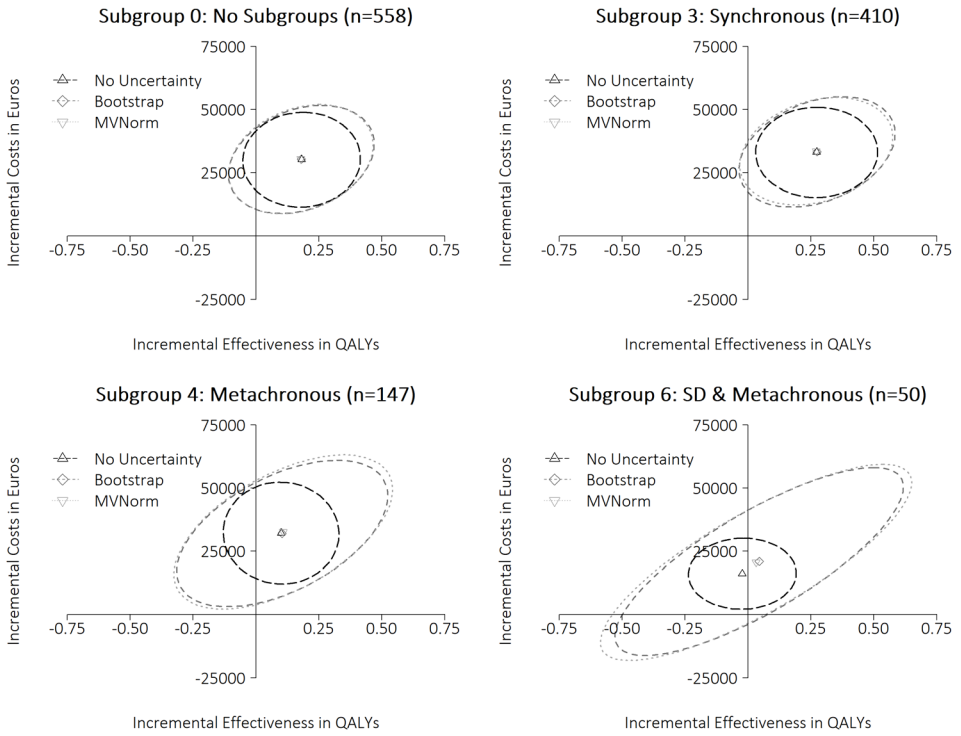


Figure 7. *Incremental cost-effectiveness planes for the cohort (Subgroup 0) and Subgroups 3, 4, and 6 of the case study, showing the mean point-estimates and corresponding 95% confidence ellipses.*

rounding health economic outcomes. This is particularly relevant in a personalized treatment context in which patient stratification obviously leads to an increase in uncertainty on the level of subgroups compared to the level of the full patient group due to decreasing sample sizes in subgroups. This increase in uncertainty should of course be reflected in PSA, and although it is likely that many experienced modelers already do so, clear guidance on appropriate methods was not yet available. Besides time-to-event distributions, this notion also applies to other types of distributions that are used to reflect stochastic uncertainty, e.g. Gamma distributions to describe patient-level costs. Although DES was used in both the simulation and case study, these findings apply to any patient-level modeling method used to reflect stochastic uncertainty, e.g. microsimulation state-transition models. Reflecting all parametric uncertainty in this way does require drawing values from a different distribution in each PSA sample, providing some additional work for modelers. However, the required effort is minimal and an online tool, including tutorials, has been made available alongside this paper to easily analyze individual patient data for implementation into patient-level models [29].

When accounting for parameter uncertainty in distributions' parameter estimates, the Bootstrap approach has some advantages over the MVNorm approach. The Bootstrap approach seems more robust for smaller sample sizes, does not require any assumptions regarding underlying distributions, and preserves the correlation in the whole dataset. Although not

distinctively illustrated in this study, the latter also concerns the correlation with other, non-time-to-event, related parameters, such as utilities and costs, and can be considered a major advantage. The Bootstrap approach does, however, require the definition of a feasible bootstrap sample. Especially in case of scarce events, which are inevitable in personalized medicine, this can be challenging. Moreover, the sets of parameters values probably need to be generated outside the software environment used for the simulation and need to be imported into this environment for performing the PSA.

Although the MVNorm approach is better capable of handling scarce events in theory, mainly because it does not require the definition of a feasible bootstrap sample, it has several severe downsides. The MVNorm approach requires the definition of feasible values for the distributions' parameter estimates and the assumption that these estimates follow a Normal distribution, which is not appropriate for insufficiently large sample sizes. For example, the shape and rate parameter of the Gamma distribution may be skewed, depending on the IPD on which their values are estimated. Additionally, many of the commonly used parametric distributions are defined for positive parameter values only, whereas Normal distributions are defined for any real number, including negatives. Furthermore, although this approach is better capable of handling scarce events in theory, it is likely to yield extreme parameter values for smaller sample sizes due to increasing standard errors, a scenario in which the use of the MVNorm approach is advised against.

Given the advantages and disadvantages of both approaches, the Bootstrap approach is recommended over the MVNorm approach if constructing enough feasible bootstrap samples is possible. However, if modelers feel the need to use the MVNorm approach, for example because multivariate Normal distributions are supported in the used software environment and the Bootstrap approach is not, they should carefully check whether 1) the sample size is sufficient, i.e. no implausible and extreme parameter values are observed, and 2) whether correlation between defined distributions to reflect patient-level variation is low or negligible and the additional uncertainty introduced by using independent multivariate Normal distributions (e.g. for costs, utilities, time-to-events) is therefore limited.

Several choices in the study design may have influenced the final results. Both the simulation study and case study focus on the impact of parameter uncertainty specifically in time-to-event distributions, because primary outcomes in clinical studies are often related to time-to-events, e.g. overall survival, and these distributions characterize DES. Additionally, accounting for the parameter uncertainty in non-time-to-event related distributions may further increase the amount of uncertainty surrounding the health economic outcomes, which will contribute to the conclusion that parameter uncertainty in distributions used to describe stochastic uncertainty (i.e. patient-level variation) needs to be accounted for. Furthermore, Weibull distributions are used to describe the IPD, which is a design choice and therefore introduces structural uncertainty. This choice is not expected to meaningfully influence the outcomes, because the fitted distributions match the data well. Moreover, other distributions' parameters may have a stronger correlation, which would further stress the need to use one of the proposed approaches to appropriately account for the uncertainty in these parameters.

Further research may be directed towards evaluating the impact of using different types of distributions to describe stochastic uncertainty on health economic outcomes, the uncertainty surrounding these outcomes, and the performance of both approaches. Additionally, sensitivity analyses other than PSA, such as deterministic sensitivity analysis and structural sensitivity analysis, might be considered. Furthermore, additional guidance is desirable on how uncertainty in IPD can be appropriately combined with uncertainty in aggregated data on population level, e.g. a reported mean estimate and standard error from literature, for sensitivity analyses in patient-level models.

CONCLUSIONS

With an increasing need for patient-level models to accurately represent clinical practice, modelers should be aware that the parameter uncertainty in parametric distributions used to describe stochastic uncertainty (i.e. patient-level variation) should be accounted for in PSA performed in health economic modeling studies. This type of uncertainty could have a substantial impact on the total amount of uncertainty surrounding the health economic outcomes and may influence healthcare decision-making. At least two approaches are available to account for the parameter uncertainty in parametric distributions used to describe stochastic uncertainty. If feasible, the Bootstrap approach is recommended to account for this type of uncertainty.

AVAILABILITY OF DATA AND MATERIALS

An online tool to easily analyze individual patient data for implementation into patient-level models is available at: https://personex.nl/tools. The data of the CAIRO3 trial is not publicly available.

ACKNOWLEDGEMENTS

The authors acknowledge prof. dr. C.J.A. Punt from the Academic Medical Center of the University of Amsterdam, who was the principal investigator of the CAIRO3 trial together with prof. dr. M. Koopman, for providing the dataset of the trial as a case study.

CONTRIBUTIONS STATEMENT

KD contributed to the research design, performed the simulation study and case study, and drafted the manuscript. MIJ contributed to the research design, supervised the statistical methodology, and provided critical revisions and discussion. MK contributed to the research design of the case study, supervised the case study, and provided critical revisions and discussion. HK contributed to the research design, supervised the simulation study and statistical methodology, and drafted the manuscript. All authors read and approved the final manuscript.

FUNDING SOURCES AND CONFLICTS OF INTEREST

No funding was received for performing this research. The authors declare that they have no competing interests.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The CAIRO3 study protocol, which includes secondary data analyses such as performed in this study, was approved by the Committee on Human-Related Research Arnhem-Nijmegen in the Netherlands. Written informed consent was not required for this study separately, as written informed consent was obtained from all participants in the CAIRO3 study, which includes secondary data analyses such as performed in this study.

REFERENCES

1. Barbieri CE, et al. The emergence of precision urologic oncology: a collaborative review on biomarker-driven therapeutics. Eur Urol. 2017;71(2):237–46.

2. Annemans L, et al. Current methodological issues in the economic assessment of personalized medicine. Value Health. 2013;16(S6):S20–6.

3. Degeling K, et al. A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models. Expert Rev Pharmacoecon Outcomes Res. 2017;17(1):17–25.

4. Briggs AH, et al. Model parameter estimation and uncertainty analysis: a report of the ISPORSMDM modeling good research practices task force working group–6. Med Decis Mak. 2012;32(5):722–32.

5. Nixon RM, et al. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. Health Econ. 2010;19(3):316–33.

6. Siebert U, et al. State-transition modeling: a report of the ISPOR-SMDM modeling good research practices task Force-3. Value Health. 2012;15(6):812–20.

7. Karnon J, et al. Modeling using discrete event simulation: a report of the ISPOR-SMDM modeling good research practices task Force-4. Value Health. 2012;15(6):821–7.

8. Law AM. Simulation Modeling and Analysis. 4th ed. Singapore: McGraw-Hill Higher Education; 2007.

9. Chen T, et al. Potential health impact and cost-effectiveness of drug therapy for prehypertension. Int J Cardiol. 2017;240:403–8.

10. Montgomery SM, et al. A discrete event simulation to model the cost-utility of fingolimod and natalizumab in rapidly evolving severe relapsing-remitting multiple sclerosis in the UK. J Med Econ. 2017;20:474–82.

11. Parikh RC, et al. Cost-effectiveness of treatment sequences of chemotherapies and targeted biologics for elderly metastatic colorectal cancer patients. J Manag Care Spec Pharm. 2017;23:64–73.

12. Claxton K. Exploring uncertainty in cost-effectiveness analysis. PharmacoEconomics. 2008;26(9):781–98.

13. Davison AC, Hinkley DV. Bootstrap methods and their application. 1st ed. Cambridge: Cambridge University Press; 1997.

14. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med. 2000;19(9):1141–64.

15. Briggs AH, et al. Decision modeling for health economic evaluation. Oxford: Oxford University Press; 2006.

16. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2016. https://www.r-project. org/. [Accessed 22 Feb 2017]

17. Delignette-Muller ML, Dutang C. Fitdistrplus: an R package for fitting distributions. J Stat Softw. 2015;64(4):1–34.

18. Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer New York; 2002.

19. Cover TM, Thomas JA. Elements of information theory. 2nd ed. Hoboken: Wiley; 2006.

20. Grün B, Leisch F. Fitting finite mixtures of generalized linear regressions in R. Comput Stat Data Anal. 2007;51(11):5247–52.

21. Grün B, Leisch F. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J Stat Softw. 2008;28(4):35.

22. Leisch F. FlexMix: a general framework for finite mixture models and latent class regression in R. J Stat Softw. 2004;11(8):1–18.

23. Simkens LHJ, et al. Maintenance treatment with capecitabine and bevacizumab in metastatic colorectal cancer (CAIRO3): a phase 3 randomised controlled trial of the Dutch colorectal cancer group. Lancet. 2015;385(9980):1843–52.

24. The AnyLogic Company. AnyLogic multi-method simulation software. Available from: http://www.anylogic.com/. [Accessed 8 Mar 2016]

25. Eddy DM, et al. Model transparency and validation: a report of the ISPOR-SMDM modeling good research practices task Force-7. Value Health. 2012;15(6):843–50.

26. Vemer P, et al. AdViSHE: a validation-assessment tool of health-economic models for decision makers and model users. PharmacoEconomics. 2016;34:349–61.

27. Franken MD, et al. Cost-effectiveness of capecitabine and bevacizumab maintenance treatment after first-line induction treatment in metastatic colorectal cancer. Eur J Cancer. 2017;75:204–12.

28. Barton P, et al. The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis. Health Technol Assess. 2004;8(11):104.

29. iDist: Individual patient data analysis tool. https://personex.nl/tools. [Accessed 4 July 2017]

APPENDIX 1: detailed description for the Bootstrap approach

Non-parametric bootstrapping is a statistical technique that can be used to construct an approximate sampling distribution for a statistic of interest, without the need for assumptions regarding the distribution of this statistic [1]. Several studies investigated the use of bootstrapping in health economics, e.g. for constructing confidence intervals for the incremental cost-effectiveness ratio and the incremental net benefit [2-4]. For a parameter of interest, we are interested in the value for the whole population β, which cannot be observed. Therefore, we try to find information about the value of this population parameter by drawing a random sample Y from this population and estimate $\widehat{\beta}(Y)$ the parameter of interest based on this sample. The use of bootstrapping enables us to find information about the relation between the population parameter β and its estimate $\widehat{\beta}(Y)$ by the relationship between an observed value for the parameter of interest $\widehat{\beta}(y_{obs})$ and a value for the parameter of interest based on a bootstrap sample $\widehat{\beta}(Y^*)$ [1, 3]. If non-parametric bootstrapping is applied, bootstrap sample Y^* is constructed by resampling from the observed sample y_{obs} with replacement [1, 3].

The reasoning for applying non-parametric bootstrapping to reflect the uncertainty in distributions' parameter estimates is explained by the fact that estimates from a clinical trial are almost always obtained based on a part of the total population. Hence, it would be incorrect to assume that the distributions' parameters values estimated from the data are known with certainty, i.e. are correctly describing the entire population. There is a certain sampling error, as another clinical trial may yield different estimates for the distributions' parameters. By bootstrapping trial data, other trial results are simulated and other estimates of parameters will be found, which may generate different probabilistic sensitivity analysis outcomes when Monte Carlo simulation is applied. In fact, as the size n of the sample increases, the estimates obtained by bootstrapping converge to the population value [1, 5].

In the Bootstrap approach the distributions' parameters are repeatedly, say r times, estimated based on different bootstrap samples y_r^* of the original data set y_{obs}, resulting in a set of $\{\widehat{\beta}_{1r}^*,...,\widehat{\beta}_{ir}^*\}$ parameter bootstrap-estimates, where i equals the number of parameters required to define the distribution chosen to describe the stochastic uncertainty in the time-to-event data. These bootstrap samples y_r^* are obtained by resampling the original data set y_{obs} with replacement, such that the sizes of the original data set and the bootstrap sample are the same. These r sets of parameter bootstrap estimates $\widehat{\beta}_{ir}^*$ can then be used in the PSA, incorporating one set in each Monte Carlo sample. In short, this approach can be translated into the following four steps:

1. Generate a feasible bootstrap sample of the original dataset, by resampling this dataset with replacement, such that the sample size of the bootstrap sample equals that of the original dataset.
2. Fit the pre-specified distribution(s) to the bootstrap sample and record the estimated parameter values, e.g. of the shape and rate parameters of a Gamma distribution.
3. Repeat (1) and (2) r times, where r equals the required number of PSA runs.
4. Perform the PSA, using a different set of the r parameter values to define the distribution(s) for each PSA run.

Note that if multiple distributions are fitted in Step 2, all of them need to be fitted based on the same bootstrap sample to preserve correlation among distributions, which also applies to non-time-to-event distributions, such as Beta distributions to describe utilities, and other model parameters, such as probabilities.

The definition of a feasible bootstrap sample, as required in Step 1, may vary between studies and can be difficult to decide on. When repeatedly fitting distributions to these random bootstrap samples, there might be samples for which it is not possible to fit a distribution. For example, a bootstrap sample might be drawn that contains only one time-to-event observation for a specific event, which makes it impossible to fit a Weibull distribution for the time to this event. This particular bootstrap sample can be considered infeasible for fitting a Weibull distribution, though it does contain information on the event of interest and, therefore, might be considered feasible. In such a scenario, it needs to be decided whether the bootstrap sample is to be excluded and a new sample will be drawn. Especially in case of small sample sizes or rare events these situations are likely to occur. Modelers need to be aware of this decision and should clearly communicate their choices in the corresponding publications.

APPENDIX 2: detailed description for the MVNorm approach

The uncertainty in distributions' parameter estimates can be described by assuming these parameter estimates to be Normal distributed. This is appropriate for sufficiently large sample sizes according to the Central Limit Theorem, which states that for any population distribution of a parameter of interest, the distributions of the sample means $\bar{\beta}$ will converge to Normal distributions as the sample size increases [4]. For a sufficient large sample size this indicates that the uncertainty surrounding an independent parameter can be defined according to its estimate $\hat{\beta}(y_{obs})$ and the Standard Error of the Estimate $\text{SSE}(\hat{\beta}_{y_{obs}})$:

$$\bar{\beta} \sim \text{N}(\hat{\beta}_{y_{obs}}, \text{SSE}(\hat{\beta}_{y_{obs}})) \qquad \textit{Equation 1}$$

However, when estimating a multi-parameter distribution, these distribution's parameters are likely to be correlated and it is, therefore, incorrect to define separate Normal distributions using Equation 1 for each of the parameters. However, multivariate Normal distributions can be used to draw correlated values for the parameters of interest. Multivariate Normal distributions are defined by parameter estimates $\{\hat{\beta}_1,...,\hat{\beta}_k\}$ and their variance-covariance matrix $\sum_k$:

$$\{\bar{\beta}_1,...,\bar{\beta}_k\} \sim \text{N}(\{\hat{\beta}_1,...,\hat{\beta}_k\},\sum_k) \qquad \textit{Equation 2}$$

Using multivariate Normal distributions, i.e. the MVNorm approach, the uncertainty surrounding the parameter estimates can be approximated using the following four steps:

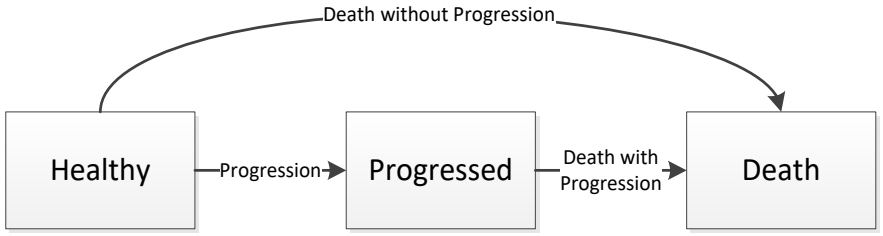
1. Fit the pre-specified distribution to the original dataset and record the estimated parameter values, e.g. of the shape and rate parameters of a Gamma distribution, and (calculate) the variance-covariance matrix.
2. Define a multivariate Normal distribution from the parameters' estimates and their variance-covariance matrix according to (1).
3. Draw r feasible sets of parameter values from the defined distribution (2), where r equals the required number of PSA runs.
4. Perform the PSA, using a different set of the r parameter values to define the distribution(s) for each PSA run.

The drawn sets of parameter values need to be assessed for their feasibility, i.e. whether the parameter values are appropriate for the pre-specified distributions. For example, if a Weibull distribution is selected to reflect stochastic uncertainty in time-to-event data, both the distribution's shape and scale parameter values need to be larger than zero, whereas (multivariate) Normal distributions are defined for any real number and may generate negative values. This theoretical definition of feasible parameter values is straightforward, though there might be scenarios in which the drawn parameter values are theoretically feasible, but rather extreme in a practical sense. For example, consider a Weibull distribution with shape and scale parameters of respectively 6 and 150, estimated based on a small sample

of time-to-event observations (e.g. $n = 25$), i.e. there might be substantial uncertainty surrounding these estimates. A draw from the corresponding multivariate Normal distribution could return a value for the scale parameter of 300, which is a theoretically feasible value but extreme in a practice sense. This value for the scale parameter will, namely, result in an expected mean value of the corresponding time-to-event distribution that is approximately twice as high with regard to the expected value of the scale parameter (i.e. 150). Although such extreme parameter values may have a substantial impact on modeling outcomes, it is hard to define which values are extreme but rather plausible and which are extreme and implausible.

APPENDIX 3: description of the model used in the simulation study

The model that is used in the simulation study is based on a basic disease progression model including three states: healthy (h), progressed (p), and death (d). The model contains two competing risks for patients in the healthy state, i.e. progression and death. A graphical representation of the disease progression model's structure is presented below:



Graphical representation of the model used in the simulation study

Based on this disease progression model, two populations of patients were simulated to sample hypothetical trials from in the simulation study. The parameters defined and the distributions used to do so are defined in the table below. In this table, the control and experimental populations are referred to by a zero (0) and an one (1) in the parameter name, respectively. As can be seen in the table, the two populations differ in terms of treatment costs and survival in the progression state. As explained in the manuscript, the parameter uncertainty surrounding non-time-to-event related variables, i.e. the utilities (u) and costs (c) was reflected according to health economic modeling good practices guidelines. The parameter uncertainty in the defined time-to-event distributions, and their related parameters such as the probability of a competing risks occurring, were addressed according to the Bootstrap or MVNorm approach.

Parameters defined and the distributions used to simulate two poputations of hypothetical patients according to the disease progression model.

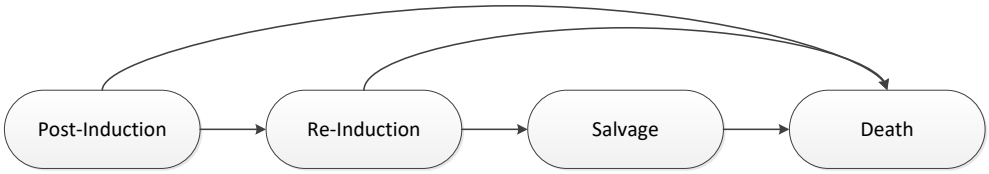
Model Parameter	Population Distribution	PSA	Description
Time to Event			
p.hp.0	Probability = 0.6	According to approach.	Probability of Progression
t.hp.0	Weibull(shape=1, scale=200)	According to approach.	Time to Progression from Healthy
t.hd.0	Weibull(shape=4, scale=150)	According to approach.	Time to Death from Health
t.pd.0	Weibull(shape=6, scale=150)	According to approach.	Time to Death from Health
p.hp.1	Probability = 0.6	According to approach.	Probability of Progression
t.hp.1	Weibull(shape=1, scale=200)	According to approach.	Time to Progression from Healthy
t.hd.1	Weibull(shape=4, scale=150)	According to approach.	Time to Death from Health
t.pd.1	Weibull(shape=5, scale=350)	According to approach.	Time to Death from Health
Effectiveness			
u.healthy	Utility = 0.9	Beta(shape1=9, shape2=1)	Utility in Healthy
u.diseased	Utility = 0.6	Beta(shape1=60, shape2=40)	Utility in Progressed
Costs			
c.progressed.0	Cost = 10	Gamma(shape=10000, rate=1000)	Costs per day
c.progressed.1	Cost = 50	Gamma(shape=500, rate=10)	Costs per day

APPENDIX 4: description of the model used in the case study

The discrete event simulation (DES) model that was used in the case study was defined on patient-level using AnyLogic software and according to the ISPOR-SMDM Modeling Good Research Practice Task Force guidelines [6]. The model was defined to have the same health states as the model that was used for the original evaluation of the CAIRO3 study: post-induction, re-induction, salvage, and death (see figure below) [7].

Weibull distributions [8] were used to define all health state-specific time-to-event parameters and were estimated from the CAIRO3 trial data using the *fitdist* function of the *fitdistrplus* [9] package in R Statistical Software [10]. Events, i.e. transitions between health states, were based on patient-specific processing times, which were randomly drawn from the estimated Weibull distributions. Competing risks were handled by selecting the first event to occur based on the respective observed event probabilities and their corresponding state-specific time-to-event distributions [11]. For example, for a patient that is entering the re-induction state a random number was compared to the chance of progression to determine whether the patient would survive and progress to the salvage therapy state. After the event was selected, the time to that event was randomly drawn from the corresponding Weibull distribution, i.e. time-to-progression or time-to-death.

In the DES model, 10,000 patients were simulated per treatment strategy. Patient-level outcomes were calculated based on the time patients had spent in each health state and were summarized to enable comparison of the two treatment strategies on population level. The DES model was validated according to good practices guidelines by assuring no unnecessary detail was present, structured "walk-throughs", comparing results with calculations by hand, extreme value analysis, trace analysis, sensitivity analysis, and cross validation with the model that was used for the original evaluation of the CAIRO3 study [12, 13].



Graphical representation of the model used in the case study.

APPENDIX 5: mean parameter estimates (including standard error) of the simulation study

* The real value refers to the value that was used to define the simulated populations (Online Resource 3). ** The true value refers to the "true" value as defined in the methods section in the manuscript, i.e. the value representing the scenario in which 2500 clinical studies were performed to estimate the parameter.

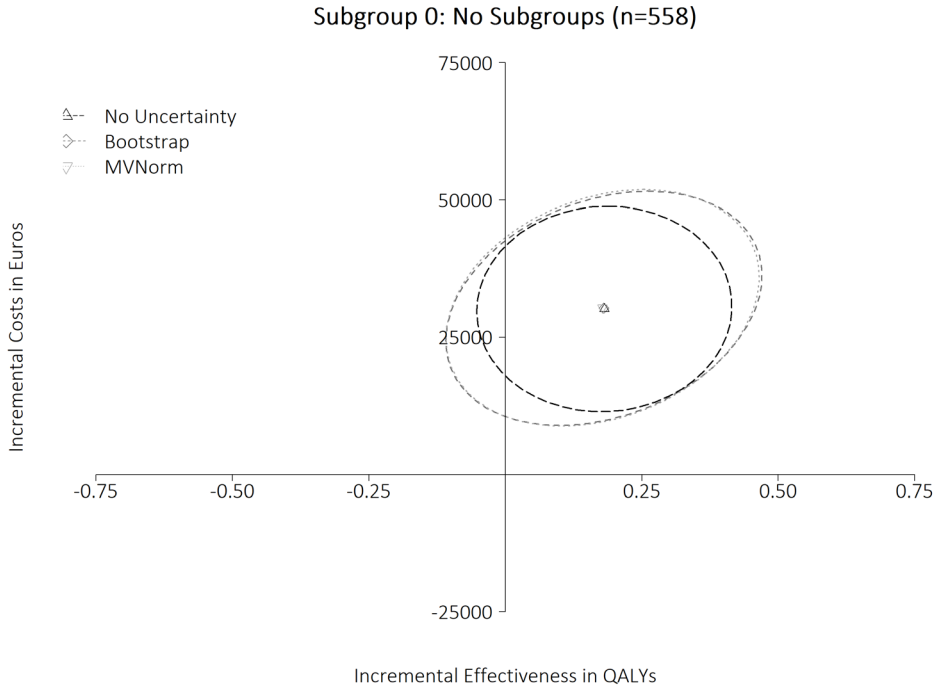
		n = 500			n = 100		
Parameter	Real Value*	True**	Boot	MVN	True**	Boot	MVN
p.hp.0	0.60	0.61	0.61	0.61	0.61	0.61	0.61
		(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
t.hp.0.shape	1.00	1.00	1.00	1.00	1.01	1.04	1.02
		(0.000)	(0.001)	(0.001)	(0.000)	(0.002)	(0.002)
t.hp.0.scale	200.00	199.08	199.40	199.29	198.91	201.10	200.43
		(0.000)	(0.237)	(0.237)	(0.000)	(0.559)	(0.558)
t.hd.0.shape	4.00	3.99	4.02	3.99	4.11	4.26	4.11
		(0.000)	(0.004)	(0.004)	(0.000)	(0.011)	(0.011)
t.hd.0.scale	150.00	149.14	149.02	149.06	149.05	148.67	148.89
		(0.000)	(0.056)	(0.056)	(0.000)	(0.128)	(0.128)
t.pd.0.shape	6.00	6.09	6.12	6.10	6.20	6.35	6.22
		(0.000)	(0.006)	(0.006)	(0.000)	(0.013)	(0.013)
t.pd.0.scale	150.00	149.54	149.51	149.51	149.42	149.35	149.40
		(0.000)	(0.029)	(0.029)	(0.000)	(0.066)	(0.066)
p.hp.1	0.60	0.61	0.61	0.61	0.61	0.61	0.61
		(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
t.hp.1.shape	1.00	1.02	1.02	1.02	1.04	1.06	1.04
		(0.000)	(0.001)	(0.001)	(0.000)	(0.002)	(0.002)
t.hp.1.scale	200.00	197.24	197.44	197.37	198.06	198.07	197.68
		(0.000)	(0.231)	(0.231)	(0.000)	(0.523)	(0.524)
t.hd.1.shape	4.00	3.97	4.00	3.97	4.08	4.22	4.08
		(0.000)	(0.004)	(0.004)	(0.000)	(0.011)	(0.011)
t.hd.1.scale	150.00	149.47	149.49	149.52	149.21	149.00	149.21
		(0.000)	(0.056)	(0.056)	(0.000)	(0.125)	(0.125)
t.pd.1.shape	5.00	4.96	4.97	4.95	5.07	5.15	5.04
		(0.000)	(0.004)	(0.004)	(0.000)	(0.011)	(0.010)
t.pd.1.scale	350.00	349.95	349.84	349.90	349.57	349.69	349.99
		(0.000)	(0.085)	(0.085)	(0.000)	(0.190)	(0.190)
incr.costs		9444.90	8921.80	8914.76	9421.60	8928.78	8909.90
		(0.000)	(7.357)	(7.350)	(0.000)	(16.884)	(16.832)
incr.effects		0.18	0.18	0.18	0.18	0.18	0.18
		(0.000)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)

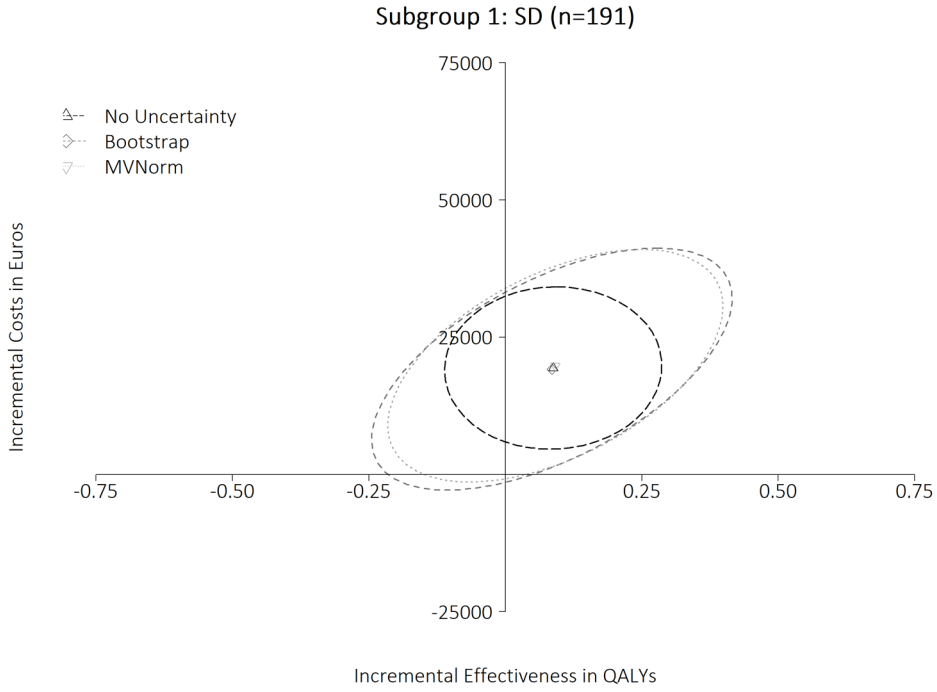
	n = 50			n = 25	
True**	Boot	MVN	True**	Boot	MVN
0.61	0.61	0.61	0.61	0.60	0.60
(0.000)	(0.001)	(0.001)	(0.000)	(0.002)	(0.002)
1.04	1.09	1.04	1.09	1.22	1.10
(0.000)	(0.003)	(0.003)	(0.000)	(0.006)	(0.005)
199.24	202.28	200.92	199.99	205.64	203.33
(0.000)	(0.788)	(0.786)	(0.000)	(1.155)	(1.152)
4.28	4.63	4.29	4.85	6.56	4.75
(0.000)	(0.019)	(0.017)	(0.000)	(0.142)	(0.035)
148.73	148.14	148.61	148.13	146.86	147.88
(0.000)	(0.182)	(0.182)	(0.000)	(0.268)	(0.269)
6.36	6.65	6.37	6.74	7.45	6.72
(0.000)	(0.020)	(0.020)	(0.000)	(0.038)	(0.032)
149.28	149.03	149.16	149.12	148.29	148.63
(0.000)	(0.094)	(0.094)	(0.000)	(0.139)	(0.139)
0.61	0.61	0.61	0.61	0.60	0.60
(0.000)	(0.001)	(0.001)	(0.000)	(0.002)	(0.002)
1.06	1.11	1.06	1.12	1.24	1.12
(0.000)	(0.003)	(0.003)	(0.000)	(0.006)	(0.005)
198.91	199.70	198.86	198.94	201.34	199.87
(0.000)	(0.728)	(0.728)	(0.000)	(1.068)	(1.068)
4.25	4.59	4.25	4.70	8.11	4.69
(0.000)	(0.018)	(0.016)	(0.000)	(1.828)	(0.033)
149.17	148.64	149.11	148.36	147.56	148.57
(0.000)	(0.187)	(0.187)	(0.000)	(0.276)	(0.277)
5.20	5.41	5.17	5.43	6.02	5.43
(0.000)	(0.017)	(0.016)	(0.000)	(0.029)	(0.025)
349.53	348.39	349.00	348.79	346.40	347.67
(0.000)	(0.274)	(0.274)	(0.000)	(0.403)	(0.403)
9419.53	8923.27	8898.46	9448.78	8852.58	*unrealistic*
(0.000)	(24.560)	(24.415)	(0.000)	(33.291)	*unrealistic*
0.18	0.18	*unrealistic*	0.18	0.17	*unrealistic*
(0.000)	(0.002)	*unrealistic*	(0.000)	(0.003)	*unrealistic*

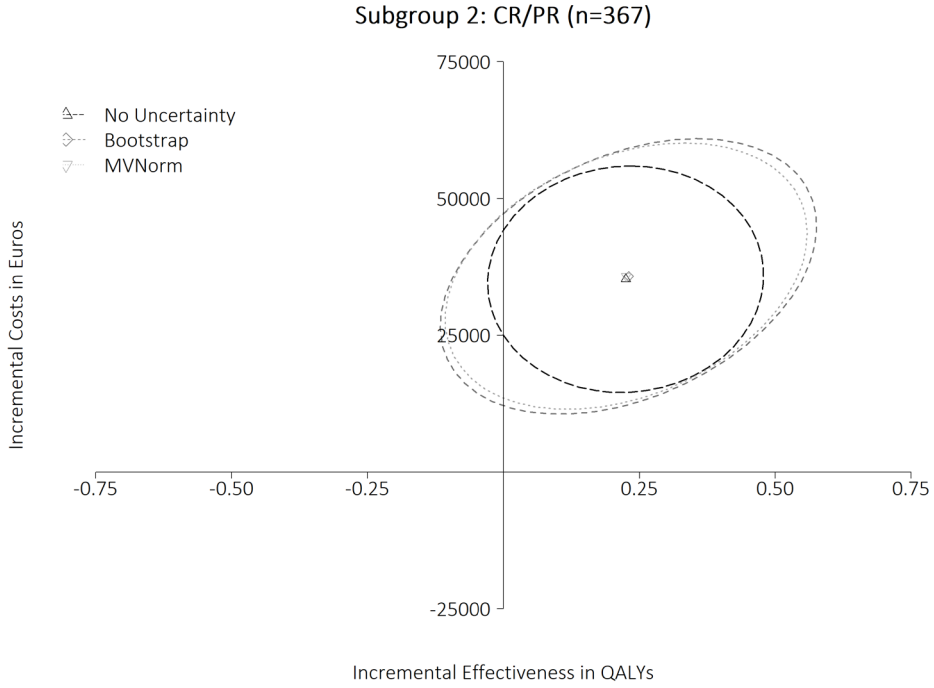
APPENDIX 6: mean Kullback-Leibler Divergence results (including standard error) for the simulation study

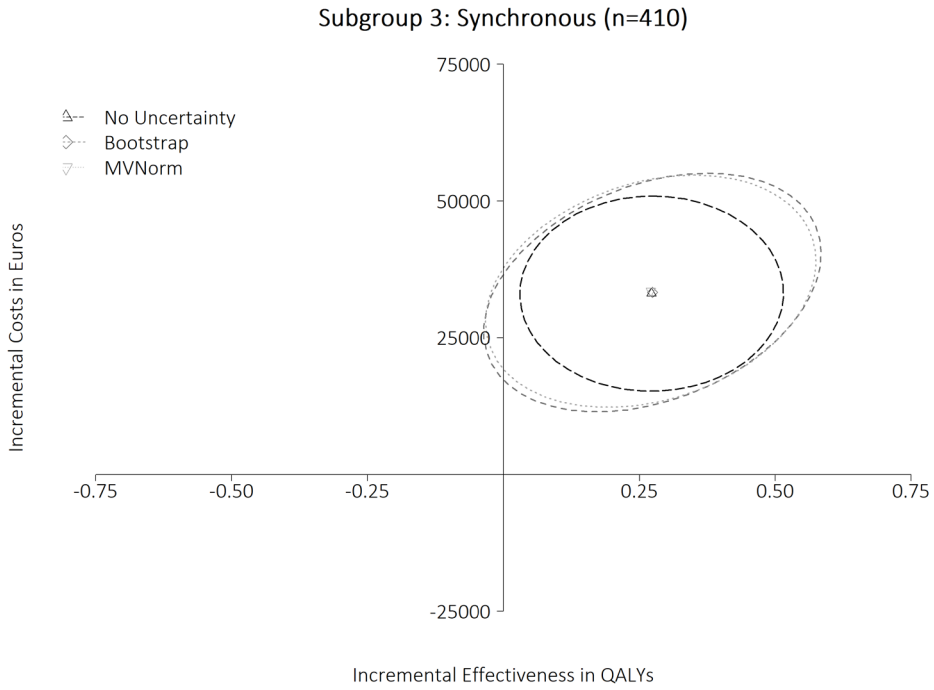
	n = 500		n = 100		n = 50		n = 25	
Parameter	**Boot**	**MVN**	**Boot**	**MVN**	**Boot**	**MVN**	**Boot**	**MVN**
p.hp.0	0.46	0.48	0.49	0.50	0.48	0.50	0.50	0.52
	(0.013)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.015)	(0.015)
t.hp.0.shape	0.50	0.49	0.59	0.55	0.70	0.61	0.84	0.67
	(0.015)	(0.014)	(0.017)	(0.016)	(0.021)	(0.016)	(0.026)	(0.017)
t.hp.0.scale	0.45	0.46	0.48	0.50	0.48	0.50	0.50	0.52
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.013)
t.hd.0.shape	0.50	0.48	0.60	0.55	0.72	0.62	0.60	0.50
	(0.015)	(0.014)	(0.017)	(0.015)	(0.020)	(0.015)	(0.020)	(0.011)
t.hd.0.scale	0.48	0.48	0.52	0.52	0.56	0.56	0.68	0.68
	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)	(0.017)	(0.017)
t.pd.0.shape	0.51	0.48	0.58	0.51	0.66	0.56	0.75	0.61
	(0.015)	(0.013)	(0.017)	(0.014)	(0.018)	(0.014)	(0.021)	(0.014)
t.pd.0.scale	0.46	0.47	0.49	0.49	0.52	0.51	0.59	0.58
	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)	(0.013)	(0.015)	(0.015)
p.hp.1	0.48	0.46	0.52	0.50	0.54	0.52	0.52	0.50
	(0.013)	(0.013)	(0.015)	(0.014)	(0.016)	(0.015)	(0.016)	(0.014)
t.hp.1.shape	0.50	0.49	0.58	0.55	0.66	0.59	0.82	0.65
	(0.014)	(0.014)	(0.017)	(0.015)	(0.020)	(0.015)	(0.025)	(0.016)
t.hp.1.scale	0.45	0.47	0.46	0.48	0.44	0.46	0.47	0.50
	(0.012)	(0.012)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)
t.hd.1.shape	0.49	0.47	0.61	0.51	0.69	0.55	0.84	0.64
	(0.014)	(0.013)	(0.018)	(0.014)	(0.020)	(0.014)	(0.024)	(0.015)
t.hd.1.scale	0.46	0.46	0.48	0.48	0.56	0.57	0.69	0.70
	(0.013)	(0.012)	(0.012)	(0.012)	(0.014)	(0.015)	(0.017)	(0.017)
t.pd.1.shape	0.49	0.48	0.56	0.53	0.65	0.57	0.77	0.61
	(0.015)	(0.014)	(0.016)	(0.014)	(0.019)	(0.014)	(0.023)	(0.015)
t.pd.1.scale	0.48	0.48	0.48	0.48	0.52	0.52	0.57	0.57
	(0.013)	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)	(0.014)	(0.014)
incr.costs	0.39	0.39	0.14	0.14	0.09	0.09	0.05	0.05
	(0.006)	(0.006)	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)
incr.effects	0.39	0.39	0.51	0.49	0.53	0.48	0.05	0.05
	(0.011)	(0.011)	(0.014)	(0.014)	(0.015)	(0.013)	(0.005)	(0.004)
average	0.47	0.46	0.50	0.49	0.55	0.51	0.58	0.52

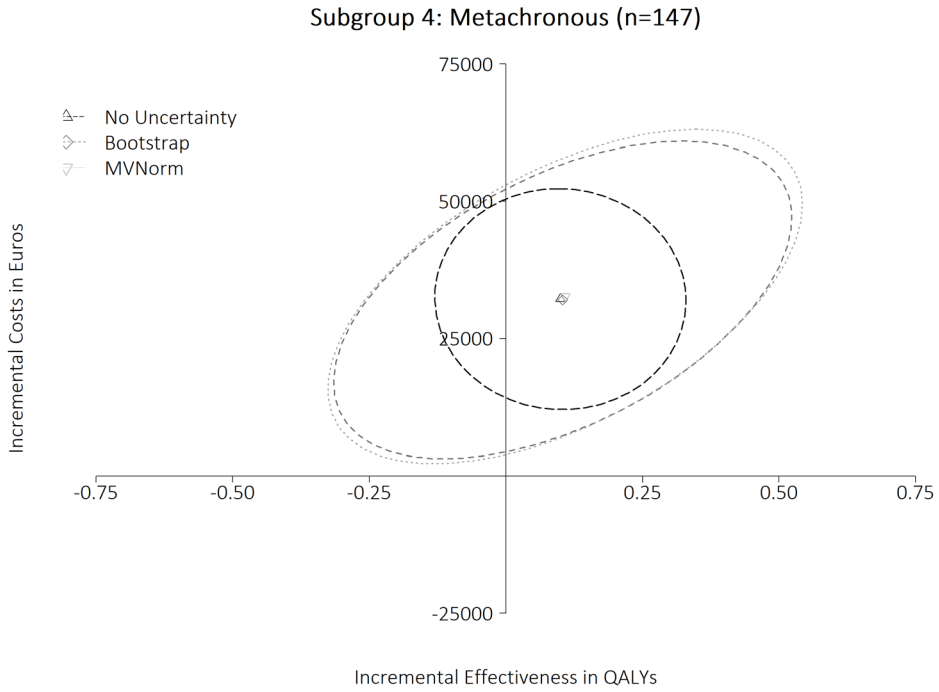
APPENDIX 7: incremental cost-effectiveness planes for the cohort and all subgroups of the case study

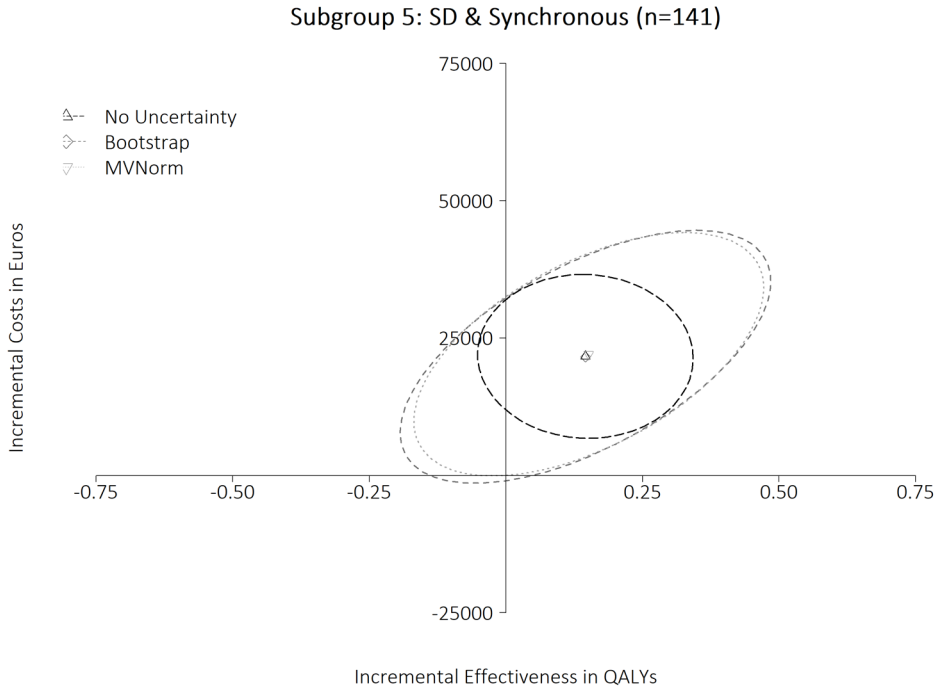


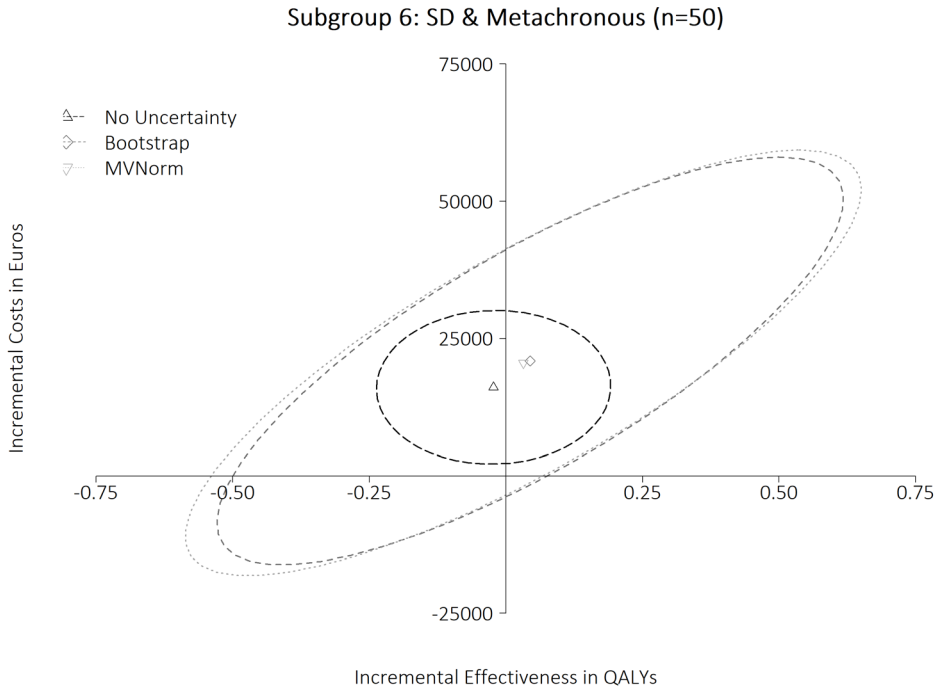


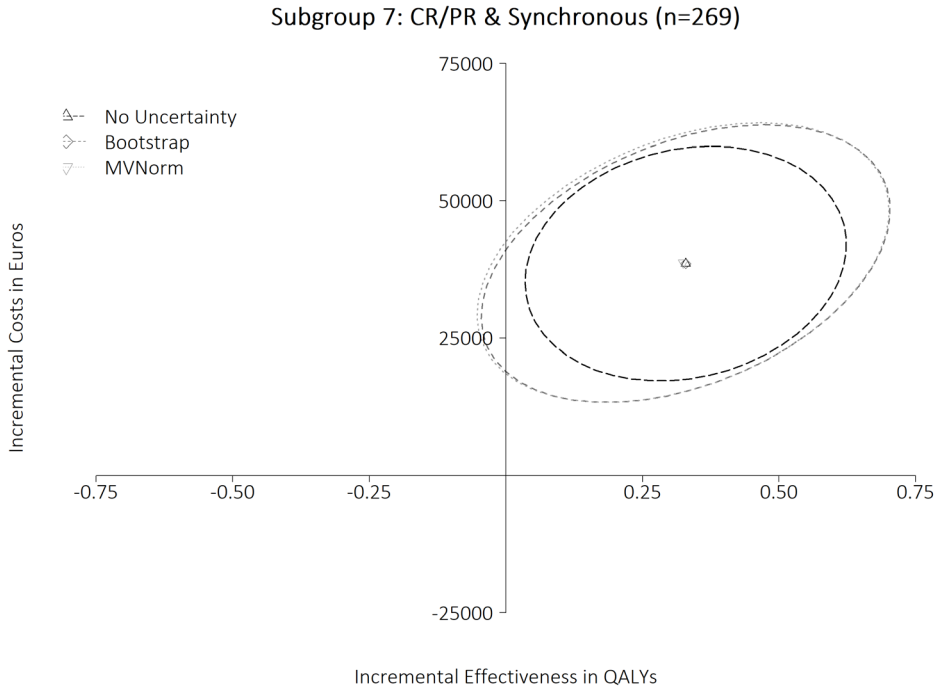

Subgroup 2: CR/PR (n=367)
No Uncertainty
Bootstrap
MVNorm
Incremental Costs in Euros
75000
50000
25000
-25000
-0.75
-0.50
-0.25
0.25
0.50
0.75
Incremental Effectiveness in QALYs

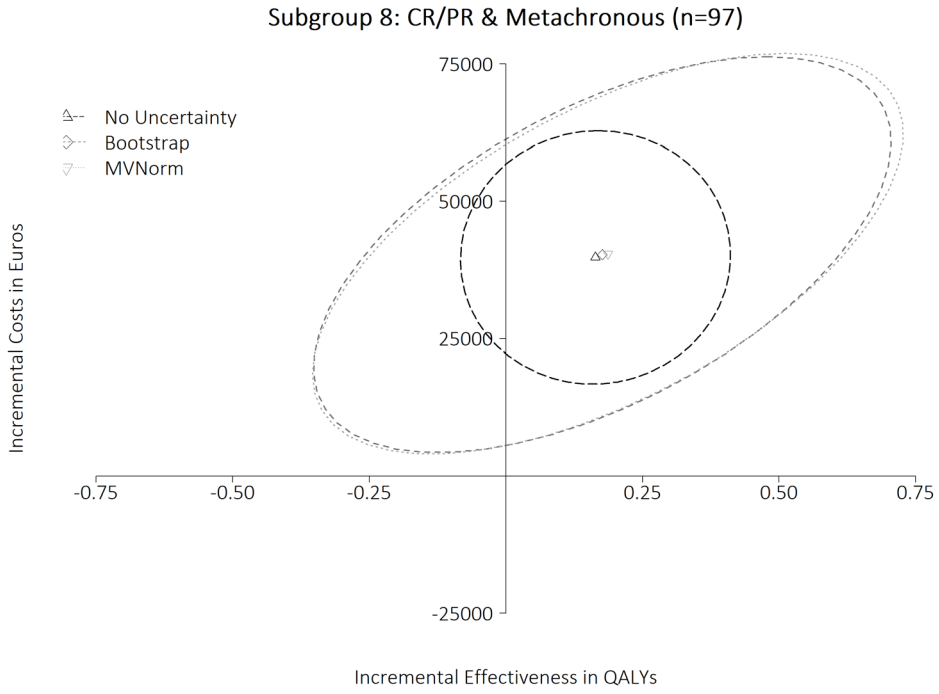

Subgroup 3: Synchronous (n=410)
No Uncertainty
Bootstrap
MVNorm
Incremental Costs in Euros
75000
50000
25000
-25000
-0.75
-0.50
-0.25
0.25
0.50
0.75
Incremental Effectiveness in QALYs


Subgroup 4: Metachronous (n=147)
No Uncertainty
Bootstrap
MVNorm
75000
50000
25000
-25000
-0.75
-0.50
-0.25
0.25
0.50
0.75
Incremental Costs in Euros
Incremental Effectiveness in QALYs

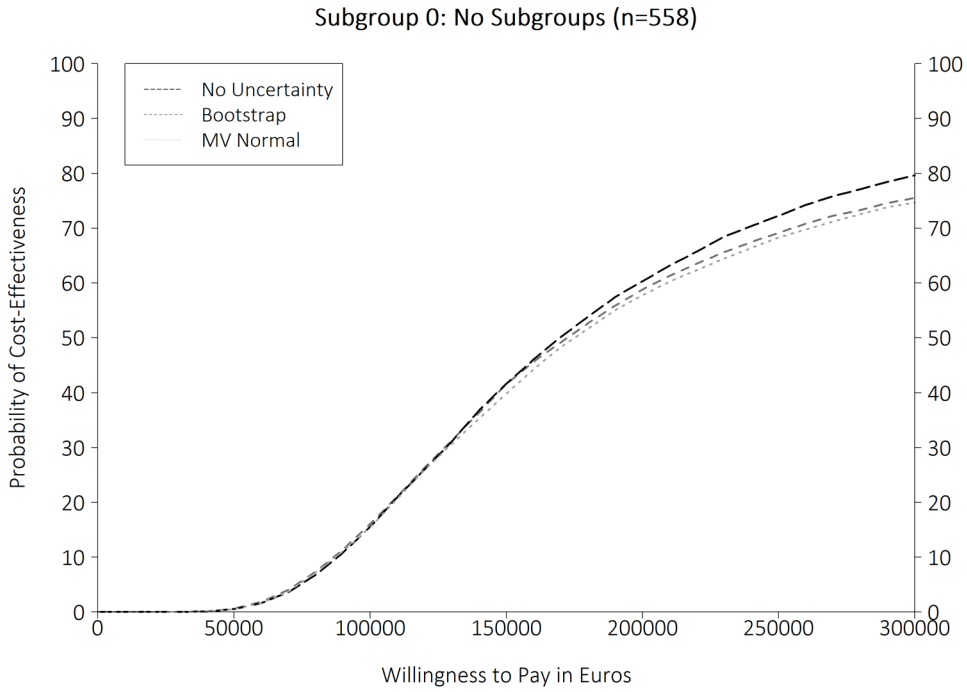

Subgroup 5: SD & Synchronous (n=141)
No Uncertainty
Bootstrap
MVNorm
75000
50000
25000
-25000
-0.75
-0.50
-0.25
0.25
0.50
0.75
Incremental Costs in Euros
Incremental Effectiveness in QALYs

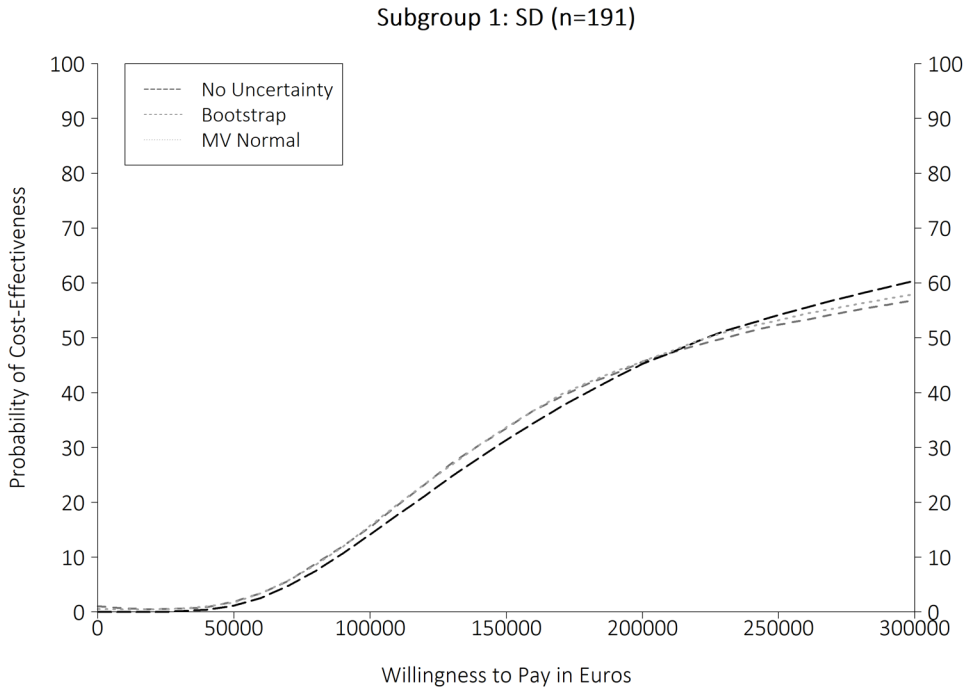

Subgroup 6: SD & Metachronous (n=50)
No Uncertainty
Bootstrap
MVNorm
75000
50000
25000
-25000
-0.75
-0.50
-0.25
0.25
0.50
0.75
Incremental Costs in Euros
Incremental Effectiveness in QALYs

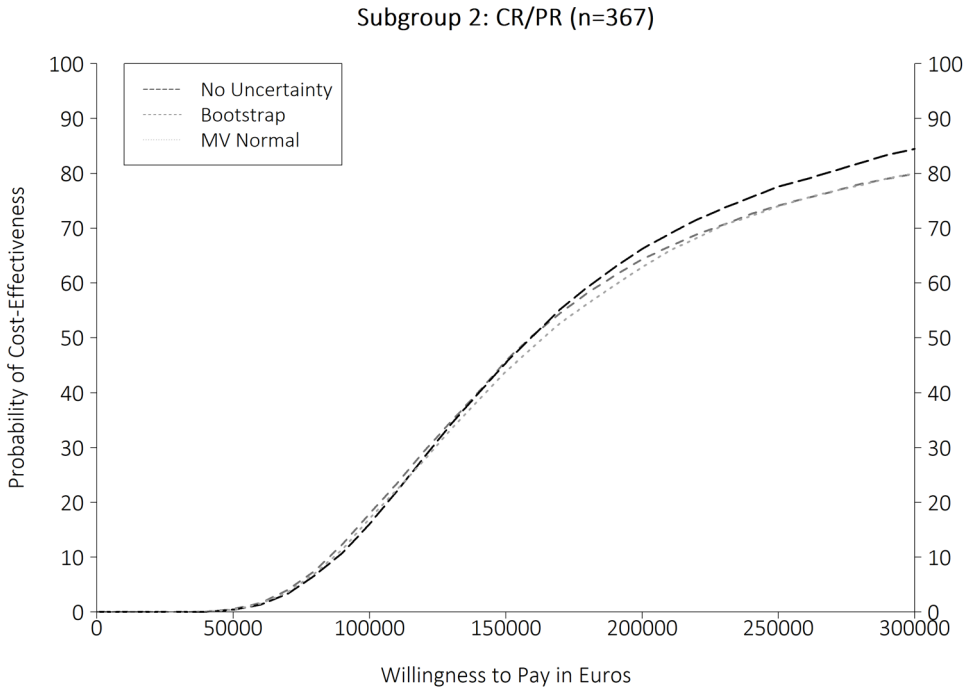

Subgroup 7: CR/PR & Synchronous (n=269)
No Uncertainty
Bootstrap
MVNorm
75000
50000
25000
-25000
-0.75
-0.50
-0.25
0.25
0.50
0.75
Incremental Costs in Euros
Incremental Effectiveness in QALYs

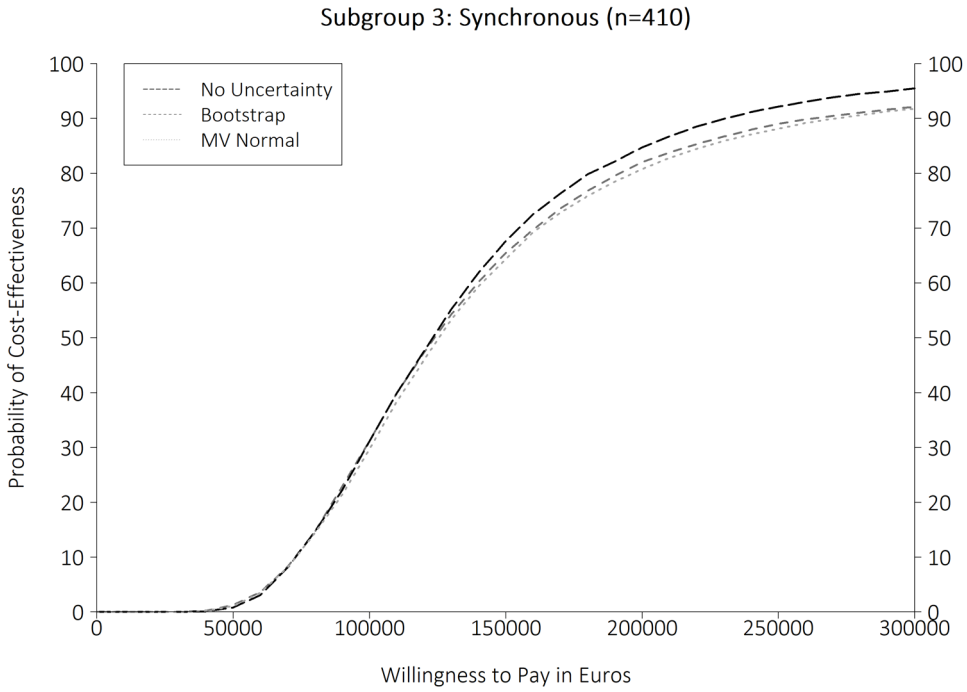

Subgroup 8: CR/PR & Metachronous (n=97)
No Uncertainty
Bootstrap
MVNorm
75000
50000
25000
-25000
-0.75
-0.50
-0.25
0.25
0.50
0.75
Incremental Costs in Euros
Incremental Effectiveness in QALYs

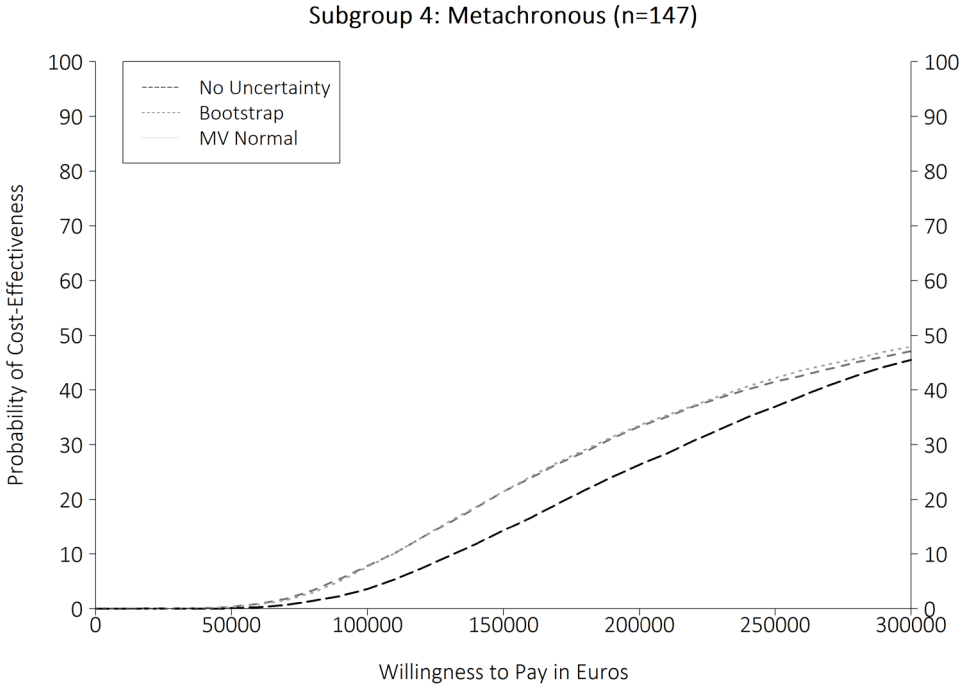
APPENDIX 8: Cost-Effectiveness Acceptability Curves for the cohort and all subgroups of the case study

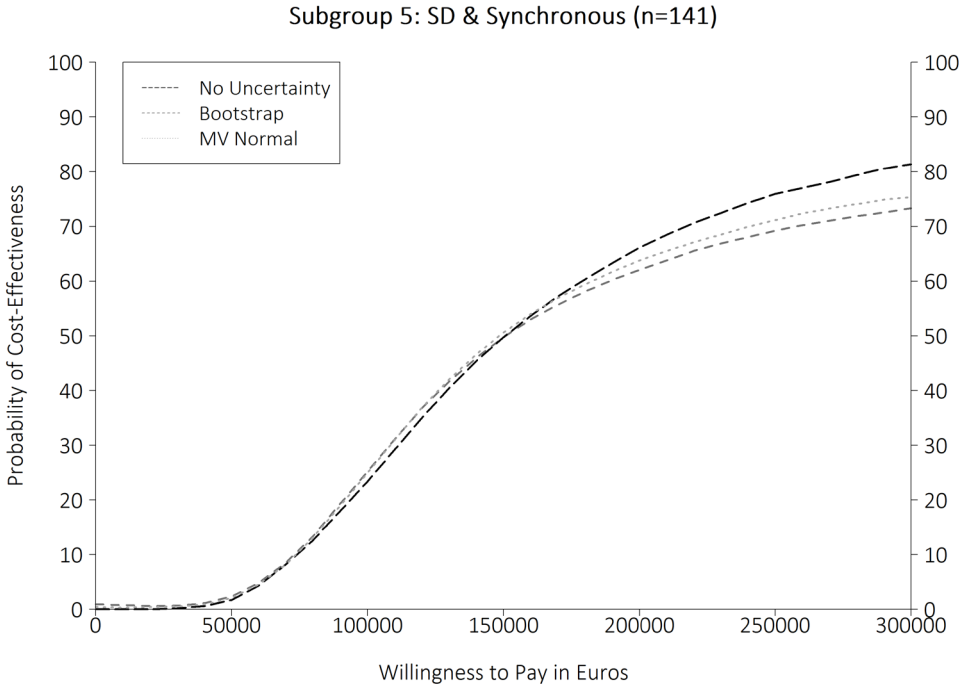


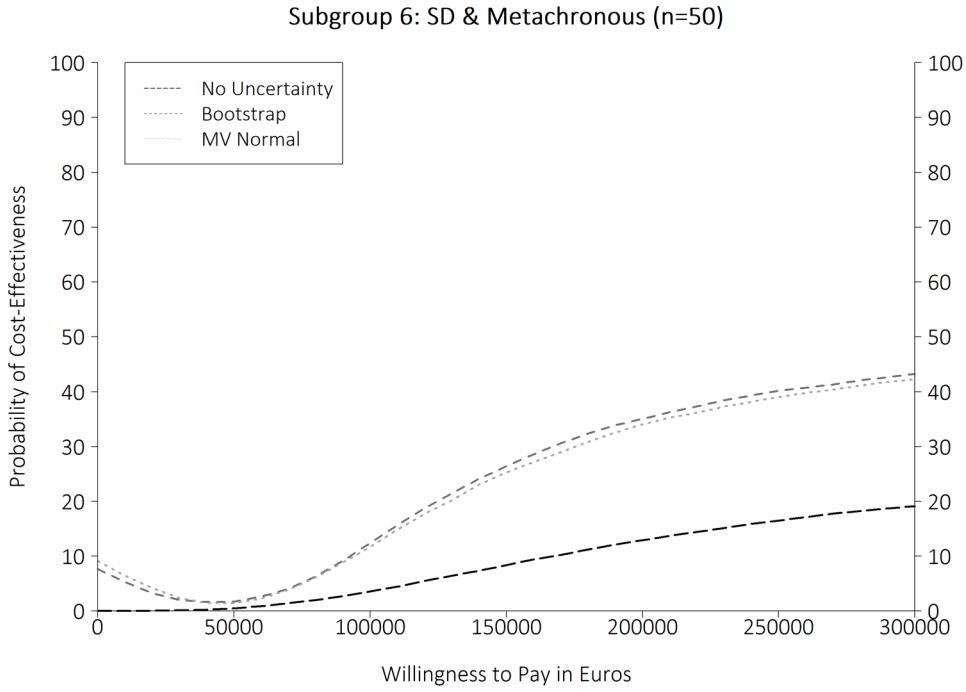


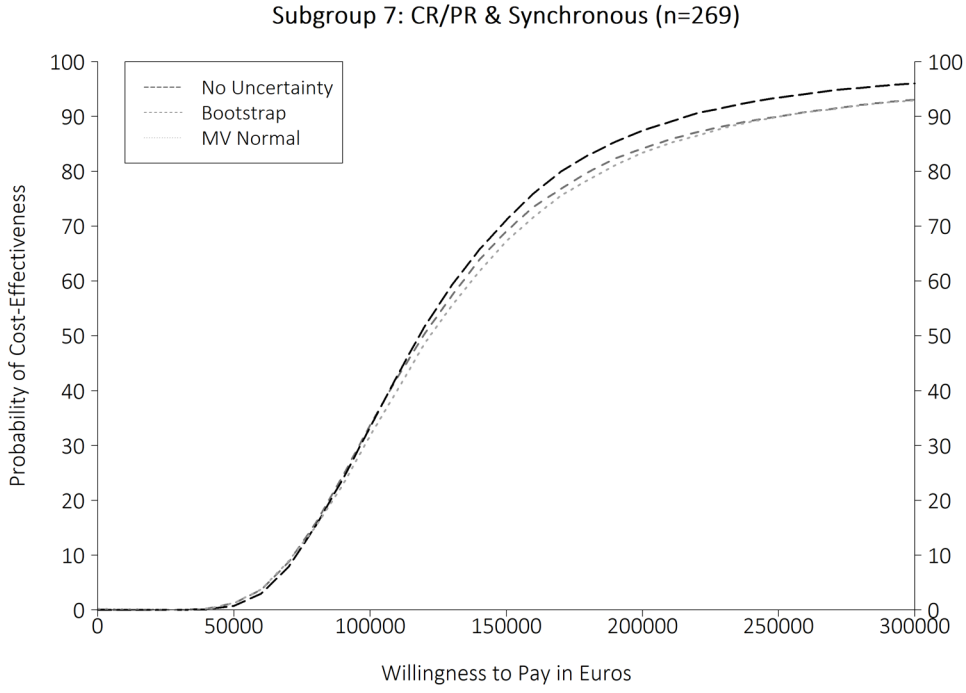

Subgroup 2: CR/PR (n=367)
No Uncertainty
Bootstrap
MV Normal
Probability of Cost-Effectiveness
Willingness to Pay in Euros
0
10
20
30
40
50
60
70
80
90
100
0
50000
100000
150000
200000
250000
300000

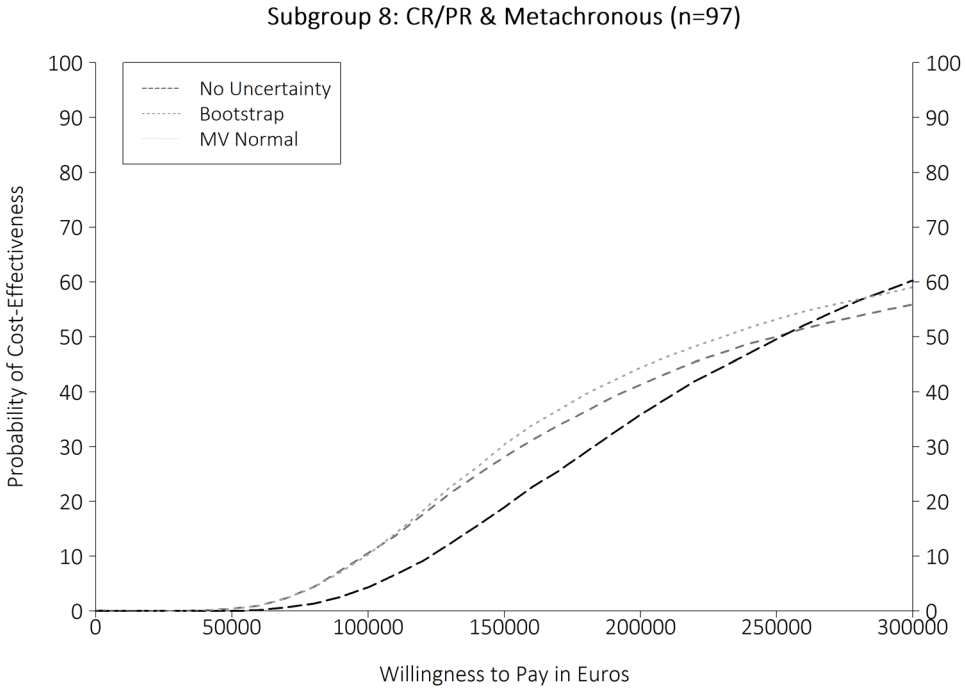

Subgroup 3: Synchronous (n=410)
No Uncertainty
Bootstrap
MV Normal
Probability of Cost-Effectiveness
Willingness to Pay in Euros
0
10
20
30
40
50
60
70
80
90
100
0
50000
100000
150000
200000
250000
300000


Subgroup 4: Metachronous (n=147)
No Uncertainty
Bootstrap
MV Normal
Probability of Cost-Effectiveness
Willingness to Pay in Euros
0
10
20
30
40
50
60
70
80
90
100
0
50000
100000
150000
200000
250000
300000


Subgroup 5: SD & Synchronous (n=141)
No Uncertainty
Bootstrap
MV Normal
Probability of Cost-Effectiveness
Willingness to Pay in Euros
0
10
20
30
40
50
60
70
80
90
100
0
50000
100000
150000
200000
250000
300000


Subgroup 6: SD & Metachronous (n=50)
No Uncertainty
Bootstrap
MV Normal
Probability of Cost-Effectiveness
Willingness to Pay in Euros
0
10
20
30
40
50
60
70
80
90
100
0
50000
100000
150000
200000
250000
300000


Subgroup 7: CR/PR & Synchronous (n=269)
No Uncertainty
Bootstrap
MV Normal
Probability of Cost-Effectiveness
Willingness to Pay in Euros
0
10
20
30
40
50
60
70
80
90
100
0
50000
100000
150000
200000
250000
300000


Subgroup 8: CR/PR & Metachronous (n=97)
No Uncertainty
Bootstrap
MV Normal
Probability of Cost-Effectiveness
Willingness to Pay in Euros
0
10
20
30
40
50
60
70
80
90
100
0
50000
100000
150000
200000
250000
300000

APPENDIX 9: references used in the Appendices

1. Davison AC, Hinkley. DV. Bootstrap Methods and their Application. 1st ed. Cambridge University Press; 1997.

2. Briggs AH, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. Health economics. 1997;6(4):327-40.

3. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med. 2000;19(9):1141-64.

4. Nixon RM, Wonderling D, Grieve RD. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. Health Economics. 2010;19(3):316-33.

5. Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC press; 1994.

6. Karnon J, Stahl J, Brennan A, Caro JJ, Mar J, Möller J. Modeling using Discrete Event Simulation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. Value in Health. 2012;15(6):821-7.

7. Franken MD, van Rooijen EM, May AM, Koffijberg H, van Tinteren H, Mol L, ten Tije AJ, Creemers GJ, van der Velden AMT, Tanis BC, Uyl-de Groot CA, Punt CJA, Koopman M, van Oijen MGH. Cost-effectiveness of capecitabine and bevacizumab maintenance treatment after first-line induction treatment in metastatic colorectal cancer. European Journal of Cancer. 2017;75:204-12.

8. Collett D. Modelling Survival Data in Medical Research, Second Edition. Taylor & Francis; 2003.

9. Delignette-Muller ML, Dutang C. fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software. 2015;64(4):1-34.

10. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

11. Barton P, Jobanputra P, Wilson J, Bryan S, Burls A. The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis. Health Technology Assessment. 2004;8(11):104.

12. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Value in Health. 2012;15(6):843-50.

13. Vemer P, Corro Ramos I, van Voorn GAK, Al MJ, Feenstra TL. AdViSHE: A Validation-Assessment Tool of Health-Economic Models for Decision Makers and Model Users. PharmacoEconomics. 2016;34:349-61.

CHAPTER 6

Comparing strategies for modeling competing risks in discrete-event simulations: a simulation study and illustration in colorectal cancer

Koen Degeling
Hendrik Koffijberg
Mira D. Franken
Miriam Koopman
Maarten J. IJzerman

This chapter has been published as:
Degeling K, Koffijberg H, Franken MD, Koopman M, IJzerman MJ. Comparing strategies for modeling competing risks in discrete-event simulations: a simulation study and illustration in colorectal cancer. Medical Decision Making. 2019;39(1):57-73.

ABSTRACT

Background: Different strategies toward implementing competing risks in discrete-event simulation (DES) models are available. This study aims to provide recommendations regarding modeling approaches that can be defined based on these strategies by performing a quantitative comparison of alternative modeling approaches.

Methods: Four modeling approaches were defined: 1) event-specific distribution (ESD), 2) event-specific probability and distribution (ESPD), 3) unimodal joint distribution and regression model (UDR), and 4) multimodal joint distribution and regression model (MDR). Each modeling approach was applied to uncensored individual patient data in a simulation study and a case study in colorectal cancer. Their performance was assessed in terms of relative event incidence difference, relative absolute event incidence difference, and relative entropy of time-to-event distributions. Differences in health economic outcomes were also illustrated for the case study.

Results: In the simulation study, the ESPD and MDR approaches outperformed the ESD and UDR approaches, in terms of both event incidence differences and relative entropy. Disease pathway and data characteristics, such as the number of competing risks and overlap between competing time-to-event distributions, substantially affected the approaches' performance. Although no considerable differences in health economic outcomes were observed, the case study showed that the ESPD approach was most sensitive to low event rates, which negatively affected performance.

Conclusions: Based on overall performance, the recommended modeling approach for implementing competing risks in DES models is the MDR approach, which is defined according to the general strategy of selecting the time-to-event first and the corresponding event second. The ESPD approach is a less complex and equally performing alternative if sufficient observations are available for each competing event (i.e., the internal validity shows appropriate data representation).

INTRODUCTION

The use of discrete-event simulation (DES) for evaluating health care interventions is expected to increase due to the complexity of many novel personalized treatment options [1]. In particular its ability to model dynamic pathways based on patient-level histories and patient characteristics makes DES particularly useful for representing personalized treatment processes [2,3]. Besides dynamic model structures and flexibility toward defining different levels of abstraction throughout a model, DES embodies alternative approaches for modeling the occurrence of events. In contrast to more commonly applied patient-level discrete-time state transition models (STMs) [4], DES does not require time to be defined by fixed discrete cycles but allows for parametric distributions to be directly implemented to represent patient-level time-to-event variation. Uncertainty in parametric distributions' parameters can be accounted for in probabilistic sensitivity analyses, so that both stochastic uncertainty (i.e., first-order uncertainty) and parameter uncertainty (i.e., second-order uncertainty) are reflected [5]. Although parametric distributions can also be used to populate STMs, this requires an additional discretization step, that is, evaluation of the cumulative density functions at fixed time points, to obtain discrete-time transition probabilities.

Health economic models that are informed with individual patient data (IPD) and employ DES for patient-level simulations are subject to several design choices [6], one of which relates to the approach taken to implement competing risks. Since competing risks are present in every clinical pathway and may affect the observation of outcomes in clinical studies, it is important to appropriately represent competing risks in health economic models [7]. Modelers are provided with a high degree of flexibility with regard to selecting one or multiple strategies for implementing competing risks in DES models. According to Barton et al. [8], 4 general strategies are available: 1) simulating times for all competing events and selecting the event that is the first to occur, 2) selecting the event to occur first and the corresponding time-to-event second, 3) selecting the time-to-event first and the corresponding event second, and 4) using discretized cyclic probabilities to resemble discrete-time STM.

According to the Professional Society for Health Economics and Outcomes Research (ISPOR) and Society for Medical Decision Making (SMDM) modeling good research practices guidelines, the recommended strategy toward implementing competing risks in DES models is Strategy 3 as proposed by Barton et al. [8]: selecting the time-to-event first, based on a joint time-to-event distribution, and then selecting the corresponding event [6]. However, a thorough quantitative comparison of the strategies available has not yet been performed, which might increase variation in approaches taken to represent competing risks in published health economic DES models. For example, recently published studies used approaches based on the strategy of using discretized cyclic probabilities [9,10], selecting the event first and the time-to-event second [11,12], or selecting the event that is the first to occur [13-15]. Moreover, the specification and motivation of the approach taken to implement competing risks are not always made explicit in modeling publications [16,17]. This may be due to space limitations or, alternatively, due to limited awareness of the availability of different approaches.

The objective of this study is to describe, illustrate, and compare different approaches for handling competing risks in DES models informed by uncensored IPD. The comparison will ideally lead to generalized recommendations, so modelers can make informed and deliberative decisions regarding the handling of competing risk data in DES models. To achieve this objective, approaches are compared in a simulation study to assess their accuracy in representing the incidence of competing events and corresponding event-specific time-to-event distributions, in terms of bias and relative entropy, respectively. Subsequently, the approaches are applied in a case study based on uncensored patient-level data obtained from the randomized controlled CAIRO3 trial [18] to illustrate potential differences in health economic outcomes of a cost-effectiveness analysis in colorectal cancer.

METHODS

We focus on 3 general strategies for handling competing risks identified by Barton et al. [8]: 1) select the event that is the first to occur, 2) select the event first and the time-to-event second, and 3) select the time-to-event first and the event second. The strategy of using discretized cyclic probabilities was deliberately not included, as this approach would discard major advantages of using DES and resemble a STM. Moreover, this would require arbitrary decisions about the cycle length and time dependency of transition probabilities, creating the need to also assess the impact of these decisions on simulation outcomes. Based on the 3

Table 1. *Hypothetical patient-level time-to-event data used to illustrate the modeling approaches included in this study. The data concerns nine hypothetical time-to-event observations of patients that were subject to two competing risks: progression and death. Progr., Progression; C , censored at the time of observing the competing event; T represents a random variable to describe a time-to-event; D represents a certain unimodal time-to-event distribution (e.g., Weibull or Gamma) defined by parame-*

Observed Data			**ESD Approach**		**ESPD Approach**	
Patient ID	**Time to Event**	**Observed Event**	**Time to Death**	**Time to Progr.**	**Event is Death**	**Event is Progr.**
1	154	Death	154	C	Yes	No
2	351	Progr.	C	351	No	Yes
3	294	Progr.	C	294	No	Yes
4	310	Progr.	C	310	No	Yes
5	257	Death	257	C	Yes	No
6	211	Death	211	C	Yes	No
7	438	Progr.	C	438	No	Yes
8	80	Progr.	C	80	No	Yes
9	347	Progr.	C	347	No	Yes
Example of Data Analysis Output:			$T_{Death} \sim D(\beta_{Death1}, \beta_{Death2})$	$T_{Progr} \sim D(\beta_{Progr1}, \beta_{Progr2})$	$P(Death)$	$P(Progr)$

included strategies, 4 specific modeling approaches for implementing competing risks in DES models informed by uncensored IPD were defined, which are described below in more detail. Pseudo-algorithms for data analysis and simulation according to these modeling approaches are provided in Boxes 1, 2, 3, and 4. Additional code illustrating how these approaches can be implemented in R Statistical Software [19] is provided online at www.personex.nl/research/competing-risks.

Modeling Approach 1: ESD

The ESD approach implements a cause-specific hazards model [20,21] and uses event-specific time-to-event distributions to draw random times at which the competing events would occur and subsequently selects the first event to occur to be simulated. To estimate the event-specific distributions, observations of competing events are considered censored observations because those competing events prohibit observing the event of interest [21]. Box 1 reviews how IPD is analyzed according to the ESD modeling approach and how the resulting event-specific time-to-event distributions can be implemented in simulation models. To illustrate this for the illustrative data presented in Table 1, a distribution D_e needs to be fitted for each competing event e by estimating n parameters $\beta_1,...,\beta_n$ that define D_e, resulting in separate distributions for both the time-to-death and time-to-progression. When estimating the time-to-death distribution D_{Death}, the 3 patients who died (i.e., Patient 1, Patient 5, and Patient 6) are considered observations and those who progressed are considered right-cen-

ters β_1 and β_2; $P(...)$ represents the probability that the corresponding competing event will occur; f represents a (multinomial) regression model defined by parameters $\beta_1,...,\beta_n$ predicting the competing event to occur (dependent variable) based on the time-to-event T (independent variable); MD represents a multimodal distribution defined by n parameters $\beta_1,...,\beta_n$.

		UDR Approach		MDR Approach	
Time to Death	**Time to Progr.**	**Time to Event**	**Observed Event**	**Time to Event**	**Observed Event**
154	-	154	Death	154	Death
-	351	351	Progr.	351	Progr.
-	294	294	Progr.	294	Progr.
-	310	310	Progr.	310	Progr.
257	-	257	Death	257	Death
211	-	211	Death	211	Death
-	438	438	Progr.	438	Progr.
-	80	80	Progr.	80	Progr.
-	347	347	Progr.	347	Progr.
$T \mid \text{Death} \sim D(\beta_{Death1}, \beta_{Death2})$	$T \mid \text{Progr} \sim D(\beta_{Progr1}, \beta_{Progr2})$	$T \sim D(\beta_1, \beta_2)$	$Event \mid T \sim f(\beta_1,...,\beta_n)$	$T \sim MD(\beta_1, \beta_2)$	$Event \mid T \sim f(\beta_1,...,\beta_n)$

Box 1. *ESD Approach.*

Data Analysis (see Table 1)

1.1 For each competing event e, fit a time-to-event distribution D_e:

- Observations of patients are censored (C) at the time a competing event occurs
- Select a distribution type to represent the time-to-event T_e for event e
- Estimate n parameters $\beta_{e1},...,\beta_{en}$ that define distribution D_e

Simulation

1.2 Obtain time-to-events for each competing event:

- Draw a time t_e for each competing event e by performing a random draw from the corresponding distribution D_e

1.3 Select the competing event to occur:

- Select event k with the lowest time-to-event across all events (i.e., the first event to occur)

1.4 Simulate the selected event k at the corresponding time t_k

Box 2. *ESPD Approach.*

Data Analysis (see Table 1)

2.1 For each competing event e, estimate the probability p_e this event occurs:

- Estimate probability p_e as the proportion of patients experiencing competing event e

2.2 For each competing risk e, fit a time-to-event distribution D_e:

- Only include observations of patients that experienced competing event e
- Select a distribution type to represent the time-to-event T_e for event e
- Estimate n parameters $\beta_{e1},...,\beta_{en}$ that define distribution D_e

Simulation

2.3 Select the competing event to occur:

- Draw a random number r from a Uniform distribution $U[0,1]$
- Select the event to k occur, by comparing event probabilities p_e and random number r

2.4 Obtain a time-to-event for the selected event:

- Draw a time t for selected event k by performing a random draw from the corresponding distribution D_k

2.5 Simulate the selected event k at the corresponding time t

sored observations (C), as these patients would have died from the initial state at some point in time if they would not have progressed [20]. In the corresponding simulation model, a time t_e to each of the competing events e needs to be drawn randomly from each time-to-event distribution D_e . Subsequently, the event that is the first to occur (i.e., the event e corresponding to the lowest drawn time-to-event t_e) is selected and will be simulated.

Modeling Approach 2: ESPD

The ESPD approach implements a subdistribution hazards model [20,21] and first selects the event to be simulated based on event-specific probabilities, and it subsequently draws the time at which that event will occur from an event-specific time-to-event distribution. Box 2 reviews how IPD is analyzed according to the ESPD modeling approach and how the resulting event-specific probabilities and time-to-event distributions can be implemented in simulation models. Event-specific probabilities p_e are defined as the cumulative incidence function limit [20], simply representing the probability that corresponding competing event e occurs, which is 3 out of 9 for death in the exemplary data of Table 1. Notice that probabilities p_e of competing events always add up to 1 over the lifetime of a patient. The event-specific time-to-event distributions D_e are estimated solely based on observations of patients who experienced competing event of interest e . Considering the data in Table 1, this indicates that the time-to-death distribution D_{Death} is estimated based on the time-to-events of the 3 patients who died (i.e., Patient 1, Patient 5, and Patient 6). In the corresponding simulation model, a random number r needs to be compared to the event-specific probabilities p_e to select the event k that a hypothetical patient will experience. Subsequently, a time-to-event t needs to be drawn randomly from distribution D_k corresponding to the selected event k .

Modeling Approach 3: UDR

The UDR approach first selects the time at which an event will be simulated based on a joint time-to-event distribution, representing all competing events, and selects the event corresponding to the selected time-to-event using a (multinomial) logistic regression model. Multinomial logistic regression models are required for cases involving more than 2 competing risks, as standard logistic regression models can only account for binary data (i.e., 2 competing risks). When estimating the joint time-to-event distribution, the UMR approach assumes the joint time-to-event distribution to be unimodal. Box 3 reviews how IPD is analyzed according to the UDR modeling approach and how the resulting time-to-event distribution and (multinomial) logistic regression model can be implemented in simulation models. For the exemplary data presented in Table 1, this indicates that to represent time-to-event distribution D , for example, a single Weibull distribution [22] is estimated based on all time-to-event observations. Next, logistic regression model f [23], predicting the type of event (dependent variable), needs to be estimated based on the time-to-event (independent variable). In the corresponding simulation model, a random time-to-event t needs to be drawn from the joint distribution D , which then can be used together with a random number r to select the corresponding event k to occur using probabilities p_e obtained from the logistic regression model f .

Box 3. *UDR Approach.*

Data Analysis (see Table 1)

3.1 For all events combined, fit a joint unimodal time-to-event distribution D:
- Include the observations of all patients
- Select an unimodal distribution to represent the time-to-event T for all competing events
- Estimate n parameters $\beta_1,...,\beta_n$ that define distribution D

3.2 Fit a (multinomial) logistic regression model to predict the competing event to occur:
- Estimate the parameters $\beta_1,...,\beta_n$ of (multinomial) logistic regression model f that predicts the probabilities p_e of each competing event e to occur (dependent variable) based on the time-to-event t (independent variable)

Simulation

3.3 Obtain a time-to-event for the event to occur:
- Draw a time t for the event to occur by performing a random draw from the joint distribution D

3.4 Select the competing event to occur:
- Obtain probabilities p_e for each competing event to occur based on time-to-event t, using (multinomial) logistic regression model f
- Draw a random number r from a Uniform distribution $U[0,1]$
- Select the event k to occur, by comparing event probabilities p_e and random number r

3.5 Simulate the selected event k at the corresponding time t

Modeling Approach 4: MDR

The MDR approach is similar to the UDR approach, except for the fact that the MDR approach does not assume the time-to-event data to be unimodal distributed but allows the joint time-to-event distribution MD to be multimodal. This implies that, for example, a phase type [24] or mixture [25] distribution can be estimated to represent the patient-level variation in time-to-event values. Box 4 reviews how IPD is analyzed according to the MDR modeling approach and how the resulting time-to-event distribution and (multinomial) logistic regression model can be implemented in simulation models. Except for the type of distribution used to represent the time-to-event data, the data analysis and simulation processes are the same as those for the UDR modeling approach illustrated in Box 3.

Simulation Study to Compare the Performance of the Modeling Approaches

A simulation study was performed to compare the accuracy of the approaches and assess whether data and disease pathway characteristics, such as the number of competing risks, affect the performance. This simulation study included the analysis and simulation of the incidence and time-to-event distributions of competing risks from an initial health state (Figure

Box 4. *MDR Approach.*

Data Analysis (see Table 1)

4.1 For all events combined, fit a multimodal time-to-event distribution MD:
- Include the observations of all patients
- Select a multimodal distribution to represent the time-to-event T for all competing events
- Estimate n parameters $\beta_1,...,\beta_n$ that define distribution MD

4.2 Fit a (multinomial) logistic regression model to predict the competing event to occur:
- Estimate the parameters $\beta_1,...,\beta_n$ of (multinomial) logistic regression model f that predicts the probabilities p_e of each competing event e to occur (dependent variable) based on the time-to-event t (independent variable)

Simulation

4.3 Obtain a time-to-event for the event to occur:
- Draw a time t for the event to occur by performing a random draw from the joint distribution MD

4.4 Select the competing event to occur:
- Obtain probabilities p_e for each competing event to occur based on time-to-event t, using (multinomial) logistic regression model f
- Draw a random number r from a Uniform distribution $U[0,1]$
- Select the event k to occur, by comparing event probabilities p_e and random number r

4.5 Simulate the selected event k at the corresponding time t

1a). As illustrated in the simulation study overview presented in Figure 2 and as reviewed in Box 5, the simulation study was run for i = 9 different patient populations P_i, which were simulated according to unique combinations of the number of competing risks (i.e., 2, 3, or 4 competing risks) and the degree of overlap between the corresponding competing time-to-event distributions (i.e., low ~ 10%, medium ~ 50%, and high ~ 90% overlap). Figure 3 illustrates the different levels of overlap for a population defined by 3 competing risks. Details on the exact population definitions and calculation of overlap between distributions are provided in Appendix 1 and Appendix 2, respectively. In addition, the simulation study was performed for different hypothetical trial arm sample sizes n_{sample} (n_{sample} = 50, 100, 200, 500) to assess sample size impact on the performance of the approaches, resulting in a total of 36 unique scenarios.

A total of j = 10,000 simulation runs were performed for each unique population P_i and sample size n_{sample} combination. In each of these runs, a hypothetical trial arm sample p_{ijn} of the applicable sample size n_{sample} was randomly sampled from the corresponding population P_i. Next, the hypothetical trial sample p_{ijn} was analyzed according to the m = 1, 2, 3, 4 mode-

ling approaches, based on which the incidence and time-to-event distributions were simulated for n_{sim} = 10,000 new patients to obtain simulation sample s_{ijnm}. Finally, the performance of the approaches was assessed by comparing the event incidence and time-to-event distributions in these newly simulated patients s_{ijnm} to those in the population P_i (i.e., external validation) and hypothetical trial sampled from this population p_{ijn} (i.e., internal validation). Regarding the incidence of events, the bias in terms of relative incidence difference (%) and relative absolute incidence difference (%) of the approaches in s_{ijnm} compared to the population P_i and trial sample p_{ijn} were assessed. The performance with regard to the simulated time-to-event distributions was obtained by comparing the simulated event-specific distributions in s_{ijnm} to those of the population P_i and trial sample p_{ijn} based on the relative entropy (i.e., the Kullback-Leibler divergence) [26]. The relative entropy is a measure of the difference between 2 probability distributions, for which lower values indicate a better performance. To summarize the relative absolute incidence difference and Kullback-Leibler divergence for each of the 36 scenarios (i.e., combinations of P_i and n_{sample}), event-specific performance outcomes were weighted according to event incidences in the population. Event-specific relative incidence differences were not weighted to obtain summarized performance measures, because weighing relative incidence differences based on the incidence does not result in meaningful outcomes (i.e., outcomes of approximately zero).

The simulation study was performed in R Statistical Software version 3.3.2 [19]. All time-to-event data were simulated and analyzed using Weibull distributions [22] to rule out potential bias due to mismatching distributions. Weibull distributions were selected to represent patient-level time-to-event variation in the simulation study, because these distributions are commonly used in survival analysis and accurately represent the IPD of the case study. Univariate Weibull distributions were estimated using the *fitdistrplus* package [27] for the ESD, ESPD, and UDR approaches. For the MDR approach, the *mixtools* package [28] was used to estimate Weibull mixture distributions, providing the parameter estimates of the ESPD approach as starting values to increase the likelihood of convergence. If the algorithm for estimating a mixture distribution did not converge in a specific simulation run, the parameter estimates of the ESPD approach were used to define the corresponding mixture distribution. The *nnet* [29] package was used to estimate (multinomial) logistic regression models. The Kullback-Leibler divergence was determined using the *flexmix* package [30-32].

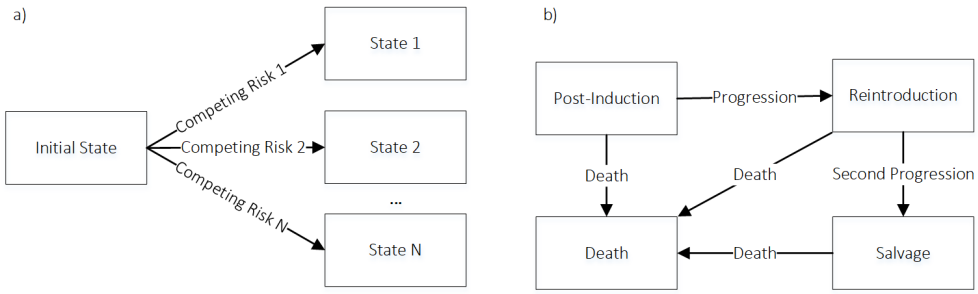


Figure 1. *Overview of a) the structure of the discrete-event simulation (DES) model used in the simulation study and b) the structure of the case study DES model.*

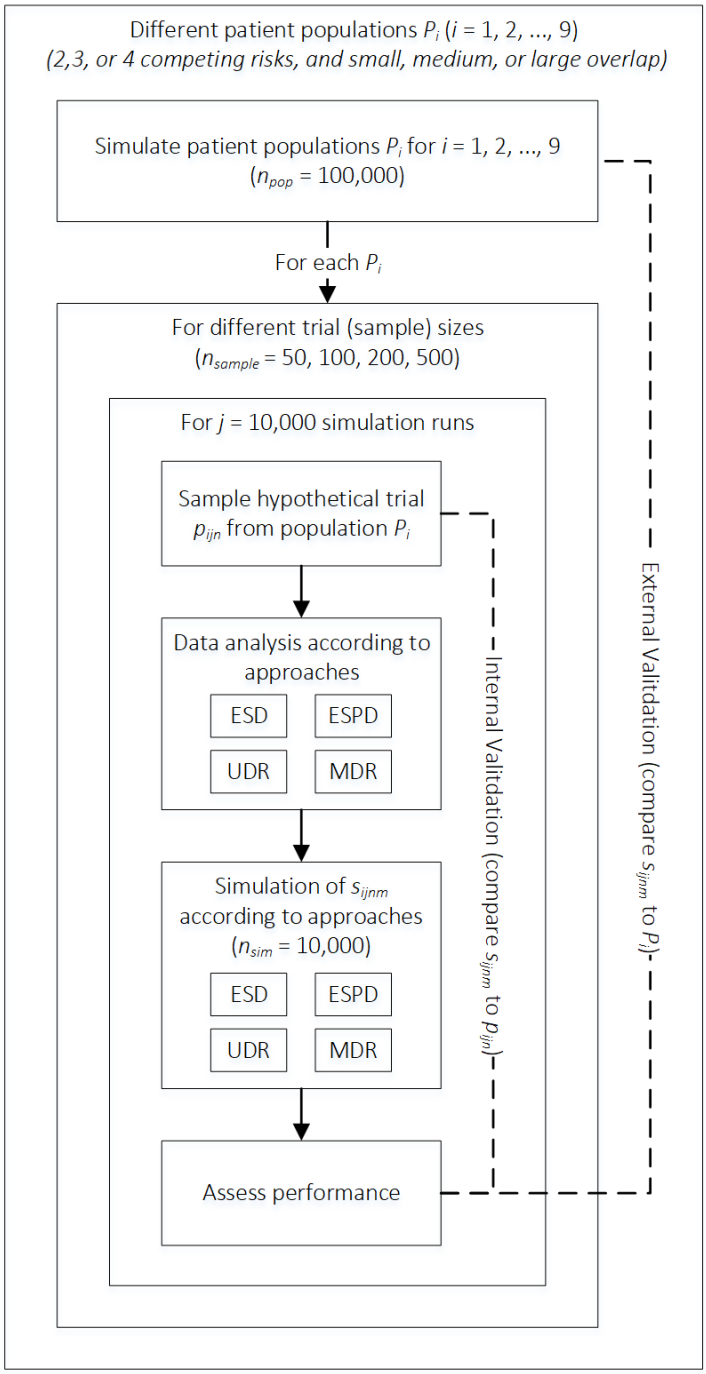


Figure 2. *Overview of the simulation study. ESD, event-specific distribution; ESPD, event-specific probability and distribution; MDR, multimodal joint distribution and regression model; UDR, unimodal joint distribution and regression model.*

Box 5. *Simulation Study (also see Figure 2).*

5.1 Simulate different patient populations according to the number of competing risks and time-to-event distribution overlap:

- Define patient population i as P_i for i = 1, 2, ... , 9
- Each P_i represents a combination of 2, 3, or 4 competing risks and a small, medium, or large overlap of time-to-event distributions (see Figure 3)
- Simulate n_{pop} = 100,000 hypothetical patients to define P_i according to the number of competing risks and time-to-event distribution overlap

5.2 For each population P_i (i = 1, 2, ... , 9):

5.3 For different trial sizes n_{sample} = 50, 100, 200, 500:

5.4 For j = 10,000 simulation runs:

5.5 Sample a hypothetical trial of size n_{sample} from patient population P_i:

- Define p_{ijn} as the j^{th} trial sample of size n_{sample} of population P_i

5.6 Analyze p_{ijn} according to each modeling approach (see Boxes 1, 2, 3, and 4)

5.7 Simulate a sample of size n_{sim} = 10,000 patients according to each modeling approach (see Boxes 1, 2, 3, and 4):

- Define s_{ijnm} as the j^{th} simulation sample according to modeling approach m (m = 1, 2, 3, 4) based on a sample p_{ijn} of size n_{sample} and population P_i

5.8 Assess the performance of each modeling approach:

- Calculate the relative incidence difference, relative absolute incidence difference, and relative entropy based on internal and external validation
- Internal validation: compare simulation samples s_{ijnm} for m = 1, 2, 3, 4 to trial sample p_{ijn}
- External validation: compare simulation samples s_{ijnm} for m = 1, 2, 3, 4 to population P_i

Illustration of Competing Risks Modeling Approaches in Colorectal Cancer

To assess the modeling approaches' potential performance and impact on health economic outcomes in real-word scenarios, a case study was performed based on an anonymized data set from the randomized phase 3 CAIRO3 study (NCT00442637) of the Dutch Colorectal Cancer Group. The CAIRO3 study randomized 558 metastatic colorectal cancer patients with stable disease or better after 6 cycles of capecitabine, oxaliplatin, and bevacizumab (CAPOX-B) induction therapy to either capecitabine and bevacizumab (CAP-B) maintenance treatment (intervention) or observation (control) until progression of disease [18]. For both the maintenance and observation strategy, CAPOX-B treatment was to be re-induced upon progression and continued until second progression (PFS2), the primary end point of the study. The orig-

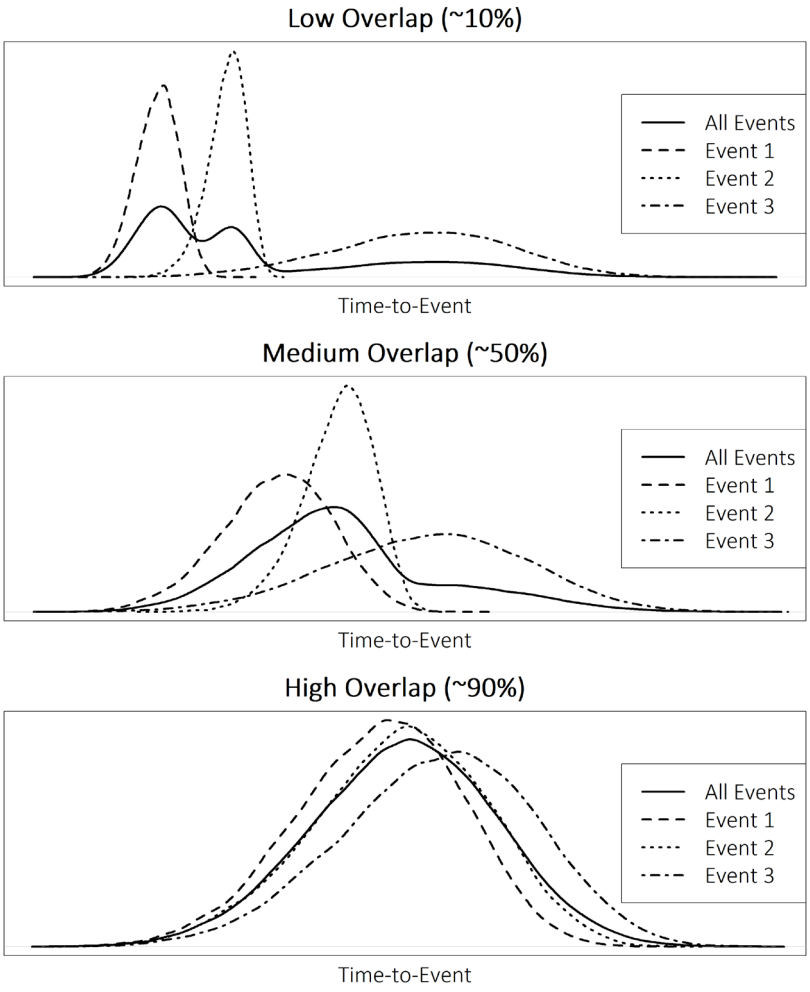


Figure 3. *Illustration of the different levels of overlap between competing time-to-event distributions used in the simulation study.*

inal health economic evaluation of the CAIRO3 study was based on a discrete-time cohort STM and has been published elsewhere [33].

A previously developed DES model [5] was implemented in R Statistical Software version 3.3.2 [19], according to the structure of the CAIRO3 study: postinduction, reintroduction, salvage, and death (Figure 1b). Model state postinduction refers to observation (control) or CAP-B maintenance treatment (intervention) after 6 cycles of CAPOX-B. Reintroduction of CAPOX-B refers to modeling state reintroduction for both treatment strategies (i.e., observation and CAP-B maintenance treatment). If patients progress from the reintroduction state (i.e., the cancer no [longer] responds to reintroduction of CAPOX-B), salvage therapy (i.e., alternative

treatment options) is provided, which refers to the salvage state. As illustrated in Figure 1b, patients were subject to competing risks progression and death while in the postinduction state and to second progression and death in the reintroduction state. The original DES model handled these competing risks according to the ESPD approach; here, 3 alternative versions of the model were created to reflect the other approaches (i.e., ESD, UDR, and MDR). For the case study, all modeling approaches were implemented using the same R packages as were used for the simulation study.

Stochastic uncertainty (i.e., patient-level variation) in health state durations was accounted for by using Weibull (mixture) distributions. Parameter uncertainty in probabilities, parametric distributions, and regression models used to reflect time-to-event evidence according to the 4 approaches was accounted for by bootstrapping the CAIRO3 data set [5], averting potential bias by applying all approaches to the exact same bootstrap samples. For the other model parameters, parameter uncertainty was accounted for as in the original health economic evaluation of the CAIRO3 study [33]. Probabilistic sensitivity analysis was performed based on 10,000 runs of 10,000 patients per treatment strategy in each run.

Clinical relevant subgroup analyses were performed to illustrate potential sample size impact on modeling outcomes for the different approaches. A total of 8 subgroups with sample sizes ranging from 50 to 410 were defined according to patient characteristics that were found relevant in the evaluation of the CAIRO3 study [18], that is, treatment response (stable disease (SD) v. complete or partial response (CR/PR)) and stage of disease (synchronous v. metachronous) (Appendix 3). This stratification strategy resulted in subgroups for which events were observed only once, or not all, which resembles the personalized medicine context.

For all subgroups, the accuracy of the approaches was compared based on the internal validity in terms of relative incidence difference (%), relative absolute incidence difference (%), and Kullback-Leibler divergence over all probabilistic sensitivity analysis runs. To obtain these performance measures, simulations according to the approaches were compared to the bootstrap sample based on which of the approaches' parameters had been estimated. Incremental cost-effectiveness planes and cost-effectiveness acceptability curves presented the health economic outcomes.

RESULTS

Simulation Study to Compare the Performance of the Modeling Approaches

Mean weighted results for the relative absolute incidence difference and relative entropy over all runs of the simulation study are presented in Table 2. Detailed event-specific results for the relative absolute incidence difference and relative incidence difference are not presented in this article to enhance readability but are available in the Supplementary Materials.

The bias in terms of relative incidence difference is substantially better for the ESPD and MDR approaches compared to the ESD and UDR approaches. Using the ESD or UDR approach results in a substantial amount of bias for higher numbers of competing risks and lower overlap

between their corresponding time-to-event distributions. For example, for the population defined by 4 competing risks and low overlapping distributions, the ESD and UDR approaches yield up to approximately 60% and 50% bias, respectively, whereas the ESPD and MDR approaches yield up to approximately 5% bias. For all approaches, bias is lower when simulation outcomes are compared to the hypothetical trial (i.e., internal validation) rather than the simulated population (i.e., external validation). Furthermore, overall performance is better for lower numbers of competing risks, higher overlap between time-to-event distributions, and larger sample sizes, although performance of the ESD and UDR approaches is much more sensitive to changes in these data and disease pathway characteristics.

Although the bias in terms of relative absolute event incidence difference shows less extreme outcomes for the ESD and UDR approaches, these approaches are again outperformed by the ESPD and MDR approaches. Furthermore, overall performance is better with regard to the hypothetical trial (i.e., internal validation) than the simulated population (i.e., external validation). Also, with respect to the bias in terms of relative absolute incidence difference, mainly the ESD and UDR approaches benefit from lower numbers of competing risks, higher overlap between the competing time-to-event distributions, and higher sample sizes.

The performance in terms of relative entropy shows the same trends in differences between approaches. The ESPD and MDR approaches strongly outperform their ESD and UDR equivalents, although the Kullback-Leibler divergence also shows that the ESPD approach slightly outperforms the MDR approach. Once more, the performance in terms of internal validity is better compared to the external validity, and especially the performance of the ESD and UDR approaches benefits from lower numbers of competing risks, higher overlap between the competing time-to-event distributions, and higher sample sizes.

Illustration of Competing Risks Modeling Approaches in Colorectal Cancer

Mean results for the relative incidence difference and relative entropy for the case study over all runs of the probabilistic sensitivity analysis are presented in Table 3. Results for the relative absolute incidence difference are not presented in the article but are available in Appendix 4. The relative absolute incidence differences showed negligible differences compared to the absolute value of the relative incidence differences, indicating that the approaches underestimated or overestimated the event incidences systematically.

The internal validation of the approaches shows similar trends for the case study as for the simulation study. The ESPD and MDR approaches overall yield slightly better relative incidence differences and relative absolute incidence differences. For example, for the cohort analysis (Subgroup 0), the mean relative incidence difference in the probability of progression from the postinduction state for the control group is 1.9%, 0.0%, 0.9%, and 0.4% for the ESD, ESPD, UDR, and MDR approaches, respectively. Interestingly, the results also suggest that the performance in terms of bias for the ESPD approach is more sensitive to low event rates compared to the other approaches, which is illustrated by the mean relative incidence differences for the intervention group of Subgroup 6. Only 2 of 17 patients in this subgroup died during the reintroduction state (Appendix 3), resulting in a mean relative incidence difference in the

Table 2. *Mean accuracy of the approaches in terms of relative absolute incidence difference and relative entropy over all simulation study runs.*

Number of Events	Distribution Overlap (%)	Sample Size	Incidence of Events *Relative Absolute Difference (%)*							
			External Validation					Internal Validation		
			Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial
2	10	50	11.5	14.2	11.6	12.3	11.5	7.4	0.8	6.5
2	10	100	7.6	10.2	7.7	8.2	7.6	6.9	0.8	4.3
2	10	200	5.7	8.5	5.7	6.1	5.7	6.8	0.8	3.4
2	10	500	3.4	7.5	3.6	4.1	3.6	6.8	0.8	3.0
2	50	50	10.8	11.0	10.9	11.0	10.8	1.5	0.8	2.0
2	50	100	7.7	7.9	7.8	8.0	7.8	1.3	0.8	1.7
2	50	200	5.5	5.7	5.6	5.5	5.5	1.2	0.8	1.5
2	50	500	3.5	3.7	3.6	3.7	3.5	1.2	0.7	1.4
2	90	50	10.9	10.9	10.9	10.9	10.8	0.9	0.8	0.8
2	90	100	7.8	7.8	7.8	7.8	7.8	0.8	0.8	0.8
2	90	200	5.7	5.7	5.7	5.7	5.7	0.8	0.8	0.8
2	90	500	3.5	3.6	3.6	3.5	3.5	0.8	0.8	0.8
3	10	50	16.0	19.5	16.0	21.6	16.0	9.8	1.2	22.0
3	10	100	10.9	14.7	11.0	20.5	11.0	9.6	1.1	21.2
3	10	200	8.0	11.9	8.1	20.0	8.0	9.5	1.1	20.1
3	10	500	5.0	10.1	5.1	18.8	5.1	9.3	1.1	19.0
3	50	50	14.9	16.1	15.0	14.7	14.9	4.1	1.1	7.3
3	50	100	11.6	12.6	11.7	12.1	11.7	3.9	1.2	6.9
3	50	200	8.1	9.0	8.2	9.5	8.2	3.8	1.1	6.6
3	50	500	4.9	6.0	5.0	7.4	5.1	3.7	1.1	6.5
3	90	50	15.0	15.1	15.1	15.1	15.2	1.4	1.1	1.2
3	90	100	10.9	11.1	11.0	11.0	11.1	1.3	1.1	1.1
3	90	200	8.0	8.1	8.1	8.0	8.0	1.2	1.1	1.2
3	90	500	4.9	5.0	4.9	5.0	5.0	1.2	1.1	1.1
4	10	50	18.2	22.6	18.3	29.7	18.3	12.3	1.4	30.5
4	10	100	13.0	18.3	13.2	28.3	13.1	12.1	1.4	28.2
4	10	200	9.2	15.4	9.3	27.1	9.3	11.8	1.3	26.9
4	10	500	5.8	13.1	5.9	26.2	5.9	11.7	1.3	25.9
4	50	50	17.2	19.8	17.3	21.0	17.3	9.8	1.3	17.1
4	50	100	12.4	15.9	12.6	17.8	12.5	9.1	1.3	15.4
4	50	200	9.4	13.3	9.5	16.1	9.5	8.9	1.3	14.4
4	50	500	5.9	10.6	6.0	15.0	6.1	8.7	1.3	14.3
4	90	50	17.8	17.7	18.0	18.1	17.9	2.8	1.3	2.0
4	90	100	12.8	12.9	12.9	12.8	12.9	2.2	1.3	1.5
4	90	200	9.2	9.5	9.3	9.3	9.4	2.0	1.4	1.4
4	90	500	5.9	6.3	6.1	6.1	6.1	1.9	1.3	1.4

	Time-to-Event Distributions Kullback-Leibler Divergence								
	External Validation					Internal Validation			
MDR vs. Trial	Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
1.2	0.073	0.237	0.044	0.117	0.064	0.227	0.041	0.115	0.045
1.0	0.038	0.224	0.020	0.101	0.029	0.211	0.024	0.093	0.024
0.9	0.021	0.218	0.010	0.095	0.014	0.206	0.015	0.085	0.014
0.8	0.011	0.215	0.005	0.092	0.007	0.206	0.008	0.084	0.008
0.9	0.073	0.035	0.043	0.045	0.054	0.055	0.042	0.061	0.050
0.9	0.039	0.021	0.020	0.030	0.025	0.034	0.025	0.041	0.028
0.8	0.022	0.014	0.010	0.023	0.013	0.022	0.015	0.030	0.017
0.8	0.011	0.010	0.005	0.019	0.007	0.014	0.008	0.023	0.010
0.8	0.069	0.035	0.045	0.034	0.045	0.049	0.038	0.048	0.048
0.8	0.037	0.019	0.021	0.018	0.021	0.029	0.023	0.028	0.028
0.8	0.020	0.012	0.010	0.010	0.011	0.019	0.014	0.018	0.017
0.8	0.010	0.007	0.005	0.006	0.006	0.012	0.008	0.010	0.010
1.8	0.095	0.623	0.062	0.224	0.092	0.616	0.048	0.228	0.054
1.4	0.047	0.611	0.028	0.206	0.043	0.593	0.028	0.202	0.030
1.2	0.026	0.607	0.013	0.198	0.021	0.586	0.017	0.189	0.018
1.2	0.012	0.605	0.006	0.197	0.011	0.587	0.009	0.187	0.010
1.4	0.104	0.304	0.068	0.153	0.124	0.312	0.054	0.173	0.107
1.2	0.053	0.292	0.030	0.135	0.081	0.288	0.032	0.145	0.077
1.2	0.029	0.288	0.015	0.127	0.065	0.277	0.019	0.128	0.063
1.1	0.014	0.285	0.007	0.123	0.057	0.275	0.010	0.123	0.056
1.2	0.109	0.060	0.070	0.056	0.077	0.085	0.053	0.076	0.077
1.1	0.053	0.038	0.031	0.031	0.036	0.054	0.031	0.045	0.043
1.1	0.028	0.030	0.015	0.021	0.021	0.039	0.018	0.030	0.028
1.1	0.014	0.025	0.007	0.016	0.014	0.029	0.010	0.020	0.018
2.1	0.115	0.664	0.084	0.229	0.117	0.658	0.053	0.255	0.057
1.7	0.057	0.654	0.035	0.220	0.053	0.629	0.032	0.217	0.033
1.5	0.028	0.649	0.016	0.215	0.025	0.624	0.018	0.206	0.018
1.4	0.013	0.648	0.007	0.213	0.012	0.627	0.009	0.202	0.011
1.7	0.145	0.448	0.102	0.211	0.176	0.452	0.066	0.226	0.118
1.5	0.072	0.430	0.043	0.186	0.096	0.419	0.040	0.194	0.081
1.4	0.036	0.425	0.020	0.173	0.066	0.406	0.022	0.172	0.060
1.4	0.017	0.423	0.009	0.167	0.052	0.406	0.012	0.166	0.050
1.4	0.146	0.094	0.103	0.080	0.116	0.128	0.066	0.103	0.102
1.4	0.071	0.068	0.043	0.048	0.053	0.084	0.038	0.065	0.058
1.3	0.037	0.057	0.021	0.036	0.031	0.064	0.022	0.045	0.038
1.3	0.017	0.050	0.009	0.028	0.020	0.053	0.012	0.033	0.024

probability of progression from the reintroduction state of 0.1%, 5.1%, 0.0%, and 0.1% for the ESD, ESPD, UDR, and MDR approaches, respectively. With regard to the relative entropy, the ESPD and MDR approaches generally outperformed the ESD and UDR approaches. Comparing the ESPD and MDR approaches, they both alternately outperformed the other, making it difficult to state which is the best-performing approach overall.

Table 3. *Mean accuracy of the approaches in terms of relative incidence difference and relative entropy over all probabilistic sensitivity analysis runs for the case study.*

Subgroup Number	Treatment Response	Stage of Disease	Treatment Strategy[1]	Subgroup Size	Incidence of Events[2] (Relative Difference (%))								
					Post-Induction: Progression				Reintroduction: Progression				
					ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
0	-	-	0	279	1.9	0.0	0.9	0.4	-1.7	-1.6	-1.6	-1.6
			1	279	1.4	0.0	0.2	0.3	-3.2	-3.0	-3.1	-3.0
1	SD	-	0	95	-	-	-	-	-1.9	-1.7	-1.7	-1.6
			1	96	0.2	0.0	0.1	0.1	-3.4	-3.4	-3.5	-3.4
2	CR/PR	-	0	184	2.6	0.0	0.4	0.4	-1.6	-1.5	-1.5	-1.5
			1	183	2.0	0.0	0.0	0.4	-3.0	-2.9	-2.9	-2.8
3	-	Synch.	0	191	1.1	0.1	0.3	0.3	-1.4	-1.2	-1.2	-1.2
			1	219	1.3	0.0	0.2	0.3	-2.9	-2.8	-2.8	-2.8
4	-	Meta.	0	88	4.2	0.2	2.9	0.3	-2.6	-2.6	-2.7	-2.6
			1	59	1.6	0.1	0.2	0.1	-2.1	-2.6	-2.6	-2.6
5	SD	Synch.	0	67	-	-	-	-	-1.6	-1.4	-1.3	-1.2
			1	74	0.3	0.1	0.1	0.1	-3.1	-3.0	-3.1	-3.0
6	SD	Meta.	0	28	-	-	-	-	-2.6	-2.6	-4.5	-3.4
			1	22	0.1	5.1	0.0	0.1	1.5	2.3	-1.0	0.3
7	CR/PR	Synch.	0	124	1.5	0.1	0.2	0.4	-1.2	-1.1	-1.1	-1.1
			1	145	1.9	1.9	0.0	0.0	0.4	-2.8	-2.7	-2.7
8	CR/PR	Meta.	0	61	5.7	0.3	3.1	0.3	-2.5	-2.6	-2.8	-2.5
			1	37	2.5	0.2	0.4	0.3	-2.5	-1.9	-3.1	-3.0

[1]Treatment Strategy: 0) observation strategy, 1) maintenance treatment. SD, stable disease; CR/PR, complete or partial response; Meta., metachronous; Synch., synchronous.

[2]A hyphen (-) indicates that the corresponding probabilities were either zero or one, i.e. one of the competing events was not observed, making a comparison of the relative incidence irrelevant.

Table 3 (continued). *Mean accuracy of the approaches in terms of relative incidence difference and relative entropy over all probabilistic sensitivity analysis runs for the case study.*

Subgroup Number	Time-to-Event Distributions[3] (Kullback-Leibler Divergence)															
	Post-Induction: Progression				Post-Induction: Death				Reintroduction: Progression				Reintroduction: Death			
	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
0	0.075	0.017	0.066	0.017	0.815	0.112	0.694	0.087	0.027	0.022	0.024	0.018	0.058	0.047	0.068	0.074
	0.012	0.012	0.011	0.012	0.110	0.040	0.048	0.039	0.075	0.073	0.063	0.056	0.060	0.027	0.043	0.056
1	0.023	0.023	0.023	0.023	-	-	-	-	0.024	0.033	0.020	0.021	0.150	0.094	0.154	0.172
	0.027	0.026	0.027	0.027	0.365	0.252	0.366	0.355	0.035	0.057	0.025	0.028	0.155	0.106	0.142	0.147
2	0.097	0.022	0.083	0.022	0.692	0.108	0.628	0.085	0.048	0.037	0.043	0.035	0.110	0.112	0.121	0.116
	0.018	0.014	0.012	0.014	0.141	0.071	0.085	0.069	0.091	0.086	0.080	0.072	0.084	0.051	0.056	0.069
3	0.075	0.022	0.067	0.022	0.732	0.170	0.550	0.246	0.026	0.023	0.024	0.021	0.087	0.069	0.098	0.101
	0.012	0.009	0.009	0.009	0.127	0.038	0.051	0.039	0.047	0.031	0.038	0.027	0.059	0.029	0.047	0.061
4	0.075	0.017	0.065	0.017	0.321	0.145	1.227	0.145	0.042	0.048	0.036	0.025	0.104	0.084	0.091	0.099
	0.050	0.052	0.052	0.051	0.216	0.155	0.209	0.166	0.183	0.181	0.187	0.191	0.198	0.112	0.180	0.158
5	0.028	0.028	0.028	0.028	-	-	-	-	0.025	0.034	0.024	0.024	0.319	0.219	0.317	0.358
	0.021	0.021	0.021	0.021	0.332	0.210	0.306	0.292	0.033	0.050	0.025	0.023	0.237	0.197	0.231	0.242
6	0.059	0.059	0.059	0.059	-	-	-	-	0.131	0.124	0.106	0.115	0.220	0.200	0.242	0.200
	0.079	0.075	0.078	0.077	-	-	-	-	0.125	0.149	0.150	0.160	0.280	0.199	0.256	0.214
7	0.102	0.029	0.090	0.030	0.620	0.160	0.472	0.225	0.049	0.048	0.049	0.047	0.162	0.157	0.170	0.166
	0.017	0.015	0.012	0.015	0.147	0.061	0.080	0.059	0.068	0.047	0.058	0.045	0.089	0.056	0.061	0.076
8	0.098	0.026	0.081	0.026	0.227	0.143	1.203	0.143	0.053	0.049	0.044	0.027	0.149	0.092	0.116	0.128
	0.066	0.057	0.063	0.057	0.239	0.182	0.255	0.185	0.223	0.221	0.231	0.233	0.232	0.141	0.214	0.200

[3] *A hyphen (-) indicates that insufficient observations were available in de original trial data to calculate the Kullback-Leibler divergence.*

The extent to which health economic modeling outcomes are affected by differences in performance is illustrated by cost-effectiveness planes for selected subgroup analyses in Figure 4 and in cost-effectiveness planes and cost-effectiveness acceptability curves for all subgroup analyses in Appendix 5 and Appendix 6, respectively. The cost-effectiveness planes show that cost-effectiveness point estimates are similar for large sample sizes (e.g., n = 300 or larger), illustrated by the overlapping points representing the incremental cost-effectiveness estimate for the different approaches. For analysis Subgroup 0 (i.e., based on the complete patient cohort), incremental cost-effectiveness ratios are €177,317, €166,997, €172,811, and €169,024 per quality-adjusted life year (QALY) gained, which translates to a net monetary benefit of –€26,601, –€26,656, –€26,579, and –€26,503 at a willingness to pay of €20,000 per QALY for the ESD, ESPD, UDR, and MDR approaches, respectively. However, the point estimates of the incremental costs and effects show small differences between approaches for smaller sample sizes, which is illustrated by the results for Subgroup 4 in Figure 4, for example. For this subgroup analysis, incremental cost-effectiveness ratios are €494,517, €342,726, €420,222, and €322,343 per QALY gained, which translates to a net monetary benefit of –€29,854, –€29,966, –€29,195, and –€29,939 at a willingness to pay of €20,000 per QALY for the ESD, ESPD, UDR, and MDR approaches, respectively. With regard to the uncertainty surrounding the point estimates, represented by the confidence ellipses, small differences are observed for certain subgroup analyses (e.g., Subgroup 4). In agreement, the cost-effectiveness acceptability curves show modest differences between approaches (Appendix 6), of which the magnitude increases when sample sizes decrease.

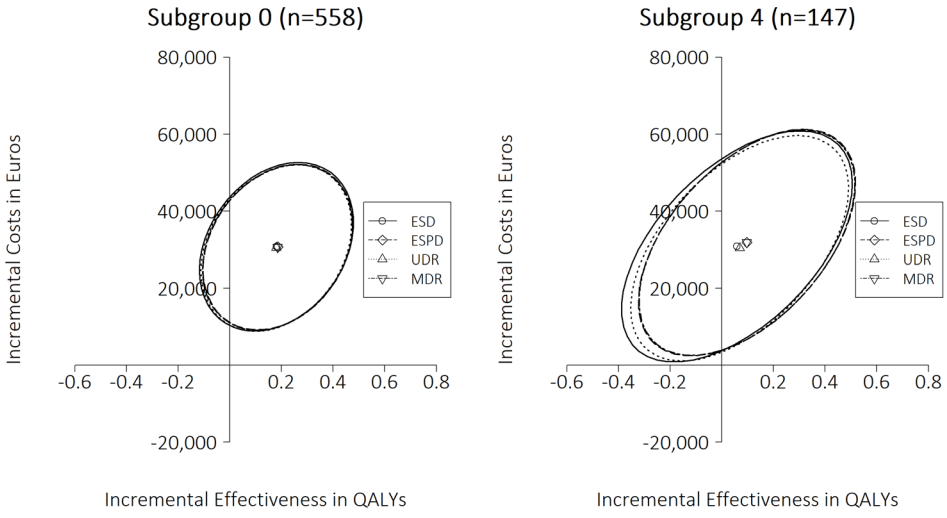


Figure 4. *Cost-effectiveness planes based on the probabilistic sensitivity analysis for selected subgroup analyses. ESD, event-specific distribution; ESPD, event-specific probability and distribution; MDR, multimodal joint distribution and regression model; UDR, unimodal joint distribution and regression model.*

DISCUSSION

This article provides a thorough quantitative comparison of strategies available for implementing competing risks in DES models informed by uncensored IPD and, thereby, contributes to the general, already existing guidance [6]. According to these general modeling good research practices guidelines, the preferred modeling approach is based on the strategy of selecting the time-to-event first, from a joint time-to-event distribution, and then to select the corresponding event, which corresponds to the UDR and MDR approaches illustrated in this article. Although the results for the UDR approach question this recommendation, the results for the MDR approach, which attempts to reflect the true nature of the data by allowing for multimodal joint distributions, support this recommendation. Interestingly, the results also show that the ESPD approach, which might be considered more straightforward for the novice, provides an attractive alternative to the rather complex MDR approach.

The superior performance of the ESPD and MDR approaches can be explained by the assumptions underlying the ESD and UDR approaches. For the ESD approach, observations of competing events are considered censored observations when estimating the event-specific time-to-event distributions. Given that the analysis of censored time-to-event data is more complex compared to uncensored data, the resulting increase in uncertainty is likely to negatively affect performance. The main assumption for the UDR approach is that the joint time-to-event data, representing all competing risks, is unimodal distributed. Consequently, the (negative) impact of this assumption is illustrated in the simulation study, as the performance of the UDR approach deteriorates stronger compared to the performance of the ESPD and MDR approaches when overlap between time-to-event distributions decreases.

Based on the performance in the simulation study and internal validity in the case study, the ESPD and MDR approaches are expected to yield more accurate health economic outcomes. However, relatively small differences in cost-effectiveness point estimates and magnitude of the uncertainty surrounding these estimates between approaches are observed for the case study, which can be explained by the characteristics of the CAIRO3 data. First of all, the maximum number of competing risks for a specific state in the CAIRO3 model is 2 (Figure 1b). For health economic models in which the number of competing risks is higher, more substantial differences in health economic outcomes are expected, as the simulation study shows that differences in the approaches' performance increase according to higher numbers of competing risks. Second, the CAIRO3 study represents a relatively large sample size (i.e., n = 558), which contributes to the overall performance of the approaches. Third, time-to-event distributions for the case study substantially overlap (e.g., 81% for the reintroduction state). Larger differences in health economic outcomes are expected for models informed by IPD comprising lower overlap between competing time-to-event distributions. Finally, health economic models defined by more health states may yield more substantial differences in health economic outcomes due to a cumulative effect of subsequent competing risks including states.

Based on the performance in the simulation study, either the ESDP or MDR approach would be recommended for implementing competing risks in DES models. However, the case study

illustrates that the ESPD approach might be more sensitive to low event rates in practice. Contrarily, if it is infeasible to apply the MDR approach (e.g., due to difficulties with regard to estimating a joint multimodal distribution that appropriately represents the observed IPD), the ESPD approach is a legitimate alternative. Regarding the complexity of implementing the approaches, the flexible multimodal distributions and multinomial regression models required for the MDR approach are more difficult to estimate and implement compared to the event-specific probabilities and distributions of the ESPD approach. Consequently, the ESPD approach would be preferred over its MDR equivalent if the IPD comprises sufficient observations for each competing event. However, a general definition of ''sufficient observations'' cannot be stated and modelers should, therefore, ensure that applying the ESPD approach is appropriate by internal validation. If the internal validity of the ESPD approach indicates its use might be inappropriate (e.g., due to substantial differences in event incidence), both the ESPD and MDR approaches should be applied and their internal validity compared to select the best-performing modeling approach. If time does not allow for such a comparison, modelers are advised to apply the MDR approach. Except if there are evident reasons for using the ESD or UDR approach, the use of either of these approaches is advised against.

All methods presented in and findings of this study are applicable to data sets in which a competing event is observed for each patient (i.e., the data are uncensored). Although clinical trial data may often be right censored due to limited time horizons, certain clinical trials do capture all (competing) events of interest for each patient. Examples of clinical contexts for which uncensored clinical trials are generally feasible include nausea and vomiting after surgery and metastatic cancers, as was illustrated in the current case study. In addition, retrospective data sets (e.g., from registries) are potentially uncensored for specific cohorts of patients. Performance of the modeling approaches may be different for censored IPD (e.g., because all modeling approaches will include survival analysis of censored data), whereas this only applies to the ESD approach for uncensored IPD or because a different implementation of the ESPD approach is required to calculate event-specific probabilities if a competing event is not observed for each patient. Consequently, the current focus on uncensored IPD is relevant due to a lack of interchangeability of the modeling approaches' implementation and performance for censored and uncensored IPD. Furthermore, this highlights a need for comparing modeling approaches toward implementing competing risks in DES models informed by censored IPD, which is part of a subsequent study.

Besides general limitations relating to the external validity of simulation studies and single-case studies, this study has certain additional limitations. Findings and conclusions presented in this article apply to DES modeling studies informed by uncensored IPD and for which a decision on the approach taken to implement competing risks needs to be made. If aggregated data (e.g., from literature) are used to populate a model, this evidence is structured according to a specific approach and can only be implemented accordingly. In addition, Weibull distributions were used in the simulation study to simulate and represent patient-level time-to-event variation, which allowed unbiased comparison of the modeling approaches. On the other hand, this design choice limits the generalizability of the results, as in practice, underlying distributions are unknown and performance of the approaches may vary, depend-

ing on their flexibility to describe the actually observed data. Using different types of distributions, such as Gamma or lognormal distributions, therefore, may affect the performance of the modeling approaches. Finally, mixture models are used to reflect the multimodality induced by competing underlying time-to-event distributions in the joint distributions for the MDR approach, whereas alternative strategies are available, such as phase-type distributions.

CONCLUSION

Substantial differences were observed in the accuracy of the competing risks modeling approaches in terms of event incidence and time-to-event distributions. The performance of the approaches is strongly conditional upon various data and disease pathway characteristics, such as the numbers of competing risks and overlap between competing time-to-event distributions. Performance dissimilarities between approaches, however, did not result in substantially different health economic outcomes in this specific case study. The recommended modeling approach for implementing competing risks in DES models informed by uncensored IPD is the MDR approach, which is based on the strategy of selecting the time-to-event first, based on a joint multimodal time-to-event distribution, and then selecting the corresponding event using probabilities obtained from a (multinomial) logistic regression model. However, if sufficient observations of all competing events are available, or if use of the MDR approach is infeasible, use of the less complex ESPD approach, which is based on the strategy of selecting the event to occur first and the corresponding time-to-event second, is also appropriate.

SUPPLEMENTARY MATERIALS

Supplementary material related to this article can be found in the online version at:
https://doi.org/10.1177/0272989X18814770

REFERENCES

1. Caro JJ, et al. Discrete event simulation: the preferred technique for health economic evaluations? Value Health. 2010;13(8):1056–60.

2. Karnon J, Haji Ali Afzali H. When to use discrete event simulation (DES) for the economic evaluation of health technologies? A review and critique of the costs and benefits of DES. Pharmacoeconomics. 2014;32(6):547–58.

3. Standfield L, et al. Markov modeling and discrete event simulation in health care: a systematic comparison. Int J Technol Assess Health Care. 2014;30(2): 165–72.

4. Siebert U, et al. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–3. Med Decis Making. 2012;32(5):690–700.

5. Degeling K, et al. Accounting for parameter uncertainty in the definition of parametric distributions used to describe individual patient variation in health economic models. BMC Med Res Methodol. 2017;17(1):170.

6. Karnon J, et al. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–4. Med Decis Making. 2012;32(5):701–11.

7. Caro JJ, Moller J. Decision-analytic models: current methodological challenges. Pharmacoeconomics. 2014;32(10): 943–50.

8. Barton P, et al. The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis. Health Technol Assess. 2004;8(11):104.

9. Jiang Y, Ni W. Expected lifetime numbers, risks, and burden of osteoporotic fractures for 50-year old Chinese women: a discrete event simulation incorporating FRAX. J Bone Miner Metab. 2016;34(6):714–22.

10. Erenay FS, et al. Cost-effectiveness of alternative colonoscopy surveillance strategies to mitigate metachronous colorectal cancer incidence. Cancer. 2016;122(16):2560–70.

11. Wang HI, et al. Treatment cost and life expectancy of diffuse large B-cell lymphoma (DLBCL): a discrete event simulation model on a UK population-based observational cohort. Eur J Health Econ. 2017;18(2):255–67.

12. Igarashi A, et al. Cost-effectiveness analysis of smoking cessation interventions in Japan using a discrete-event simulation. Appl Health Econ Health Policy. 2016;14(1):77–87.

13. Davis S, et al. A systematic review and economic evaluation of bisphosphonates for the prevention of fragility fractures. Health Technol Assess. 2016;20(78):1–406.

14. Standfield LB, et al. An empirical comparison of Markov cohort modeling and discrete event simulation in a capacity-constrained health care setting. Eur J Health Econ. 2017;18(1):33–47.

15. Thanh NX, et al. Cost effectiveness of eplerenone for the treatment of systolic heart failure with mild symptoms in Alberta, Canada. Am J Cardiovasc Drugs. 2016;16(5):365–76.

16. Bessen T, et al. Does one size fit all? Cost-utility analyses of alternative mammographic follow-up schedules, by risk of recurrence. Int J Technol Assess Health Care. 2015;31(5):281–8.

17. Crossan C, et al. Cost effectiveness of case-finding strategies for primary prevention of cardiovascular disease: a modelling study. Br J Gen Pract. 2017;67(654):e67–e77.

18. Simkens LHJ, et al. Maintenance treatment with capecitabine and bevacizumab in metastatic colorectal cancer (CAIRO3): a phase 3 randomised controlled trial of the Dutch Colorectal Cancer Group. Lancet. 2015;385(9980):1843–52.

19. R Core Team. R: a language and environment for statistical computing. Version 3.3.2 [software]. Available from: https://www.r-project.org/

20. Pintilie M. Competing Risks: A Practical Perspective. Hoboken, NJ: Wiley; 2006.

21. Austin PC, et al. Introduction to the analysis of survival data in the presence of competing risks. Circulation. 2016;133(6):601–9.

22. Law AM. Simulation Modeling and Analysis. 4th ed. Singapore: McGraw-Hill Higher Education; 2007.

23. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. Hoboken, NJ: John Wiley; 2000.

24. Altiok T. On the phase-type approximations of general distributions. IIE Trans. 1985;17(2):110–6.

25. Bordes L, Chauveau D. Stochastic EM algorithms for parametric and semiparametric mixture models for right-censored lifetime data. Comput Stat. 2016;31(4):1513–38.

26. Cover TM, Thomas JA. Elements of Information Theory. 2nd ed. Hoboken, NJ: John Wiley; 2006.

27. Delignette-Muller ML, Dutang C. Fitdistrplus: an R package for fitting distributions. J Stat Softw. 2015;64(4):1–34.

28. Benaglia T, et al. mixtools: an R package for analyzing finite mixture models. J Stat Softw. 2009;32(6):1–29.

29. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York, NY: Springer; 2002.

30. Grün B, Leisch F. Fitting finite mixtures of generalized linear regressions in R. Comput Stat Data Anal. 2007;51(11): 5247–52.

31. Grün B, Leisch F. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J Stat Softw. 2008;28(4):35.

32. Leisch F. FlexMix: a general framework for finite mixture models and latent class regression in R. J Stat Softw. 2004;11(8):18.

33. Franken MD, et al. Cost-effectiveness of capecitabine and bevacizumab maintenance treatment after first-line induction treatment in metastatic colorectal cancer. Eur J Cancer. 2017;75:204–12.

APPENDIX 1: exact definitions of the populations used in the simulation study

	Number of Competing Risks	Distribution Overlap (%)	Event 1			Event 2			Event 3			Event 4		
			Incidence	Shape	Scale	Incidence	Shape	Scale	Incidence	Shape	Scale	Incidence	Shape	Scale
Population 1	2	10	0.61	5	250	0.39	10	418	-	-	-	-	-	-
Population 2	2	50	0.61	5	250	0.39	10	302	-	-	-	-	-	-
Population 3	2	90	0.61	5	250	0.39	5	265	-	-	-	-	-	-
Population 4	3	10	0.41	5	250	0.25	10	418	0.34	4.5	950	-	-	-
Population 5	3	50	0.41	5	250	0.25	10	302	0.34	4.5	409	-	-	-
Population 6	3	90	0.41	5	250	0.25	5	265	0.34	5	291	-	-	-
Population 7	4	10	0.41	5	250	0.20	10	418	0.29	4.5	950	0.10	7	2070
Population 8	4	50	0.41	5	250	0.20	10	302	0.29	4.5	409	0.10	7	637
Population 9	4	90	0.41	5	250	0.20	5	265	0.29	5	291	0.10	5	337

APPENDIX 2: calculation of the overlap between distributions

No gold standard approach for calculating the overlap between multiple (competing) distributions is available. The approach used in the current study defines the overlap between distributions as the mean overlap percentage over all distribution-specific overlaps, for which the distribution-specific overlap is defined as the percentage of the distribution that is overlapping with any other distribution. More precisely, the calculation can be performed according to the following steps:

1. Define a pre-specified number of equally sized bins that cover the observations of all distributions
2. Calculate the density of all distributions for each bin (1)
3. For each distribution, calculate the "overlap density" in each of the bins (1) by determining how much the density (2) overlaps with the corresponding highest density of the other distributions
4. Calculate the distribution-specific overlap by dividing the area under the "overlap density" curve (3) by the area under overall density curve (2)
5. (Calculate the overall overlap between all distributions as the average of the distribution-specific overlaps (4)

R code used to calculate the overlap between competing time-to-event distributions is provided below.

```
# This function calculates the overlap between competing
# time-to-event distributions based on a 'data' object
# containing two columns, i.e. 'time' and 'event'.

getOverlap <- function(data, n.bins=512, digits=3) {

 # Function inputs:
 # - data:     data.frame with two columns:
 #          + time:     column containing the times
 #          + event:    column containing the events corresponding to the times
 # - n.bins:   the number of bins used for the 'density' function
 # - digits:   the number of digits to round the results to

 # Load package for calculating the Area Under the Curve (AUC)
 library(flux);

 # Get the events in the data
 events <- sort(unique(data$event));
 n.events <- length(events);

 # Get event-specific times
 times <- lapply(events, function(e) {
   subset(data, event==e)$time;
```

```
});

# Density settings
min <- min(data$time);
max <- max(data$time);
n <- n.bins;
x <- seq(from=min, to=max, length.out=n);

# Get event-specific densities
y <- sapply(1:n.events, function(i1) {
  density(times[[i1]], from=min, to=max, n=n)$y;
});

# Get event-specific density overlap
yo <- sapply(1:n.events, function(i2) {

  sapply(1:n, function(j){

    own <- y[j,i2];
    other <- max(y[j,which(c(1:n.events)!=i2)]);

    if(own<=0) {
      return(0)
    } else
      return(min(own,other))

  })
});

# Calculate event-specific overlap
o <- sapply(1:n.events, function(i3) {
  (auc(x=x, y=yo[,i3]) / auc(x=x, y=y[,i3])) * 100;
});

return(round(mean(o), digits=digits));
}
```

APPENDIX 3: subgroup definitions including sample sizes and event incidences

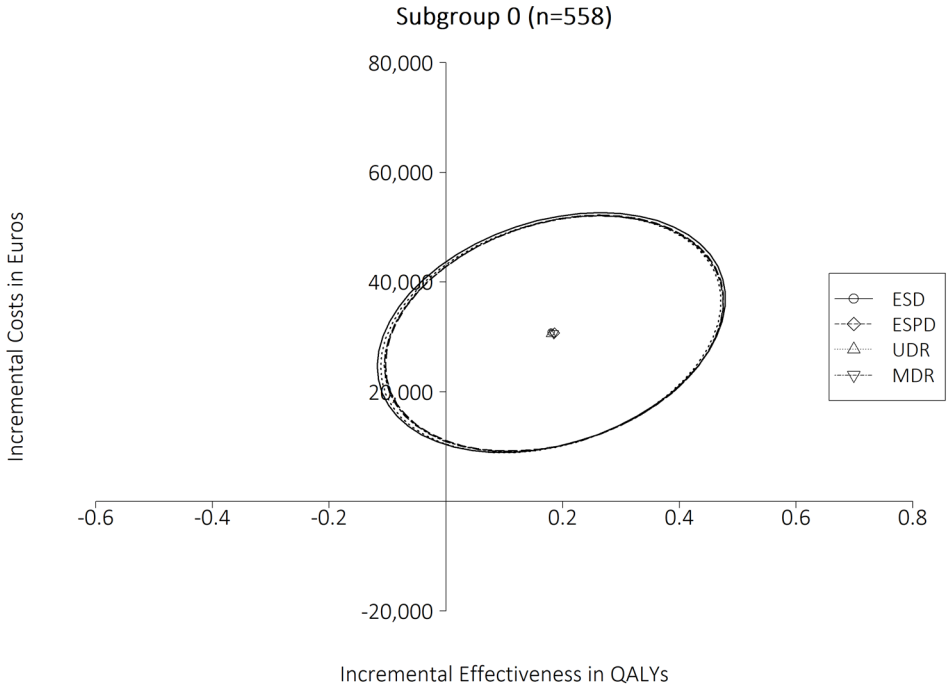
Subgroup	Sample Size	Event Observations				
		Progression *(Post-induction)*	Death *(Post-induction)*	Progression *(Re-induction)*	Death *(Re-induction)*	Death *(salvage)*
Subgroup 0: No Subgroups	558	519	39	370	109	410
Observation	279	270	9	200	54	216
Maintenance	279	249	30	170	55	194
Subgroup 1: SD	191	185	6	128	43	142
Observation	95	95	0	68	22	73
Maintenance	96	90	6	60	21	69
Subgroup 2: CR/PR	367	334	33	242	66	268
Observation	184	175	9	132	32	143
Maintenance	183	159	24	110	34	125
Subgroup 3: Synchronous	410	382	28	275	82	300
Observation	191	186	5	143	33	153
Maintenance	219	196	23	132	49	147
Subgroup 4: Metachronous	147	137	10	95	27	110
Observation	88	84	4	57	21	63
Maintenance	59	53	6	38	6	47
Subgroup 5: SD & Synchronous	141	136	5	94	34	102
Observation	67	67	0	49	15	52
Maintenance	74	69	5	45	19	50
Subgroup 6: SD & Metachronous	50	49	1	34	9	40
Observation	28	28	0	19	7	21
Maintenance	22	21	1	15	2	19
Subgroup 7: CR/PR & Synchronous	269	246	23	181	48	198
Observation	124	119	5	94	18	101
Maintenance	145	127	18	87	30	97
Subgroup 8: CR/PR & Metachronous	97	88	9	61	18	70
Observation	60	56	4	38	14	42
Maintenance	37	32	5	23	4	28

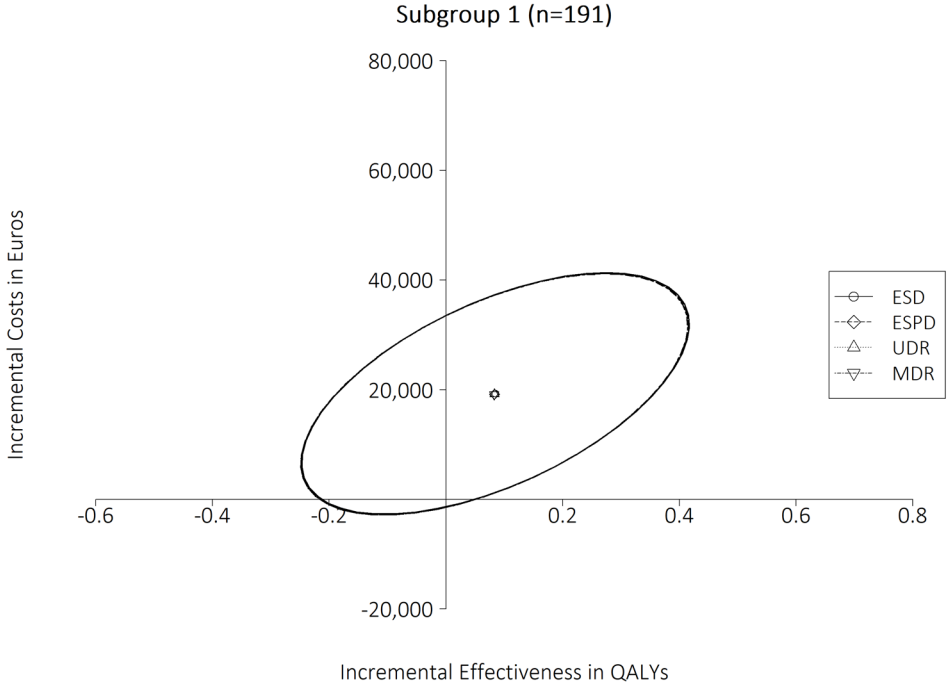
APPENDIX 4: relative absolute incidence difference for the case study

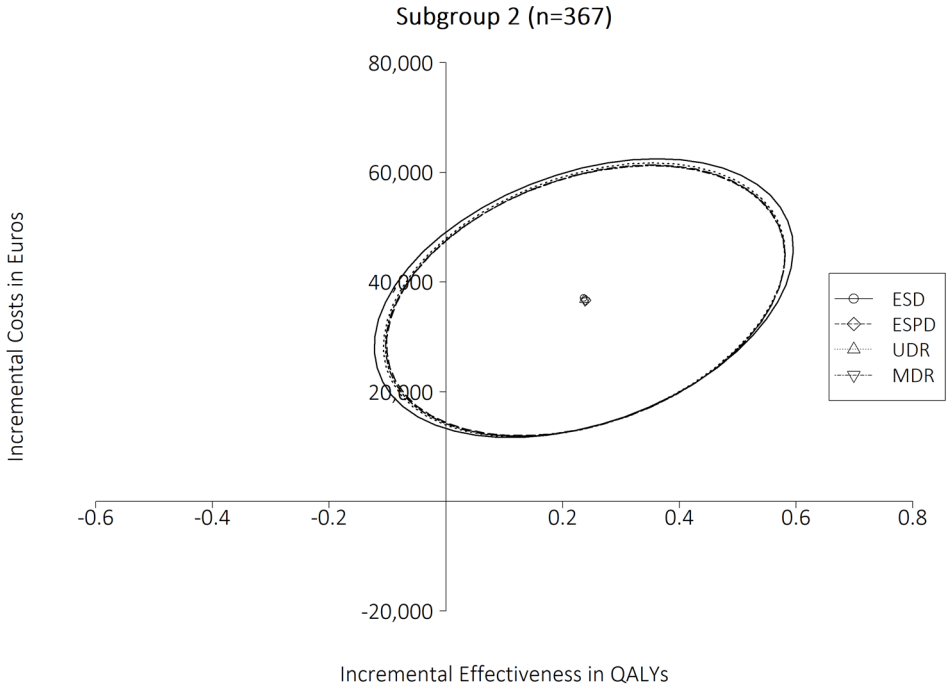
*Treatment Strategy: 0) control (observation strategy), 1) intervention (maintenance treatment). SD = stable disease, CR/PR = complete or partial response, Meta. = metachronous, Synch. = synchronous. †A hyphen (-) indicates that the corresponding probabilities were either zero or one, i.e. one of the competing events was not observed, making a comparison of the relative absolute incidence irrelevant.

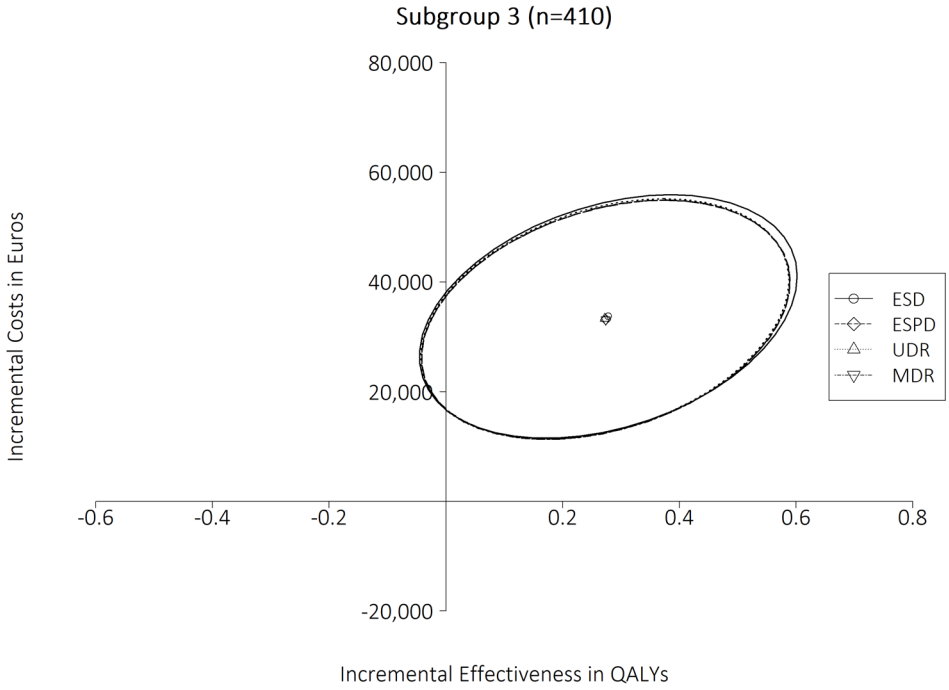
Subgroup Number	Treatment Response	Stage of Disease	Treatment Strategy*	Subgroup Size	Incidence of Events† Relative Absolute Difference (%)							
					Post-Induction: Progression				Reintroduction: Progression			
					ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
0	-	-	0	279	1.9	0.1	1.1	0.4	1.7	1.6	1.6	1.6
			1	279	1.4	0.3	0.4	0.4	3.2	3.0	3.1	3.0
1	SD	-	0	95	-	-	-	-	1.9	1.7	1.7	1.6
			1	96	0.3	0.2	0.2	0.2	3.4	3.4	3.5	3.4
2	CR/PR	-	0	184	2.6	0.2	1.2	0.5	1.6	1.5	1.5	1.5
			1	183	2.0	0.3	0.5	0.5	3.0	2.9	2.9	2.8
3	-	Synch.	0	191	1.1	0.2	0.7	0.3	1.4	1.2	1.2	1.2
			1	219	1.3	0.3	0.4	0.4	2.9	2.8	2.8	2.8
4	-	Meta.	0	88	4.2	0.4	2.9	0.4	2.6	2.6	2.7	2.6
			1	59	1.7	0.3	0.7	0.5	2.1	2.7	2.6	2.6
5	SD	Synch.	0	67	-	-	-	-	1.7	1.4	1.3	1.3
			1	74	0.4	0.3	0.3	0.3	3.1	3.0	3.1	3.0
6	SD	Meta.	0	28	-	-	-	-	2.7	2.8	4.5	3.4
			1	22	0.2	5.1	0.2	0.2	3.8	5.6	2.1	3.8
7	CR/PR	Synch.	0	124	1.5	0.2	1.0	0.5	1.2	1.1	1.1	1.1
			1	145	1.9	0.3	0.5	0.5	2.8	2.7	2.7	2.6
8	CR/PR	Meta.	0	61	5.7	0.5	3.1	0.5	2.5	2.7	2.8	2.5
			1	37	2.6	0.5	1.3	0.9	2.6	3.3	3.1	3.1

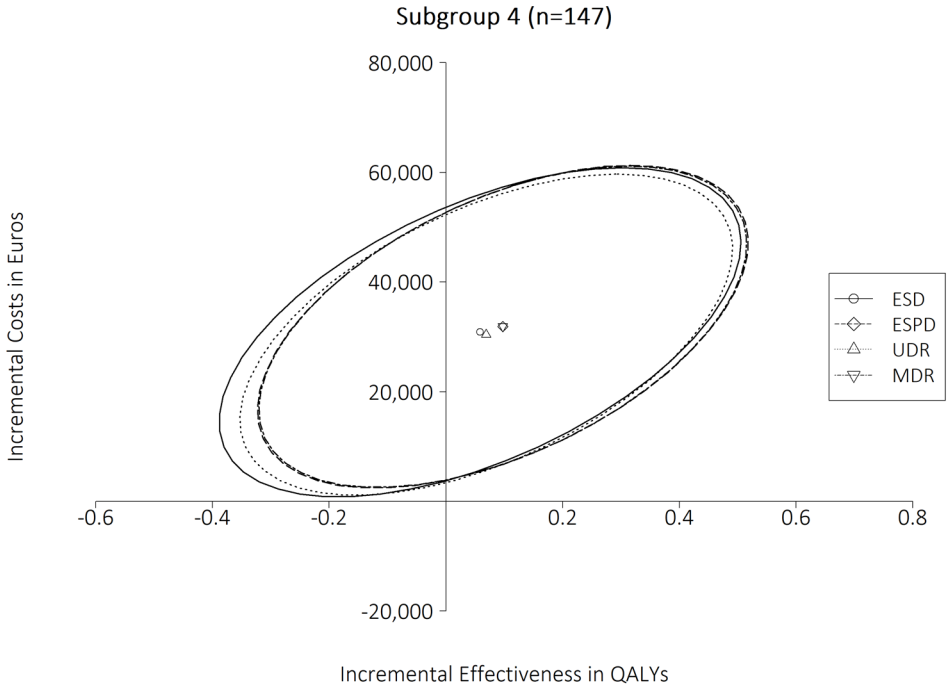
APPENDIX 5: incremental cost-effectiveness planes for all case study subgroup analyses

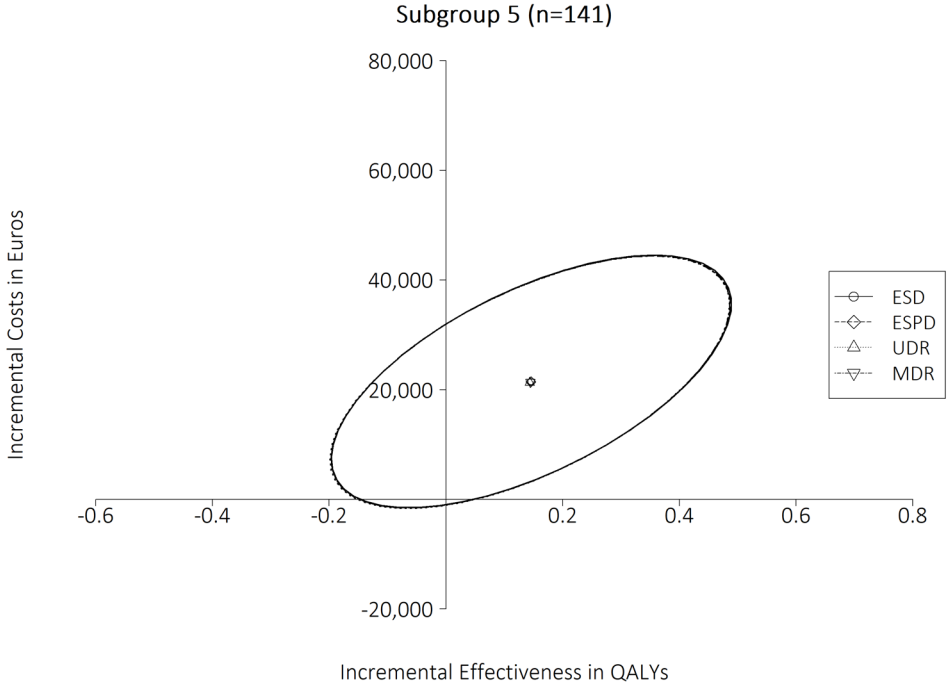


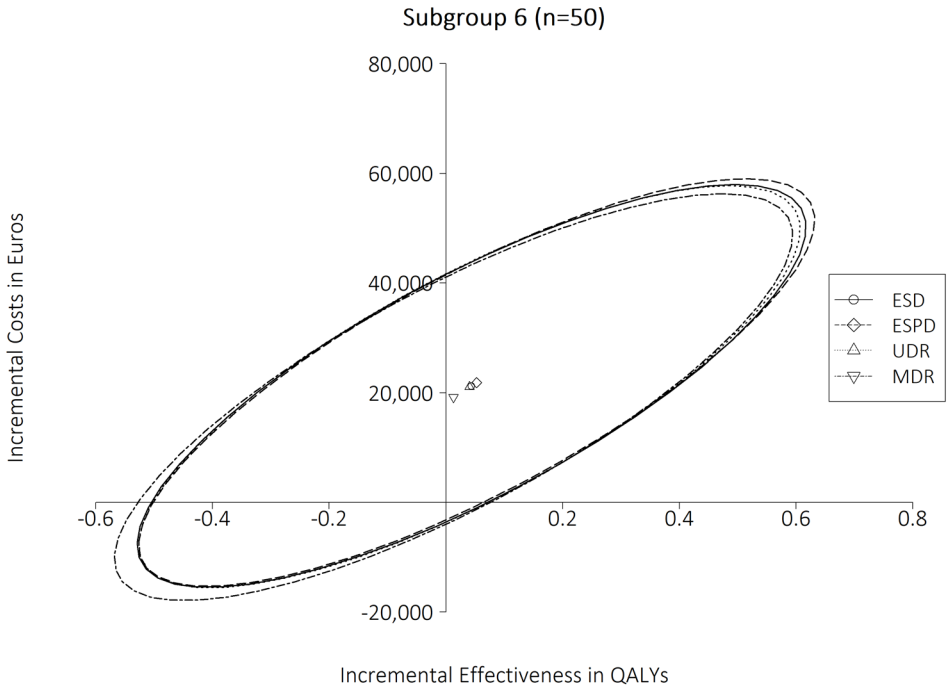


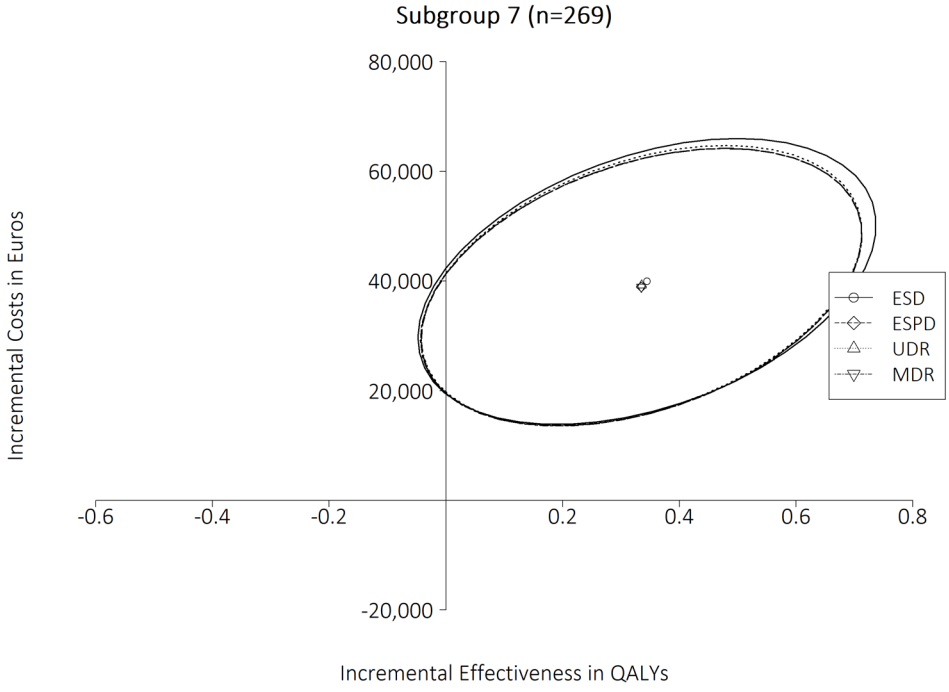

Subgroup 2 (n=367)
80,000
60,000
40,000
20,000
-20,000
-0.6
-0.4
-0.2
0.2
0.4
0.6
0.8
Incremental Costs in Euros
Incremental Effectiveness in QALYs
ESD
ESPD
UDR
MDR

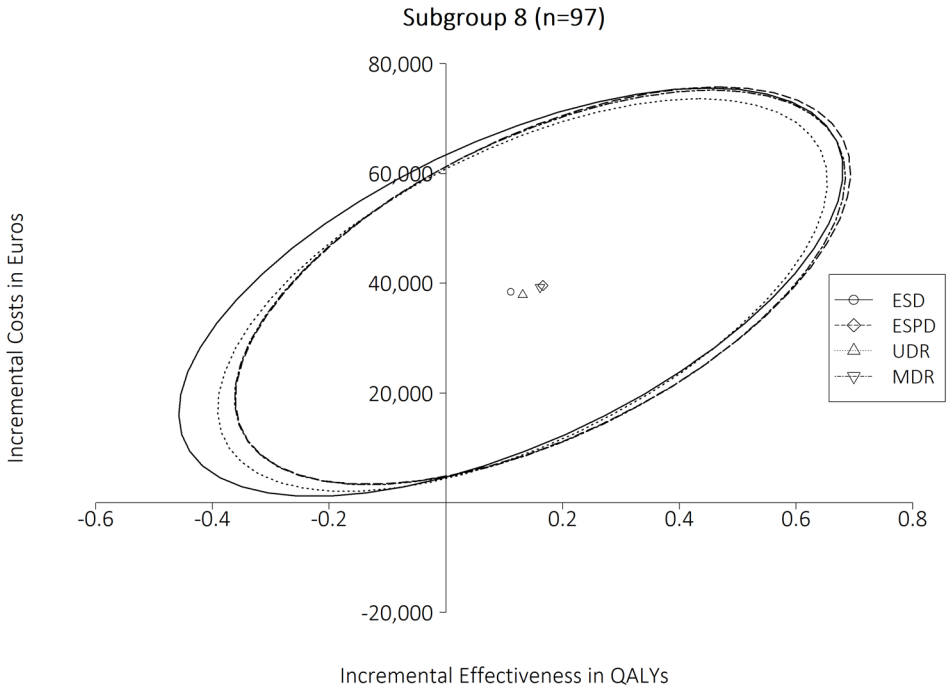

Subgroup 3 (n=410)
80,000
60,000
40,000
20,000
-20,000
-0.6
-0.4
-0.2
0.2
0.4
0.6
0.8
Incremental Costs in Euros
Incremental Effectiveness in QALYs
ESD
ESPD
UDR
MDR


Subgroup 4 (n=147)
80,000
60,000
40,000
20,000
-20,000
-0.6
-0.4
-0.2
0.2
0.4
0.6
0.8
Incremental Costs in Euros
Incremental Effectiveness in QALYs
ESD
ESPD
UDR
MDR

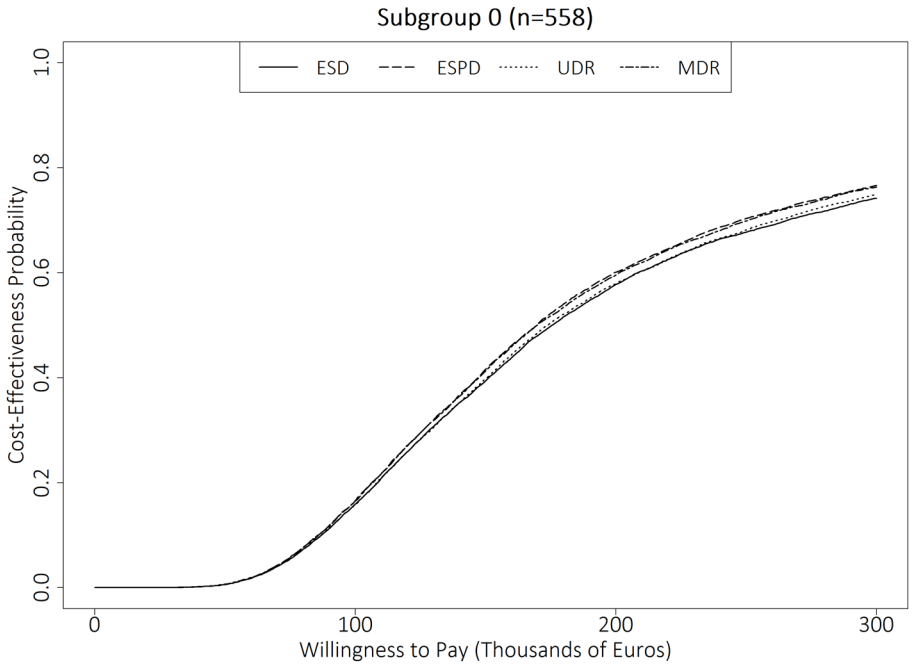

Subgroup 5 (n=141)
80,000
60,000
40,000
20,000
-20,000
-0.6
-0.4
-0.2
0.2
0.4
0.6
0.8
Incremental Costs in Euros
Incremental Effectiveness in QALYs
ESD
ESPD
UDR
MDR

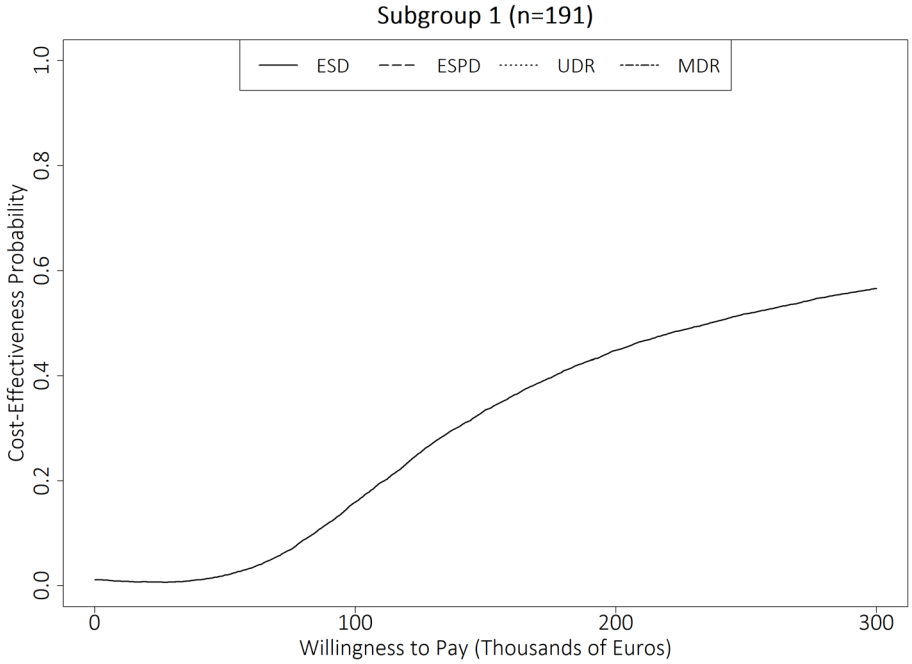

Subgroup 6 (n=50)
80,000
60,000
40,000
20,000
-20,000
-0.6
-0.4
-0.2
0.2
0.4
0.6
0.8
Incremental Costs in Euros
Incremental Effectiveness in QALYs
ESD
ESPD
UDR
MDR

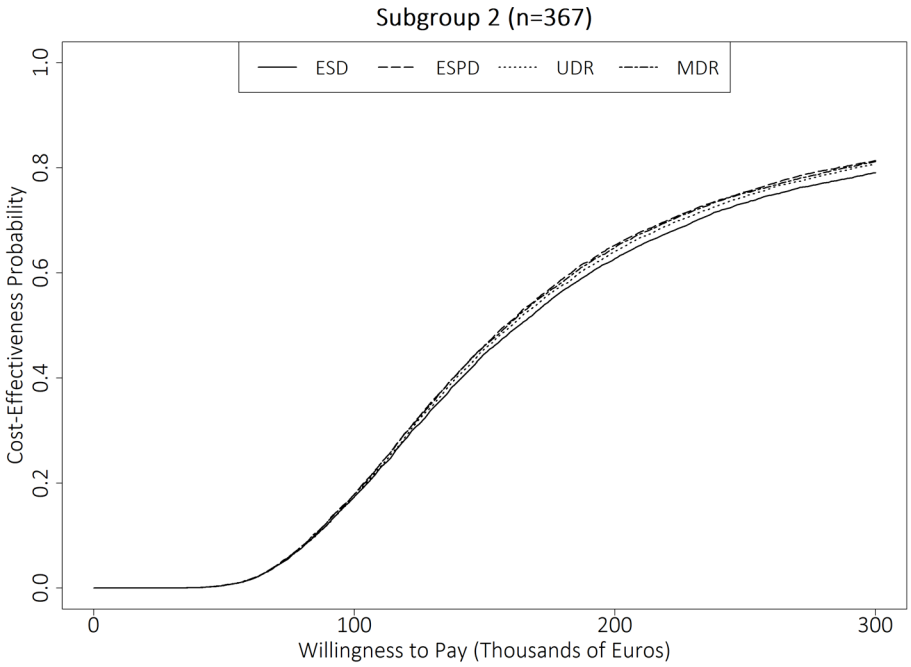

Subgroup 7 (n=269)
80,000
60,000
40,000
20,000
-20,000
-0.6
-0.4
-0.2
0.2
0.4
0.6
0.8
Incremental Costs in Euros
Incremental Effectiveness in QALYs
ESD
ESPD
UDR
MDR

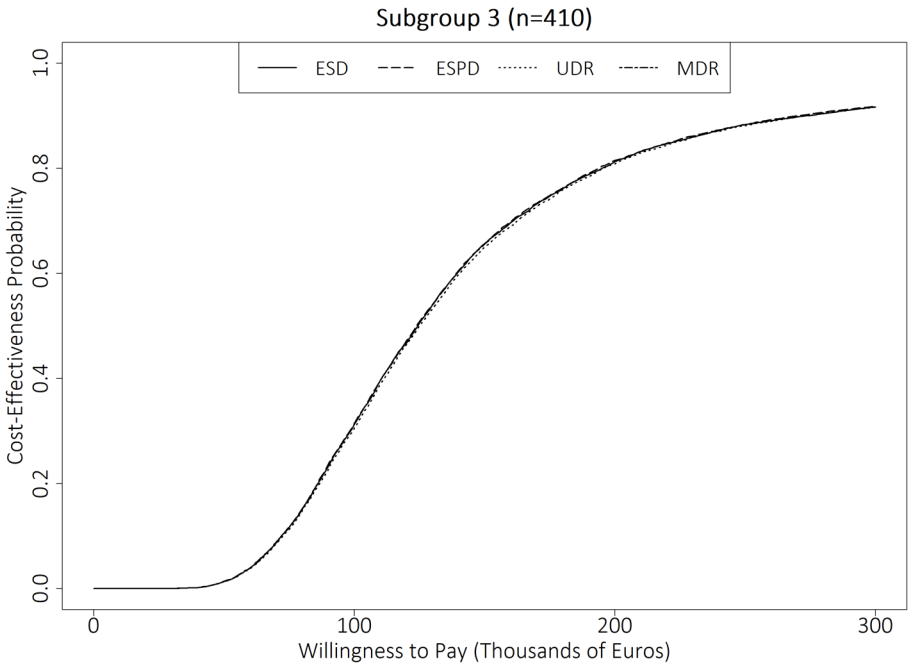

Subgroup 8 (n=97)
80,000
60,000
40,000
20,000
-20,000
-0.6
-0.4
-0.2
0.2
0.4
0.6
0.8
Incremental Costs in Euros
Incremental Effectiveness in QALYs
ESD
ESPD
UDR
MDR

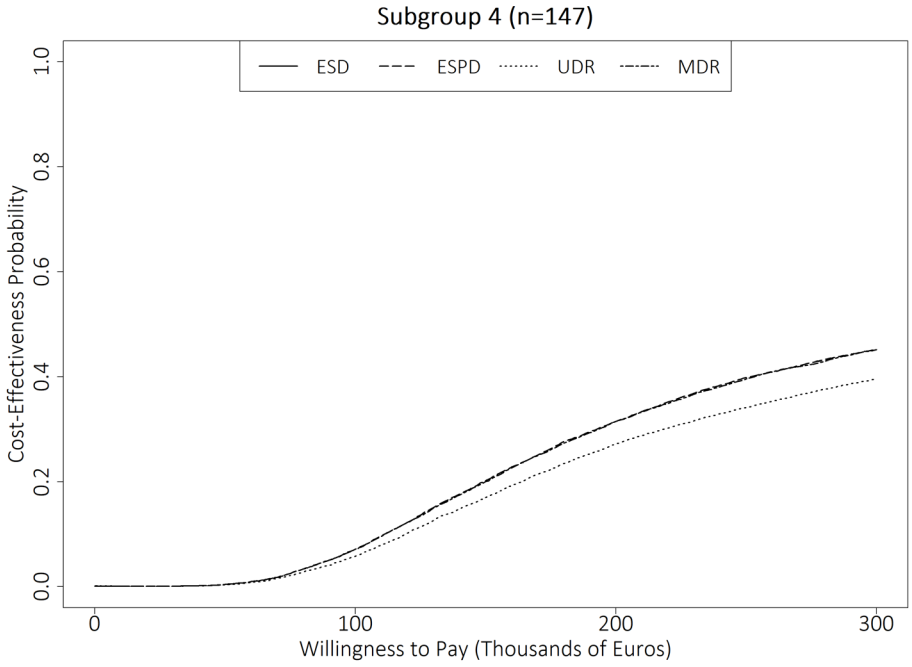
APPENDIX 6: cost-effectiveness acceptability curves for all case study subgroup analyses

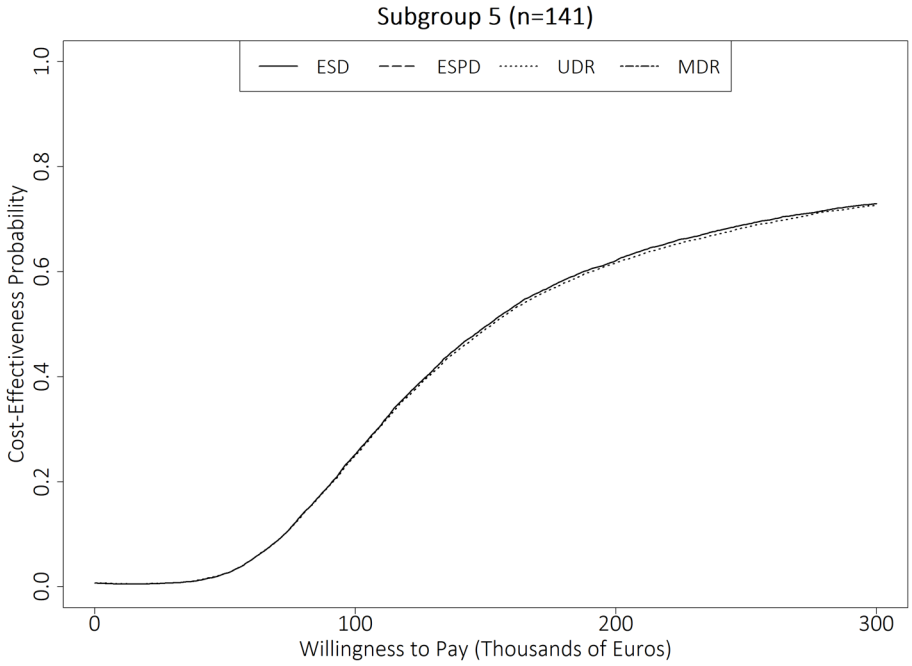


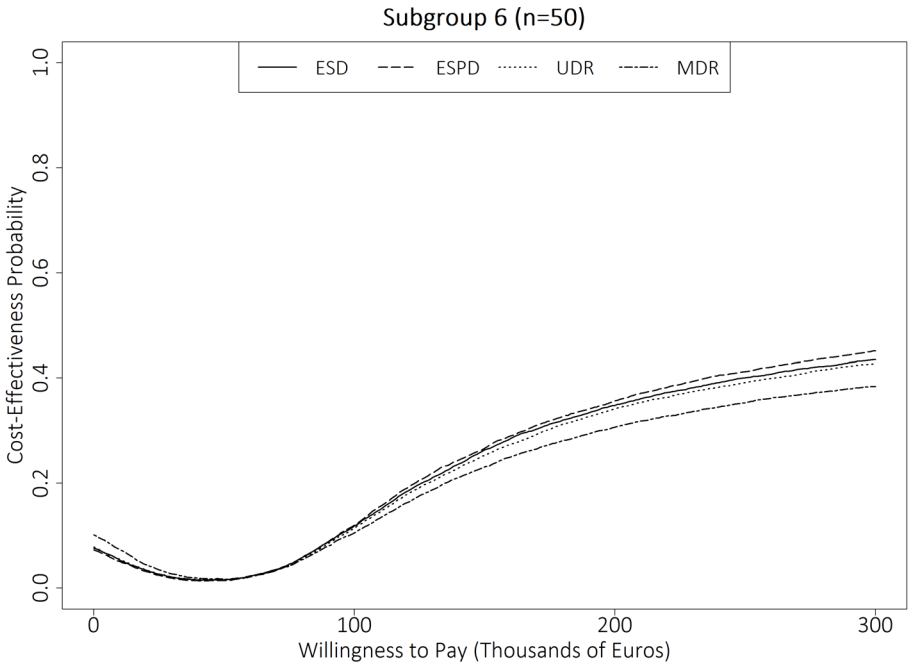


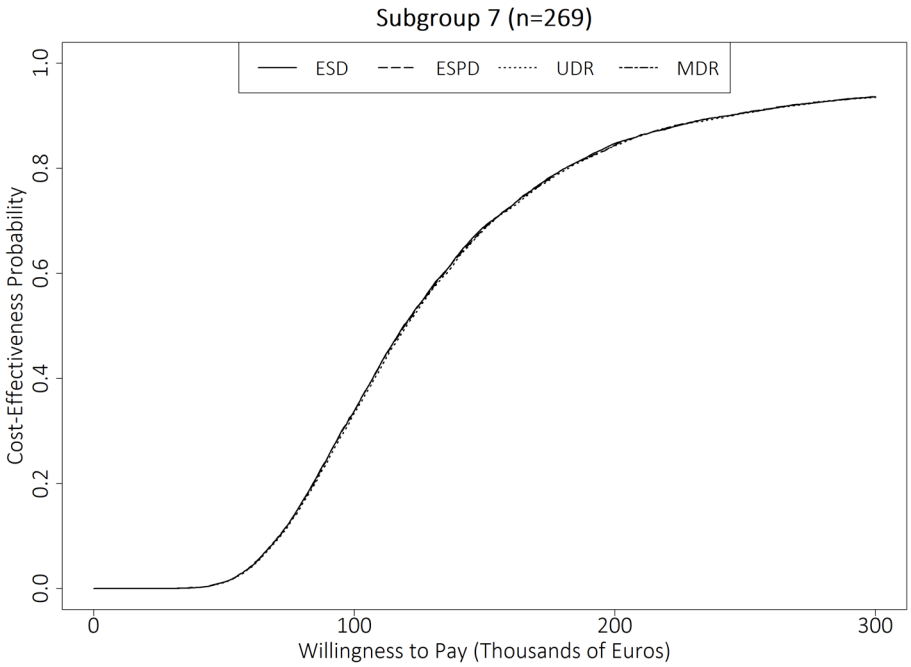

Subgroup 2 (n=367)
ESD
ESPD
UDR
MDR
Cost-Effectiveness Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
Willingness to Pay (Thousands of Euros)

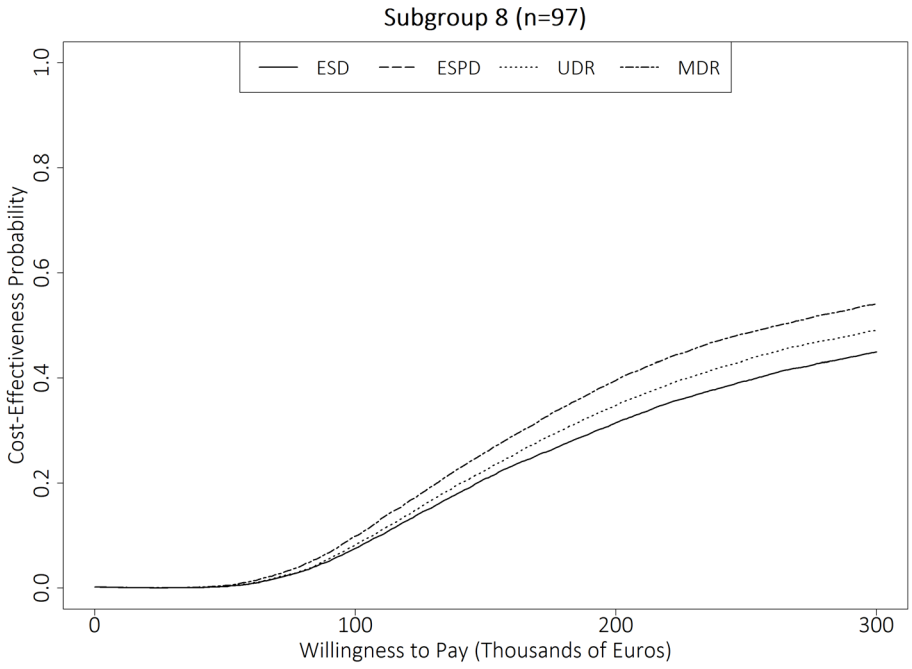

Subgroup 3 (n=410)
ESD
ESPD
UDR
MDR
Cost-Effectiveness Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
Willingness to Pay (Thousands of Euros)


Subgroup 4 (n=147)
ESD
ESPD
UDR
MDR
Cost-Effectiveness Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
Willingness to Pay (Thousands of Euros)


Subgroup 5 (n=141)
ESD
ESPD
UDR
MDR
Cost-Effectiveness Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
Willingness to Pay (Thousands of Euros)


Subgroup 6 (n=50)
ESD
ESPD
UDR
MDR
Cost-Effectiveness Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
Willingness to Pay (Thousands of Euros)


Subgroup 7 (n=269)
ESD
ESPD
UDR
MDR
Cost-Effectiveness Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
Willingness to Pay (Thousands of Euros)


Subgroup 8 (n=97)
ESD
ESPD
UDR
MDR
Cost-Effectiveness Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
Willingness to Pay (Thousands of Euros)

CHAPTER 7

Comparing modeling approaches for discrete event simulations of competing events based on censored individual patient data: a simulation study and illustration in colorectal cancer

Koen Degeling
Maarten J. IJzerman
Catharina G.M. Groothuis-Oudshoorn
Mira D. Franken
Miriam Koopman
Hendrik Koffijberg

ABSTRACT

Background: Recommendations regarding modeling approaches to implement competing events in discrete event simulations (DES) based on uncensored individual patient data (IPD) are available. This study aims to complement these recommendations with guidance for handling censored IPD by describing, illustrating, and quantitatively comparing alternative modeling approaches.

Methods: Four modeling approaches were considered: 1) event-specific distributions (ESD), 2) event-specific probabilities and distributions (ESPD), 3) a unimodal distribution and regression model (UDR), and 4) a multimodal distribution and regression model (MDR). A simulation study and a case study in colorectal cancer assessed performance of these approaches in terms of relative absolute event incidence difference and relative entropy of time-to-event distributions. Performance was assessed for 135 scenarios defined by the level of right-censoring, number of competing events, overlap of competing time-to-event distributions, and sample size. Additionally, differences in health economic outcomes were illustrated for clinically relevant subgroups in the case study.

Results: Censoring resulted in decreased performance, increased uncertainty surrounding cost-effectiveness point-estimates, and different point-estimates for several subgroup analyses. No single modeling approach performed best. For scenarios including low levels of censoring (10%), ESPD or UDR performed best, depending on the overlap of competing time-to-event distributions. For moderate levels of censoring (30%), UDR performed best. However, the case study illustrated that performance differed substantially for subgroups with similar disease pathway and data characteristics and the same level of censoring, limiting the generalizability of these findings.

Conclusions: Decision analytic modelers should be aware of different alternatives available for implementing competing events in DES based on censored IPD. Although some general recommendations could be made, decisions on which modeling approach to implement should be based on internal validity and justified.

INTRODUCTION

Decision analytic modelers have increasingly been employing alternatives to the commonly used cohort-level state-transition modeling (STM) technique following the need for models to reflect the complex dynamics of clinical pathways [1-3]. These alternative modeling techniques, such as agent-based modeling and discrete event simulation (DES), are being considered more advanced compared to cohort-level STM, mainly in terms of computational complexity and required analytical skills to implement them [4, 5]. Methodological guidance is therefore essential to inform about design choices that need to be made and need to be reported on when applying these techniques, and to provide evidence-based recommendations to support development of high quality models.

DES is a useful alternative to STM and can be used to model resource-constrained processes, but also to represent personalized treatment processes due to its ability to reflect dynamic pathways based on patient-level histories and characteristics [6, 7]. Although aggregated data can be used to develop DES models, individual patient data (IPD) are the preferred source of evidence to account for stochastic uncertainty and patient heterogeneity. An important design choice in developing DES models based on IPD concerns how competing events are implemented [8]. Competing events are those that prevent other events of interest from occurring, or fundamentally change the probability with which these other events of interest occur [9]. Four strategies have been identified to implement competing events in DES models: 1) simulating times for all competing events and selecting the event that is the first to occur, 2) selecting the event to occur first and the corresponding time-to-event second, 3) selecting the time-to-event first and the corresponding event second, and 4) using discretized cyclic probabilities to resemble discrete-time STM [10].

Based on these general strategies, four specific modeling approaches for implementing competing events in DES models informed by uncensored IPD have been defined and compared [11]. Substantial performance differences in terms of simulated event incidence and relative entropy of time-to-event distributions were demonstrated for these modeling approaches. A general recommendation to use Strategy 3 was provided, whereas an approach based on Strategy 2 also showed good performance with relatively easy implementation. Although useful, this guidance cannot be generalized to DES models informed by censored IPD.

It is important to appropriately account for censoring in analyses involving competing events, since neglecting to do so may affect model outcomes [12]. Censoring can be classified as informative or non-informative, and according to whether individuals are interval-, left-, or right-censored [13]. Here, focus is on non-informative, right-censored data. Non-informative censoring assumes that knowledge of the censoring time does not provide additional information on the occurrence of an event. Right-censored individuals are those for which it is known that an event has not yet occurred prior to a specific point in time.

This study aims to describe, illustrate, and compare different modeling approaches for implementing competing events in DES models informed by non-informative right-censored IPD, and provide recommendations so that modelers can make informed and deliberative deci-

sions regarding this design choice. Modeling approaches are compared in a simulation study and a case study in colorectal cancer, with their performance assessed for different levels of censoring in terms of relative absolute incidence difference and relative entropy of time-to-event distributions. Additionally, potential differences in terms of health economic outcomes are illustrated for the case study.

METHODS

Four modeling approaches for implementing competing events in DES models informed by censored IPD were defined in line with those defined previously for DES models informed by uncensored IPD: event-specific distributions (ESD), event-specific probabilities and distributions (ESPD), a unimodal distribution and regression model (UDR), and a multimodal distribution and regression model (MDR) [11]. A detailed discussion of these modeling approaches is provided below, using hypothetical data presented in Table 1 as an example. Pseudo algorithms for data analyses and simulations according to these modeling approaches are provided in Appendices 1-4. Code illustrating how these approaches can be implemented in R [14] is provided online: www.personex.nl/research/competing-risks.

Before discussing the modeling approaches, some general notations and definitions are provided. Let T be the time-to-event or censoring time, a random variable that is distributed according to a particular probability density function $f(t)$, with cumulative distribution function $F(t)$ and survival function $S(t)$ defined as:

$$F(t) = \Pr(T \le t) = \int_0^t f(x)dx \qquad \textit{Equation 1}$$

$$S(t) = \Pr(T > t) = 1 - F(t) = 1 - \int_0^t f(x)dx = \int_t^\infty f(x)dx \qquad \textit{Equation 2}$$

The hazard function $h(t)$ (Equation 3) can be expressed in terms of the probability density function $f(t)$ and survival function $S(t)$ (Equations 4 & 5).

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \le t + \Delta t \mid T > t)}{\Delta t} \qquad \textit{Equation 3}$$

$$h(t) = f(t) / S(t) \qquad \textit{Equation 4}$$

$$S(t) = \exp\left(-\int_0^t h(x)dx\right) \qquad \textit{Equation 5}$$

In order to extend these general survival analysis notations and definitions to a competing events setting, let $K=\{1,...,k\}$ be the index set for k mutually exclusive competing events, and let C_j be the censoring indicator that indicates whether a person experienced event $j, j=1,...,k$. Suppose $i=1,...,n$ observations of time-to-event t_i, then the censoring indicator indicates whether event j was observed (c_{ij} =1) or not (c_{ij} =0). Finally, it is assumed that time-to-event distributions D can be parametrized by a finite set of distribution parameters φ. Next, the four modeling approaches will be introduced and explained.

Modeling approach 1: event-specific distributions (ESD)

The ESD modeling approach uses event-specific time-to-event distributions to draw random time-to-events for each competing event, and subsequently selects the first event to occur to be simulated. To illustrate this according to Table 1, a distribution D_j needs to be fitted for each competing event $j=1,...,k$ by estimating distribution parameters φ_j that define D_j, resulting in distributions D_{Death} and D_{Progr} that describe the time-to-death T_{Death} and time-to-progression T_{Progr} (in the illustration k = 2). To perform a simulation, a time t_j for each competing event $j=1,...,k$ needs to be drawn randomly from corresponding time-to-event distribution D_j. Subsequently, the event that is the first to occur, i.e. the event j corresponding to the lowest drawn time-to-event t_j, is selected and will be simulated. See Appendix 1 for a pseudo-algorithm.

For the ESD modeling approach, event-specific distributions D_j are estimated according a cause-specific (CS) hazards model, which assumes that the risk set only includes individuals who are event free at the respective point in time, i.e. those who have not yet experienced *any* event [15]. Denoting the instantaneous rate of occurrence of event $j, j=1,...,k$ in individuals who are currently event free, the CS hazard function $h_j^{CS}(t)$ is defined as [15, 16]:

$$h_j^{CS}(t)=\lim_{\Delta t\to 0}\frac{\Pr\left(t<T\le t+\Delta t, C_j=1 \mid T>t\right)}{\Delta t}=\frac{f_j^{CS}(t)}{S(t)} \qquad \textit{Equation 6}$$

In a competing events setting, cumulative distribution function $F(t)$ is referred to as the cumulative incidence function $F_j(t)$, which denotes the probability with which event j occurs before or at time t. According to the CS hazards model, cumulative incidence function $F_j^{CS}(t)$ for competing event $j, j=1,...,k$ is defined as [17, 18]:

$$F_j^{CS}(t)=\int_0^t \Pr\left(T>x\right)h_j^{CS}(x)dx=\int_0^t S(x)h_j^{CS}(x)dx \qquad \textit{Equation 7}$$

$$\text{where } S(t)=\exp\left(-\int_0^t\left[\sum\nolimits_{j=1}^{k}h_j^{CS}(x)\right]dx\right) \qquad \textit{Equation 8}$$

By parameterizing the general likelihood function of the data, given parameters φ specifying the underlying hazards model, $L(t_1,...,t_n \mid \varphi)$ (Equation 9) [18] according to the CS hazards model (Equation 6), parameterized likelihood function $L^{CS}(t_1,...,t_n \mid \varphi_K^{CS})$ is obtained (Equation 10), where $\varphi_K^{CS} = \left\{\varphi_1^{CS},...,\varphi_k^{CS}\right\}$ represent the sets of event-specific distribution parameters [19].

$$L(t_1,...,t_n \mid \varphi) = \prod_{i=1}^{n}\left[\prod_{j=1}^{k}\left[f_j(t_i \mid \varphi)^{c_{ij}}\right]\left(S(t_i \mid \varphi)\right)^{1-\sum_{j=1}^{k} c_{ij}}\right] \qquad \textit{Equation 9}$$

$$L^{CS}(t_1,...,t_n \mid \varphi_K^{CS}) = \prod_{i=1}^{n}\left[\prod_{j=1}^{k}\left[\left(S(t_i \mid \varphi_K^{CS})h_j^{CS}(t_i \mid \varphi_j^{CS})\right)^{c_{ij}}\right]\left(S(t_i \mid \varphi_K^{CS})\right)^{1-\sum_{j=1}^{k} c_{ij}}\right] \qquad \textit{Equation 10}$$

Maximum likelihood estimation according to CS hazards models may be done separately for each competing event, which simplifies the estimation of the parameters, as inferences about φ_1^{CS} do not require correct specification of model $h_2^{CS}(t_i \mid \varphi_2^{CS})$, and vice versa [18, 20]. In doing so, likelihood function $L^{CS}(t_1,...,t_n \mid \varphi_K^{CS})$ can be simplified to event-specific likelihood functions $L_j^{CS}(t_1,...,t_n \mid \varphi_j^{CS})$, $j = 1,...,k$ (Equation 11), which can be used to estimate parameters φ_j^{CS} for each distribution D_j.

$$L_j^{CS}(t_1,...,t_n \mid \varphi_j^{CS}) = \prod_{i=1}^{n}\left[\left(S_j(t_i \mid \varphi_j^{CS})h_j^{CS}(t_i \mid \varphi_j^{CS})\right)^{c_{ij}}\left(S_j(t_i \mid \varphi_j^{CS})\right)^{1-c_{ij}}\right] \qquad \textit{Equation 11}$$

$$\text{where } S_j(t \mid \varphi_j) = \exp\left(-\int_0^t h_j^{CS}(x \mid \varphi_j)dx\right) \qquad \textit{Equation 12}$$

Modeling approach 2: event-specific probabilities and distributions (ESPD)

The ESPD modeling approach first uses event-specific probabilities to determine which competing event will occur, and subsequently samples a time at which this event will occur from the corresponding event-specific time-to-event distribution. Like for the ESD modeling approach, a time-to-event distribution needs to be estimated for each competing event, though for the ESPD modeling approach these distributions are conditional on a previously selected event and estimated according to the subdistribution hazards (SD) model, which will be further explained below. To illustrate this according to Table 1, event-specific probabilities $\Pr(C_{Death} = 1)$ and $\Pr(C_{Progr} = 1)$ need to be estimated, and event-specific distributions D_{Death} and D_{Progr} need to be fitted by estimating the corresponding distribution parameters φ_{Death}^{SD} and φ_{Progr}^{SD} to describe the time-to-event T_j, $j = 1,...,k$ given $C_j = 1$, i.e. $T_j \mid C_j = 1 \sim D_j\left(\varphi_j^{SD}\right)$. To perform a simulation, event specific probabilities $\Pr(C_j = 1)$ are used to determine which event will be simulated. Subsequently, a time-to-event t will be sampled from corresponding distribution D_j. See Appendix 2 for a pseudo-algorithm.

Table 1. *Hypothetical patient-level time-to-event data used to illustrate the modeling approaches included in this study. The data concerns nine hypothetical observations of patients that were subject to two competing events: progression and death. Progr = Progression; T represents a random variable to describe an event time or censoring time; D represents a certain time-to-event distribution (e.g., Weibull, Gamma or mixture) defined by parameters φ; $\Pr(C_j = 1)$ represents the probability that the competing event j will occur; $r(t)$ represents a (multinomial) logistic regression model predicting the competing event to occur (dependent variable) based on the time-to-event t (independent variable).*

Observed Data		**ESD Approach**		**ESPD Approach**		**UDR Approach**	**MDR Approach**
Time	**Event**	C_{Death}	C_{Progr}	C_{Death}	C_{Progr}	C	C
154	Death	1	0	1	0	1	1
351	Progr.	0	1	0	1	1	1
294	Progr.	0	1	0	1	1	1
310	-	0	0	0	0	0	0
257	Death	1	0	1	0	1	1
211	Death	1	0	1	0	1	1
438	-	0	0	0	0	0	0
80	Progr.	0	1	0	1	1	1
347	Progr.	0	1	0	1	1	1
Example of Data Analysis Output:		$T_{Death} \sim D_{Death}(\varphi_{Death})$	$T_{Progr} \sim D_{Progr}(\varphi_{Progr})$	$\Pr(C_{Death} = 1)$ and $T_{Death} \mid C_{Death} = 1 \sim D_{Death}(\varphi_{Death})$	$\Pr(C_{Progr} = 1)$ and $T_{Progr} \mid C_{Progr} = 1 \sim D_{Progr}(\varphi_{Progr})$	$T \sim D(\varphi_{unimodal})$ and $\Pr(C_j = 1 \mid T = t) = r(t)$	$T \sim D(\varphi_{multimodal})$ and $\Pr(C_j = 1 \mid T = t) = r(t)$

The event-specific probabilities are estimated according to the CS hazards model introduced above. Recall that cumulative incidence function $F_j^{CS}(t)$ is based on survival function $S(t)$ that includes all competing events $j=1,\ldots,k$, and not on an event-specific survival function $S_j(t)$ (Equations 7 & 8). Therefore, sub-density function $f_j^{CS}(t)$ according to the CS hazards model does not represent a proper distribution, as this function does not necessarily integrates to one [15]:

$$F_j^{CS}(\infty)=\int_0^{\infty} S(x)h_j^{CS}(x)dx=\int_0^{\infty} f_j^{CS}(x)dx\neq 1 \qquad \textit{Equation 13}$$

By parameterizing cumulative incidence function $F_j^{CS}(t)$ (Equation 7) according parameters $\varphi_K^{CS}=\left\{\varphi_1^{CS},\ldots,\varphi_k^{CS}\right\}$, estimated using Equation 11, event-specific probabilities for the ESPD modeling approach can be estimated using Equation 14. These event-specific probabilities are also referred to as absolute cause-specific risks or cause-specific failure probabilities in competing risks literature.

$$\Pr\left(C_j=1\right)=\lim_{t\to\infty}F_j^{CS}(t) \qquad \textit{Equation 14}$$

To estimate the event-specific time-to-event distributions, a SD hazards model is assumed. The SD hazard function $h_j^{SD}(t)$, $j=1,\ldots,k$, defined as [15, 16]:

$$h_j^{SD}(t)=\lim_{\Delta t\to 0}\frac{\Pr\left(t<T\leq t+\Delta t,C_j=1\,|\,T>t\right)+\Pr\left(t<T\leq t+\Delta t,C_j=1\,|\,T<t,C_j=0\right)}{\Delta t}$$

$$=\frac{f_j^{SD}(t)}{\left(1-F_j^{SD}(t)\right)} \qquad \textit{Equation 15}$$

denotes the instantaneous risk of failure from event j in individuals who have not yet experienced *an event of type* j [15]. From Equation 15 it can be seen that the risk set over which the SD hazards model is defined includes individuals who have experienced a competing event before time t, which is different from the risk set over which a CS hazards model is defined, as the latter only includes those who are completely event free at time t. According to the SD hazards model, the cumulative incidence function $F_j^{SD}(t)$, $j=1,\ldots,k$ is defined as [17]:

$$F_j^{SD}(t)=1-e^{-\int_0^t h_j^{SD}(x)dx}=1-S_j(t) \qquad \textit{Equation 16}$$

Since cumulative incidence function $F_j^{SD}(t)$ is solely based on event of interest j, integrating this function results in an absolute event probability equal to one, illustrating why the SD hazards model cannot be used to estimate event-specific probabilities. Additionally, this illustrates why the SD hazards model is used to estimate event-specific distributions for the

ESPD approach. Since the time-to-event distributions D_j, $j=1,...,k$ for the ESPD approach are conditional on the selected event, i.e. it is known that event j will occur when a time-to-event is to be drawn from distribution D_j, integrating cumulative incidence function $F_j^{SD}(t)$ should result in one.

By parameterizing general likelihood function $L(t_1,...,t_n \mid \varphi)$ (Equation 9) according the SD hazards model, parameterized likelihood function $L^{SD}(t_1,...,t_n \mid \varphi_K^{SD})$ is obtained [19], which can be used to estimate distribution parameters $\varphi_K^{SD} = \left\{\varphi_1^{SD},...,\varphi_k^{SD}\right\}$:

$$L^{SD}(t_1,...,t_n \mid \varphi_K^{SD}) = \prod_{i=1}^{n}\left[\prod_{j=1}^{k}\left[\left(h_j^{SD}(t_i \mid \varphi_j^{SD})\left(1-F_j^{SD}(t_i \mid \varphi_j^{SD})\right)\right)^{c_{ij}}\right]\left(1-\sum_{j=1}^{k}F_j^{SD}(t_i \mid \varphi_j^{SD})\right)^{1-\sum_{j=1}^{k}c_{ij}}\right]$$

Equation 17

Modeling approach 3: unimodal joint distribution and regression model (UDR)

The UDR modeling approach first selects the time at which an event will occur, using a single time-to-event distribution D, defined as the distribution of the minimum of T_j, $j=1,...,k$, i.e. $\min\left(T_1,...,T_k\right)$. In this approach it is assumed that D is unimodal and parameterized with parameters $\varphi_{unimodal}$. Subsequently, the event corresponding to a specific time-to-event is modeled using a (multinomial) logistic regression model. A logistic regression model can be used to model binary data, i.e. two competing events, whereas a multinomial logistic regression model is required to model more than two competing events. To illustrate this using Table 1, distribution D is fitted by estimating distribution parameters $\varphi_{unimodal}$, and a (multinomial) logistic regression model $r(t)$ is estimated to predict probabilities $\Pr\left(C_j=1 \mid T=t\right)$, $j=1,...,k$ with which the competing events occur given time t. Accordingly, a simulation is performed by first sampling a time t from time-to-event distribution D, and subsequently an event using regression model $r(t)$. See Appendix 3 for a pseudo-algorithm.

Since time-to-event distribution D represents all competing events, only one hazard function needs to be modeled. Therefore, distribution parameters $\varphi_{unimodal}$ will be estimated according to either the CS or SD hazards model. In doing so, all competing events are treated as one event, i.e. a single censoring indicator can be defined as C such that for n observations individual i, $i=1,...,n$ experienced any event (c_i =1) or is censored (c_i =0). Given this notation, likelihood functions $L^{CS}(t_1,...,t_n \mid \varphi_K^{CS})$ and $L^{SD}(t_1,...,t_n \mid \varphi_K^{SD})$ (Equations 10 & 17) can be simplified for a single distribution (Equations 18 & 19). These simplified likelihood functions are identical given the relation between the cumulative distribution function $F(t)$ and survival function $S(t)$ (Equation 2), and can be used to estimate parameters $\varphi_{unimodal}$.

$$L^{CS|UDR}(t_1,...,t_n \mid \varphi^{CS}) = \prod_{i=1}^{n}\left[\left(S(t_i \mid \varphi^{CS})h^{CS}(t_i \mid \varphi^{CS})\right)^{c_i}\left(S(t_i \mid \varphi^{CS})\right)^{1-c_i}\right]$$

Equation 18

$$L^{SD|UDR}(t_1,...,t_n \mid \varphi^{SD}) = \prod_{i=1}^{n}\left[\left(\left(1-F^{SD}(t_i \mid \varphi^{SD})\right)h^{SD}(t_i \mid \varphi^{SD})\right)^{c_i}\left(1-F^{SD}(t_i \mid \varphi^{SD})\right)^{1-c_i}\right]$$

Equation 19

Modeling approach 4: multimodal joint distribution and regression model (MDR)

Similar to the UDR modeling approach, the MDR modeling approach first draws a random time based on a time-to-event distribution D as defined for the UDR modeling approach, and subsequently determines the corresponding event using a (multinomial) logistic regression model. Unlike the UDR modeling approach, the MDR approach allows D to be multimodal rather than unimodal, and parameterized with parameters $\varphi_{multimodal}$. As illustrated in Table 1, a (multinomial) logistic regression model $r(t)$ is estimated and implemented as for the UDR modeling approach. See Appendix 4 for a pseudo-algorithm.

Here focus will be on the use of mixture distributions to represent multimodal data [21]. For a finite mixture of a distribution set $M=\{1,...,m\}$, the family of probability density functions $g_{mix}(t \mid \alpha_M)$ is defined as [22]:

$$g_{mix}(t \mid \alpha_M) = \sum_{j=1}^{m}\alpha_j g_j(t)$$

Equation 20

Where $\alpha_M=\{\alpha_1,...,\alpha_m\}$ represents the set of mixing proportions α_j of classes $j=1,...,m$ in the mixture, $g_j(t)$ is a probability density function for class j. Mixture cumulative distribution function $G_{mix}(t \mid \alpha_M)$ and mixture survival function $S_{mix}(t \mid \alpha_M)$ can be defined as:

$$G_{mix}(t \mid \alpha_M) = \int_0^t g_{mix}(x \mid \alpha_M)dx = \int_0^t\left(\sum_{j=1}^{m}\alpha_j g_j(x)\right)dx$$

Equation 21

$$S_{mix}(t \mid \alpha_M) = 1-G_{mix}(t \mid \alpha_M) = 1-\int_0^t\left(\sum_{j=1}^{m}\alpha_j g_j(x)\right)dx = \int_t^{\infty}\left(\sum_{j=1}^{m}\alpha_j g_j(x)\right)dx$$

Equation 22

Although mixture proportions $\alpha_M=\{\alpha_1,...,\alpha_m\}$ and distribution parameters $\varphi_M=\{\varphi_1,...,\varphi_m\}$ can be estimated by maximum likelihood [21], this is challenging due to the constraint that the sum of the mixing proportions α_M should be equal to one. This is especially challenging for scenarios involving more than two competing events or a low number of observations, for example due to censoring. Therefore, mixture distribution D here was parameterized using parameters estimated for the ESPD approach. More specifically, it was assumed that $m=k$ and event-specific probabilities $\Pr(C_j=1)$, $j=1,...,k$ were used to define mixing proportions $\alpha_M=\{\alpha_1,...,\alpha_m\}$ and time-to-event distributions parameters $\varphi_K^{SD}=\{\varphi_1^{SD},...,\varphi_k^{SD}\}$ were used to define mixture distribution parameters $\varphi_M=\{\varphi_1,...,\varphi_m\}$.

Simulation study to compare the performance of the approaches

Event incidences and time-to-event distributions of competing events were simulated to compare the modeling approaches' performance for different scenarios by varying data and disease pathway characteristics, including the sample size, number of competing risks, overlap of time-to-event distributions, and proportion of right-censoring. Figure 1a presents the conceptual model structure used for this simulation study. As illustrated in Figure 2, simulations were performed for scenarios including different numbers of competing events n_{event} (n_{event} = 2, 3, 4), level of overlap of the corresponding competing time-to-event distributions $p_{overlap}$ ($p_{overlap}$ ≈ 10%, 50%, 90%, i.e. low, medium, or high, see Appendices 5 and 6), sample sizes n_{sample} (n_{sample} = 50, 150, 500), and levels of censoring $p_{censoring}$ ($p_{censoring}$ ≈ 0%, 10%, 30%, 60%, 90%), resulting in a total of 135 unique combinations of data and disease pathway characteristics. Scenarios without censoring (i.e., $p_{censoring}$ = 0%) were included for comparison with previous work focusing on uncensored IPD [11].

First, nine uncensored populations S of 100,000 individuals were simulated according to the nine combinations of disease pathway characteristics n_{event} and $p_{overlap}$ (see Appendix 7 for exact definitions of time-to-event distributions and event incidences). For each unique combination of n_{event}, $p_{overlap}$, n_{sample}, and $p_{censoring}$, a total of 10,000 simulation runs were performed. In each simulation run, an uncensored sample $s_{uncensored}$ of corresponding sample size n_{sample} was sampled from respective population S and right-censored to censoring level $p_{censoring}$ to obtain censored sample $s_{censored}$. Right-censoring was performed here by setting a hypothetical time threshold beyond which outcomes were censored, such that the appropriate censoring level resulted. Next, all four modeling approaches were applied to analyze sample $s_{censored}$ and, subsequently, to simulate event incidences and time-to-events for 10,000 new individuals to obtain a simulation sample $s_{simulation}$ for each modeling approach. Finally, the modeling approaches' performance was assessed by comparing event incidences and time-to-event distributions in these newly simulated samples $s_{simulation}$ to those in the corresponding uncensored sample $s_{uncensored}$ (i.e., internal validation) and corresponding population S (i.e., external validation).

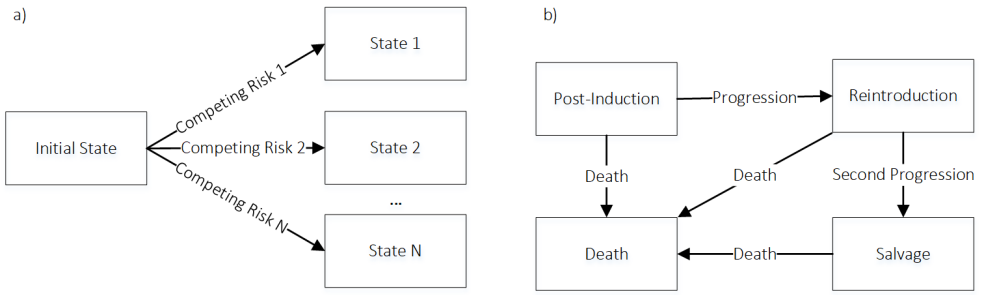


Figure 1. *Graphical representation of a) the simulation study model structure, and b) the case study model structure.*

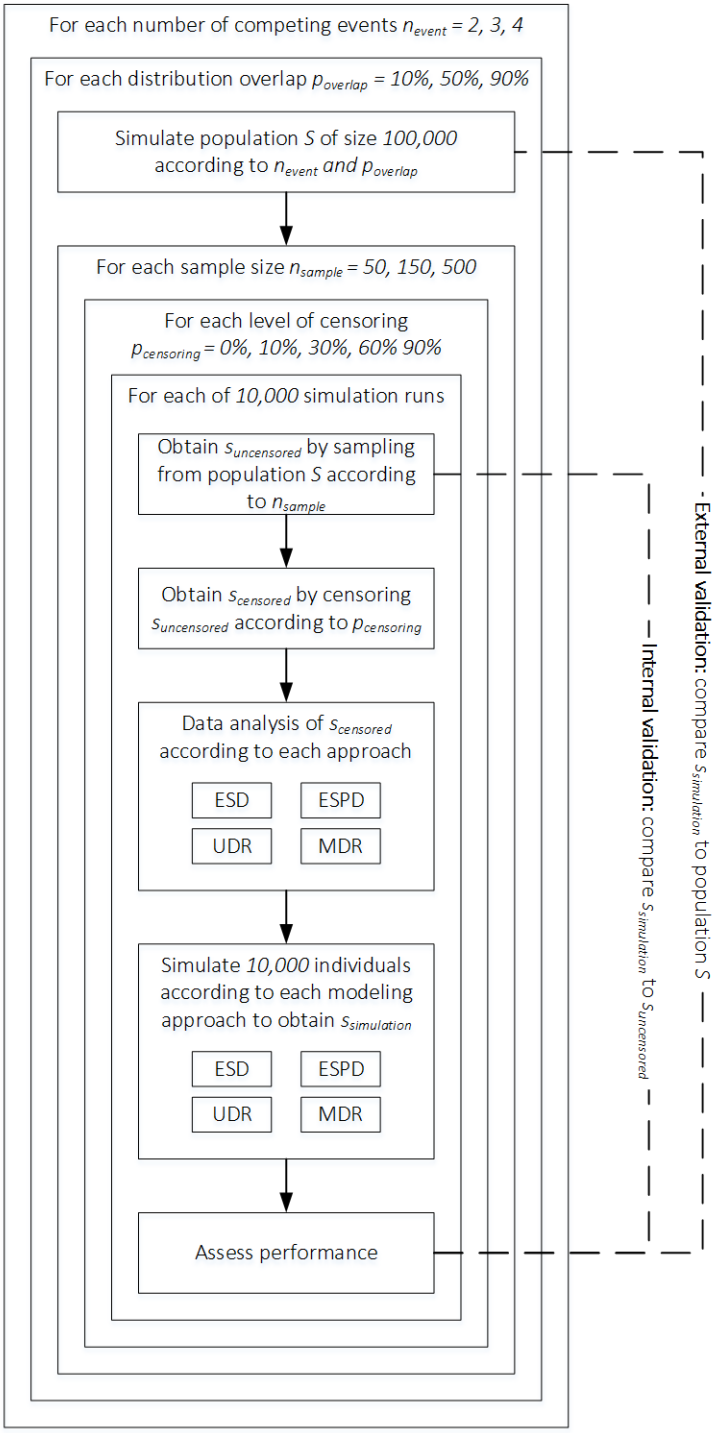


Figure 2. *Overview of the simulation study process.*

Bias in terms of relative absolute incidence difference (RAID) was used as performance measure for the event incidence:

$$RAID = \left(\frac{\left| Incidence_{simulated} - Incidence_{observed} \right|}{Incidence_{observed}} \right) x100\% \qquad \textit{Equation 23}$$

The performance with regard to simulated time-to-event distributions was based on the relative entropy, i.e. the Kullback-Leibler divergence (KLD), which is a measure of the difference between two probability distributions $f(t)$ and $g(t)$, for which lower values indicate a better performance [23]:

$$KLD(f \,||\, g) = \int f(t) \log\left(\frac{f(t)}{g(t)} \right) \qquad \textit{Equation 24}$$

Both RAID and KLD were calculated separately for each competing event and separately for internal and external validation. To summarize the RAID and KLD across competing events per scenario, event-specific performance outcomes were weighted according to event incidences as observed in corresponding population S.

The simulation study was performed in R version 3.4.1 [14]. Time-to-event data were simulated and analyzed using Weibull distributions to rule out potential bias due to mismatching distributions. Weibull distributions were selected because these distributions are commonly used in survival analysis and showed to accurately represent the distribution of the time-to-event data in the case study. In order to apply the previously introduced equations, hazard functions were parameterized according to the Weibull's shape parameter κ and rate parameter ρ [18], i.e. $\varphi_j = \{\kappa_j, \rho_j\}$, $j = 1,\ldots,k$:

$$h_j(t) = \kappa_j \left(\rho_j t \right)^{\kappa_j} / t \qquad \textit{Equation 25}$$

Several packages for R were used: the *mixtools* package to perform random draws from a mixture distribution [24], the *nnet* package to estimate (multinomial) logistic regression models [25], and the *flexmix* package to calculate KLD [26-28].

Case study to illustrate the potential impact on health economic outcomes

A case study based on data from the randomized phase III CAIRO3 study of the Dutch Colorectal Cancer Group (NCT00442637) was performed to assess the modeling approaches' performance and illustrate their potential impact on health economic outcomes in a real-world scenario. The CAIRO3 study randomized 558 metastatic colorectal cancer patients with stable disease or better after six cycles of capecitabine, oxaliplatin, and bevacizumab (CAPOX-B) induction therapy to either capecitabine and bevacizumab (CAP-B) maintenance treatment (intervention) or observation (control) until progression of disease [29]. For both the main-

tenance and observation strategy, CAPOX-B treatment was to be reintroduced upon progression and continued until second progression, which was the primary endpoint of the study.

A discrete-time cohort STM was developed previously to perform a health economic evaluation of the CAIRO3 study [30], and a DES model was developed for comparison with this discrete-time cohort STM [5]. The DES model was structured according to the CAIRO3 study: post-induction, reintroduction, salvage, and death (Figure 1b), and has been described in detail elsewhere [5]. The modeling approaches were implemented in the exact same way for the case study as for the simulation study.

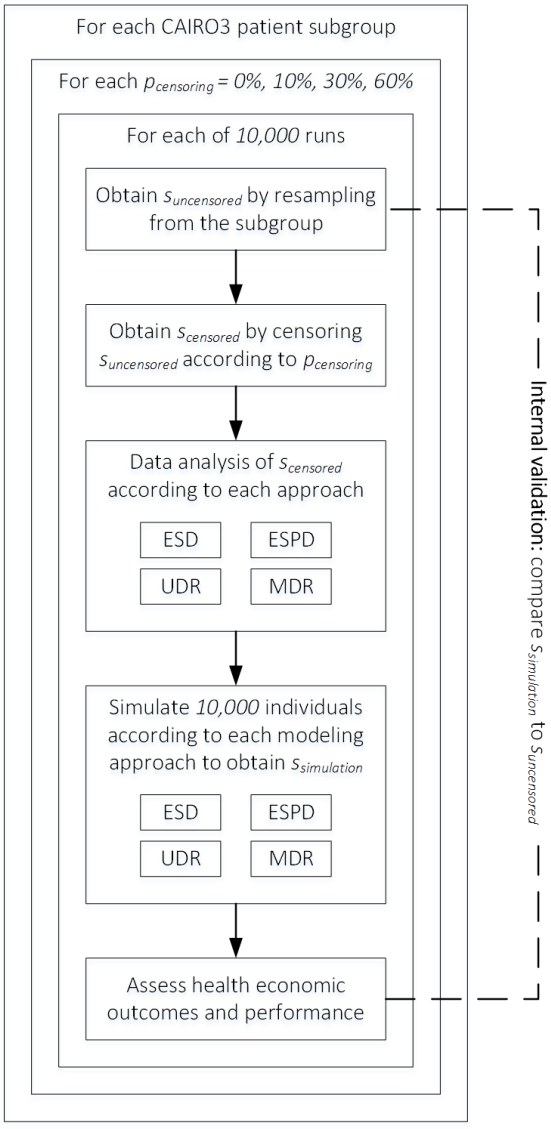


Figure 3. *Overview of the case study process.*

In addition to an analysis based on the complete CAIRO3 patient cohort, clinical relevant subgroup analyses were performed to illustrate potential sample size impact on modeling outcomes. A total of 8 subgroups with sample sizes ranging from 50 to 410 were defined according to patient characteristics that were found relevant in the evaluation of the CAIRO3 study [29], i.e. treatment response (stable disease (SD) versus complete or partial response (CR/PR)) and stage of disease (synchronous versus metachronous). See Appendix 8 for an overview of these subgroups and their event incidences. As illustrated in Figure 3, different levels of censoring $p_{censoring}$ ($p_{censoring} \approx 0\%, 10\%, 30\%, 60\%$) were applied for each subgroup analysis to assess the impact of this data characteristic on the modeling approaches' performance and health economic outcomes. Censoring was performed as for the simulation study. For each subgroup and $p_{censoring}$ combination, a probabilistic sensitivity analysis was performed based on 10,000 runs of 10,000 patients per treatment strategy in each run. As for the simulation study, performance in terms of RAID and KLD was calculated by comparing simulated samples $s_{simulation}$ for each modeling approach to corresponding the uncensored sample $s_{uncensored}$ for internal validation. Incremental cost-effectiveness planes were used to visualize potential differences in health economic outcomes.

RESULTS

In presenting the results, low, moderate, and high levels of censoring were defined as 10%, 30%, and at least 60% censoring, respectively.

Simulation study results

Results for selected scenarios are presented in Table 2, Figure 4, and Figure 5. Results for all scenarios are presented in Appendix 9. Regarding the impact of disease pathway and data characteristics on the RAID, several trends could be identified. Performance in terms of RAID was most strongly impacted by censoring, with higher levels of censoring leading to decreased performance. An increasing number of competing events decreased performance, whereas increasing sample size and overlap of competing time-to-event distributions both increased performance. These trends were observed both with regard to internal and external validation, which did not differ substantially in absolute terms.

No single modeling approach performed best in all scenarios in terms of RAID. The ESD and ESPD approaches performed similar for all scenarios. For a low level of distribution overlap and censoring, the MDR approach performed worse compared to all other approaches, which performed similar. However, no clear differences could be observed for a low level of censoring and high level of distribution overlap. For a moderate level of censoring, the UDR approach outperformed all other approaches for a low level of distribution overlap, whereas all approaches performed similar for a high level of distribution overlap. For higher levels of censoring, the MDR approach performed best, followed by the ESD and ESPD approaches. The 95% confidence intervals presented in Figure 4 were smaller for the UDR modeling approach in some scenarios, though this provided no additional insights.

Table 2. *Mean weighted performance of the modeling approaches over all simulation study runs.*

Number of Events	Distribution Overlap (%)	Sample Size	Censored Proportion (%)	Incidence of Events *Relative Absolute Difference (%)*							
				External Validation					Internal Validation		
				Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial
2	10	50	10	11,0	15,3	15,3	16,9	23,0	10,2	10,2	15,8
			30	10,0	25,2	25,2	12,9	31,1	25,4	25,4	15,0
			60	13,8	35,0	35,4	41,9	21,2	41,5	41,9	33,8
2	10	500	10	3,5	9,8	9,8	15,5	22,3	9,7	9,7	15,5
			30	3,5	26,4	26,4	14,3	33,4	26,5	26,5	14,3
			60	3,5	55,7	55,8	45,1	31,5	56,3	56,5	44,6
2	90	50	10	11,0	12,0	12,0	9,5	9,6	5,3	5,3	4,4
			30	11,1	15,3	15,3	9,4	9,1	11,6	11,6	10,4
			60	10,9	26,3	26,3	27,8	16,4	25,3	25,3	30,1
2	90	500	10	3,5	4,4	4,4	3,1	3,1	2,8	2,8	1,6
			30	3,5	6,5	6,5	7,9	6,0	5,7	5,7	7,9
			60	3,5	11,1	11,1	27,3	16,3	10,6	10,6	27,5
4	10	50	10	17,1	20,8	20,8	20,1	25,6	11,4	11,4	16,5
			30	14,5	28,3	28,3	19,9	34,2	23,2	23,2	22,0
			60	11,6	43,9	44,0	40,1	40,2	47,6	47,8	42,8
4	10	500	10	5,7	16,5	16,5	18,3	28,8	15,8	15,8	18,3
			30	4,8	23,1	23,1	19,2	35,1	22,9	22,9	19,2
			60	4,0	56,6	56,7	39,2	52,4	56,8	56,8	39,7
4	90	50	10	18,0	20,1	20,1	16,1	18,5	10,0	10,0	11,5
			30	17,8	25,8	25,8	23,4	25,3	20,9	20,9	25,4
			60	16,6	41,1	41,0	43,7	29,9	40,9	40,8	50,1
4	90	500	10	5,8	9,7	9,7	8,4	10,6	7,9	7,9	7,7
			30	5,9	16,9	16,9	24,1	24,7	16,0	16,0	24,6
			60	5,8	25,8	25,8	72,1	37,4	25,4	25,4	73,1

In general, increasing levels of censoring resulted in a decreased performance in terms of KLD. For an increasing number of competing events, performance decreased for low to moderate levels of censoring, whereas it increased for higher levels of censoring. Increasing levels of distribution overlap resulted in improved performance, especially for low to moderate levels of censoring. Increasing sample size improved performance, but only for high levels of distribution overlap. Performance in terms of internal validation was lower compared to external validation for a low level of censoring, but showed similar trends across scenarios.

	Time-to-Event Distributions Kullback-Leibler Divergence								
	External Validation					Internal Validation			
MDR vs. Trial	Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
22,6	0,071	0,246	0,054	0,202	0,132	0,225	0,045	0,194	0,114
33,6	0,071	0,395	0,261	0,211	0,301	0,357	0,217	0,199	0,256
26,0	0,122	0,388	1,278	0,316	1,368	0,377	0,730	0,358	0,914
22,4	0,010	0,215	0,006	0,192	0,089	0,206	0,009	0,177	0,085
33,4	0,010	0,248	0,036	0,160	0,073	0,240	0,039	0,145	0,071
32,1	0,045	0,296	1,418	0,574	1,283	0,270	0,968	0,594	1,039
4,4	0,069	0,039	0,087	0,032	0,039	0,050	0,079	0,047	0,051
8,6	0,070	0,064	0,294	0,051	0,114	0,073	0,277	0,065	0,115
17,4	0,068	0,181	0,948	0,154	0,508	0,188	0,877	0,157	0,476
1,6	0,010	0,008	0,028	0,006	0,009	0,012	0,033	0,010	0,013
6,0	0,010	0,011	0,212	0,009	0,061	0,015	0,217	0,014	0,060
16,5	0,010	0,021	0,832	0,028	0,332	0,025	0,819	0,033	0,318
22,7	0,089	0,575	0,173	0,581	0,469	0,568	0,158	0,577	0,447
35,0	0,088	0,556	0,222	0,502	0,418	0,546	0,186	0,501	0,382
48,4	0,080	0,350	0,423	0,223	0,452	0,341	0,281	0,220	0,320
28,8	0,015	0,586	0,304	0,581	0,515	0,564	0,269	0,559	0,481
35,0	0,013	0,545	0,147	0,504	0,373	0,527	0,131	0,484	0,347
53,0	0,037	0,278	0,707	0,136	0,604	0,287	0,439	0,149	0,422
13,8	0,137	0,097	0,219	0,065	0,118	0,115	0,204	0,108	0,151
24,6	0,126	0,134	0,506	0,102	0,340	0,150	0,457	0,145	0,344
32,9	0,110	0,296	1,030	0,241	0,833	0,303	0,885	0,259	0,747
10,0	0,018	0,050	0,115	0,032	0,071	0,054	0,117	0,041	0,077
24,8	0,022	0,048	0,503	0,057	0,406	0,053	0,433	0,078	0,382
37,7	0,024	0,072	0,940	0,370	1,164	0,080	0,828	0,390	1,105

In terms of KLD, for a low level of distribution overlap, the ESPD approach outperformed the other approaches strongly for a low level of censoring and somewhat for a moderate level of censoring. For high a level of distribution overlap, the ESD and UDR approaches outperformed the ESPD and MDR approaches slightly for a low level of censoring and more clearly for a moderate level of censoring. For high levels of censoring, the ESD and UDR approaches outperformed the ESPD and MDR approaches. No additional insights could be obtained from the 95% confidence intervals presented in Figure 5.

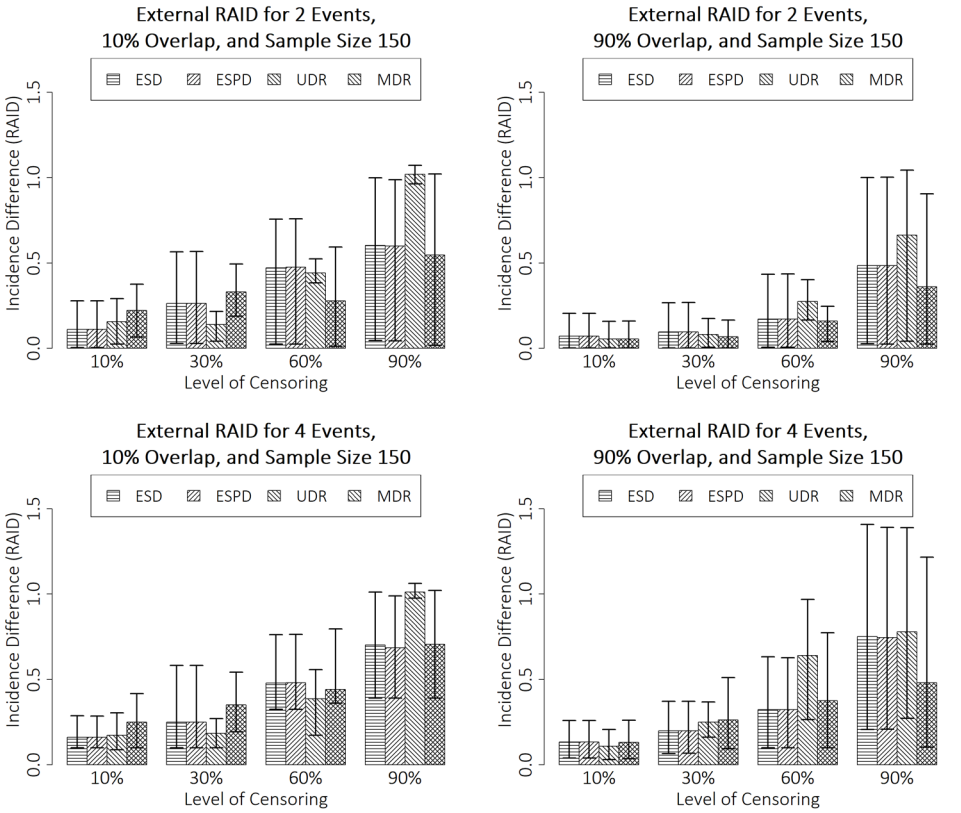


Figure 4. *External validation of the relative absolute incidence difference (RAID) for different levels of distribution overlap and censoring, based on selected experiments from the simulation study.*

Case study results

The internal validation in terms of RAID and KLD for the case study is presented in Appendix 10. Although no unambiguous trends could be observed, higher levels of censoring overall resulted in decreased performance in terms of RAID, with a stronger effect for lower sample sizes. Regarding the KLD, higher levels of censoring resulted in decreased performance in general, though this trend was not consistent over all subgroups and levels of censoring.

In terms of RAID, all approaches performed similar for moderate to high sample sizes ($n \geq 100$). However, especially for smaller samples and moderate or higher levels of censoring, the ESD and ESPD approaches generally outperformed the UDR and MDR approaches in terms of RAID. Higher levels of censoring resulted in decreased KLD performance for some subgroups, but resulted in increased performance for other subgroups. Noticeably, the ESPD approach performed substantially worse compared to the other approaches regarding the time-to-death in the experimental arm of Subgroup 6, which represents a small sample size ($n = 22$) with only 2 observations.

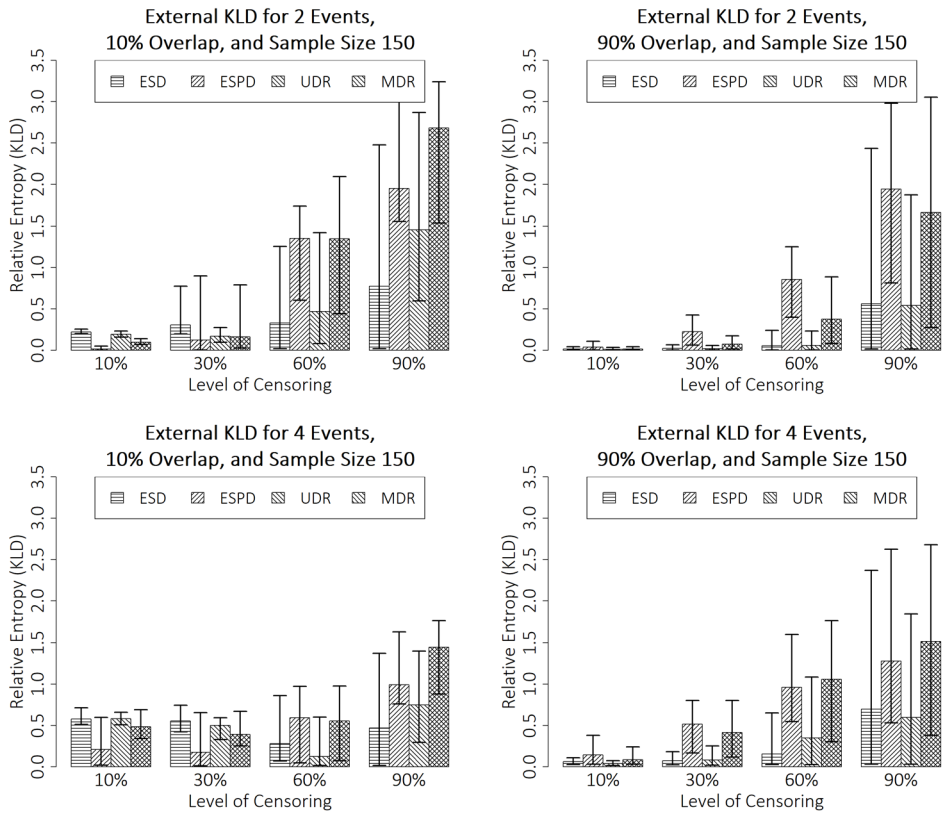


Figure 5. *External validation of the Kullback-Leibler divergence (KLD) for different levels of distribution overlap and censoring, based on selected experiments from the simulation study.*

Cost-effectiveness planes are presented for selected subgroup analyses in Figure 6 and for all subgroup analyses in Appendix 11. Although all modeling approaches yielded similar results for analyses involving moderate to high sample sizes ($n \geq 100$) and a low level of censoring, differences were observed for lower samples and higher levels of censoring. Decreasing sample size and increasing censoring resulted in increased uncertainty surrounding cost-effectiveness estimates. Overall, the ESPD and MDR approaches yielded similar results, as did the ESD and UDR approaches.

DISCUSSION

Previous research on different approaches for implementing competing events in DES models argued a limited generalizability of recommendations for scenarios without censoring to those with censoring due to anticipated differences in performance following different implementations of the modeling approaches [11]. The ESPD modeling approach indeed was implemented differently in this study, because it is not possible to estimate event-specific

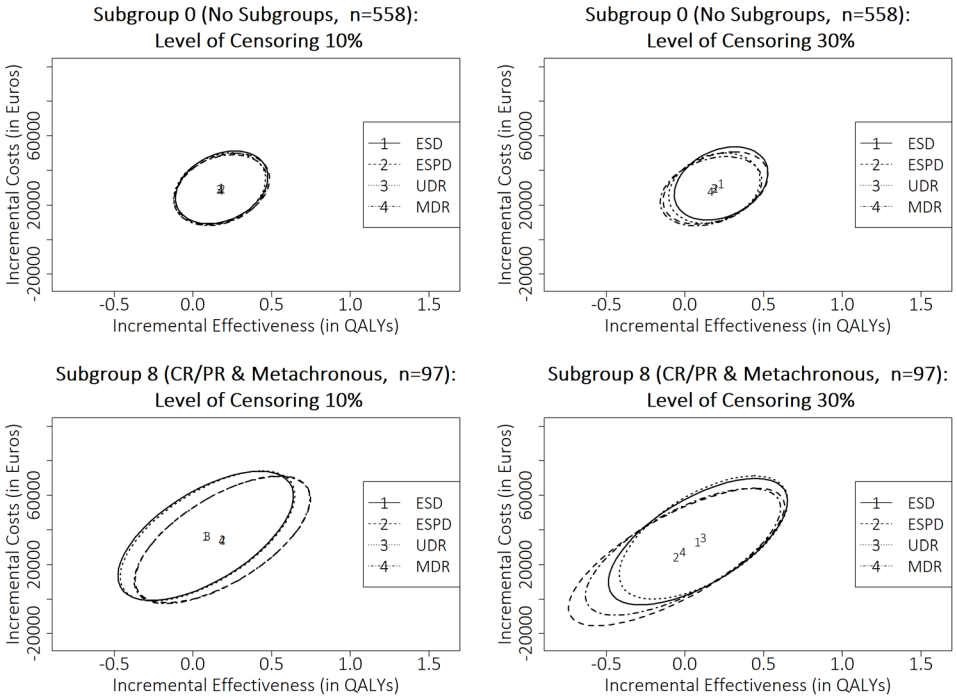


Figure 6. *Cost-effectiveness planes for selected levels of censoring and selected case study subgroup analyses.*

probabilities directly from censored data as can be done for uncensored data. Furthermore, parameters of mixture distribution used for the MDR approach were obtained differently in this study compared to the study based on uncensored data.

It is widely known that censoring needs to be accounted for and results from this study substantiate this for DES models that include competing events. Compared to scenarios without censoring, the presence of right-censoring negatively impacted performance of all modeling approaches. However, different from other disease pathway and data characteristics, increasing or decreasing levels of right-censoring did not have a clear, unambiguous impact on performance. Additionally, the case study illustrated that right-censoring may add to the uncertainty surrounding cost-effectiveness point-estimates and may result in differences between modeling approaches in terms of cost-effectiveness point-estimates when there is a substantial level of censoring (i.e., 30%).

There is no single answer to the question which modeling approach performed best in the simulation study. The ESD and ESPD approaches performed similar in terms of RAID, which is not surprising given that the ESPD approach uses event-specific probabilities estimated according to distribution parameters $\varphi_K^{CS} = \left\{\varphi_1^{CS},...,\varphi_k^{CS}\right\}$ assuming a CS hazards model, which are also used in the ESD approach. Considering both the RAID and KLD, the ESPD approach overall performed best for scenarios with a low level of censoring (i.e., 10%) and low distri-

bution overlap. For low levels of censoring combined with high levels of distribution overlap, and for moderate levels of censoring (i.e., 30%), UDR was the overall best performing approach. For high levels of censoring (i.e., ≥ 60%), the ESD approach performed best overall.

When considering censoring in addition to other data characteristics and disease pathway characteristics, it appears that making a single recommendation about which modeling approach should be used to implement competing events in DES models is not possible. According to the Professional Society for Health Economics and Outcomes Research (ISPOR) and Society for Medical Decision Making (SMDM) modeling good research practices guidelines [31], the preferred modeling approach is based on the strategy of first selecting the time-to-event from a joint time-to-event distribution and then selecting the corresponding event, which corresponds to the UDR and MDR approaches. Although the MDR approach as implemented in this study does not perform well in presence of right-censoring, the UDR approach may be considered as preferable approach for scenarios with low levels of censoring and high distribution overlap, and for scenarios with moderate levels of censoring. For scenarios including a low level of censoring and distribution overlap, use of the ESPD approach may be considered preferable. It cannot be stated that the UDR and ESPD approaches are always the best option in these scenarios, as the case study showed that modeling approaches performed differently for subgroups with similar disease pathway and data characteristics at the same level of censoring. Therefore, decision analytic modelers should be aware of alternatives that are available and that these might result in different levels of performance and health economic outcomes. Internal validity of candidate modeling approaches should be assessed to inform decisions about which approach to select for a specific case study, and these efforts should be clearly communicated about.

Internal validation can be challenging in the presence of censoring, however, since performance measures that assess performance of the modeling approaches as a whole, such as the RAID and KLD, cannot be applied to censored data. Although less straightforward to interpret and challenging to combine into one performance measure for modeling approaches as a whole, alternative measures are available to assess discrimination and calibration of single survival models, while accounting for censoring [32]. An example of such a measure is Demler et al.'s Greenwood-Nam-D'Agostino statistic [33], which is a modification of Hosmer-Lemeshow's statistic by Nam and D'Agostino [34]. Another point of consideration is whether multivariable models are being developed to reflect patient heterogeneity, so that patient characteristics or treatment histories influence the occurrence and timing of events. Multivariable models are most easily implemented for the UDR approach, especially for scenarios including more than two competing events, as the variable selection procedure, for example, only needs to be performed for one survival model and one (multinomial) logistic regression model.

This study has certain limitations. Only non-informative and right-censored data were considered. Hence, results and recommendations cannot be generalized to scenarios involving informative, interval-, or left-censored data. Generalizability is also limited by the deliberate decision to use Weibull distributions in the simulation study. Although this allowed for unbi-

ased comparison of the modeling approaches, underlying distributions are unknown in practice and performance of the approaches may vary depending on the flexibility of selected distributions to describe the data. Also, mixtures of Weibull distributions were only used for the MDR approach, whereas such multimodal distributions can also be used in combination with the other modeling approaches. Furthermore, mixture distributions were defined by parameters estimated for the ESPD approach due to issues with the maximum likelihood estimation, but there might be alternative ways to estimate these mixtures by maximum likelihood, which may result in a different performance of the MDR approach. Finally, patient heterogeneity was not considered, which limits generalization to scenarios involving multivariable (survival) models.

CONCLUSION

Modeling approaches to implement competing events in DES models perform different when these are utilizing non-informative right-censored data compared to uncensored data. Although substantial differences in the performance of the modeling approaches were observed for different scenarios of data and disease pathway characteristics, no overall best performing approach(es) could be identified. For scenarios including low levels of censoring and distribution overlap, use of the ESPD approach seems to be preferred. In presence of low levels of censoring and high distribution overlap, and for moderate levels of censoring, use of the UDR approach seems to be preferred, which is in line with general modeling good practices guidelines. However, these recommendations are limited in their generalizability, as the case study showed. Modelers should therefore base their decision on which modeling approach to implement foremost on internal validity estimates and clearly communicate about their validation efforts.

REFERENCES

1. Annemans L, et al. Current methodological issues in the economic assessment of personalized medicine. Value Health. 2013;16(S6):S20-6.

2. Caro JJ, et al. Discrete event simulation: the preferred technique for health economic evaluations? Value Health. 2010;13(8):1056-60.

3. Miller JD, et al. Current challenges in health economic modeling of cancer therapies: a research inquiry. Am Health Drug Benefits. 2014;7(3):153-62.

4. Caro JJ, et al. Modeling good research practices—overview: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Value Health. 2012;15(6):796-803.

5. Degeling K, et al. Matching the model with the evidence: comparing discrete event simulation and state-transition modeling for time-to-event predictions in a cost-effectiveness analysis of treatment in metastatic colorectal cancer patients. Cancer Epidemiol. 2018;57:60-7.

6. Karnon J, Haji Ali Afzali H. When to use discrete event simulation (DES) for the economic evaluation of health technologies? A review and critique of the costs and benefits of DES. Pharmacoeconomics. 2014;32(6):547-58.

7. Standfield L, et al. Markov modeling and discrete event simulation in health care: a systematic comparison. Int J Technol Assess Health Care. 2014;30(2):165-72.

8. Caro JJ, Moller J. Decision-analytic models: current methodological challenges. Pharmacoeconomics. 2014;32(10):943-50.

9. Pintilie M. Competing risks: a practical perspective. Hoboken, NJ: Wiley; 2006.

10. Barton P, et al. The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis. Health Technol Assess. 2004;8(11):104.

11. Degeling K, et al. Comparing strategies for modeling competing risks in discrete-event simulations: a simulation study and illustration in colorectal cancer. Med Decis Making. 2019;39(1):57-73.

12. Donoghoe MW, Gebski V. The importance of censoring in competing risks analysis of the subdistribution hazard. BMC Med Res Methodol. 2017;17(1):52.

13. Leung K-M, et al. Censoring issues in survival analysis. Annu Rev Public Health. 1997;18(1):83-104.

14. R Core Team. R: a language and environment for statistical computing. Version 3.4.1 [software]. Available from: https://www.r-project.org/

15. Austin PC, et al. Introduction to the analysis of survival data in the presence of competing risks. Circulation. 2016;133(6):601-9.

16. Latouche A, et al. Misspecified regression model for the subdistribution hazard of a competing risk. Stat Med. 2007;26(5):965-74.

17. Misspecified regression model for the subdistribution hazard of a competing risk by Latouche A, Boisson V, Chevret S and Porcher R. Stat Medicine. 2007;26(7):1649-51.

18. Jeong J-H, Fine J. Direct parametric inference for the cumulative incidence function. J R Stat Soc Ser C Appl Stat. 2006;55(2):187-200.

19. Mozumder SI, et al. Direct likelihood inference on the cause-specific cumulative incidence function: A flexible parametric regression modelling approach. Stat Med. 2018;37(1):82-97.

20. Lau B, et al. Competing risk regression models for epidemiologic data. Am J Epidemiol. 2009;170(2):244-56.

21. Bordes L, Chauveau D. Stochastic EM algorithms for parametric and semiparametric mixture models for right-censored lifetime data. Comput Stat. 2016;31(4):1513-38.

22. Davenport JW, et al. Parameter estimation for finite mixture distributions. Comput Math Appl. 1988;15(10):819-28.

23. Cover TM, Thomas JA. Elements of information theory. 2nd ed. Hoboken, NJ: John Wiley; 2006.

24. Benaglia T, et al. mixtools: An R package for analyzing finite mixture models. J Stat Softw. 2009;32(6):1-29.

25. Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York, NY: Springer; 2002.

26. Grün B, Leisch F. Fitting finite mixtures of generalized linear regressions in R. Comput Stat Data Anal. 2007;51(11):5247-52.

27. Grün B, Leisch F. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J Stat Softw. 2008;28(4):35.

28. Leisch F. FlexMix: a general framework for finite mixture models and latent class regression in R. J Stat Softw. 2004;11(8):18.

29. Simkens LHJ, et al. Maintenance treatment with capecitabine and bevacizumab in metastatic colorectal cancer (CAIRO3): a phase 3 randomised controlled trial of the Dutch Colorectal Cancer Group. Lancet. 2015;385(9980):1843-52.

30. Franken MD, et al. Cost-effectiveness of capecitabine and bevacizumab maintenance treatment after first-line induction treatment in metastatic colorectal cancer. Eur J Cancer. 2017;75:204-12.

31. Karnon J, et al. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. Value Health. 2012;15(6):821-7.

32. Rahman MS, et al. Review and evaluation of performance measures for survival prediction models in external validation settings. BMC Med Res Methodol. 2017;17(1):60-.

33. Demler OV, et al. Tests of calibration and goodness-of-fit in the survival setting. Stat Med. 2015;34(10):1659-80.

34. D'Agostino RB, Nam B-H. Evaluation of the performance of survival analysis models: discrimination and calibration measures. Handbook of Statistics. 2003;231-25.

APPENDIX 1: pseudo-algorithm for the ESD modeling approach

Data Analysis (see Table 1):

- For each competing event j, $j = 1,...,k$, fit a time-to-event distribution D_j:
  - For each individual i, $i = 1,...,n$, define observed censoring indicator c_{ij} that indicates whether individual i experienced event j (c_{ij} =1) or not (c_{ij} =0)
  - Parameterize cause-specific likelihood function $L_j^{CS}(t_1,...,t_n \mid \varphi_j^{CS})$ according to a specific distribution type (Equation 11)
  - Estimate parameters φ_j^{CS} that define distribution D_j by maximum likelihood

Simulation:

- Obtain time-to-events for each competing event:
  - Draw a time t_j for each competing event j, $j = 1,...,k$ by performing a random draw from the corresponding distribution D_j
- Select the competing event to occur:
  - Select event j with the lowest time-to-event (i.e., the first event to occur)
  - Simulate the selected event j at the corresponding time t_j

APPENDIX 2: pseudo-algorithm for the ESPD modeling approach

Data Analysis (see Table 1):

- For each competing event j, $j = 1,...,k$, estimate event-specific probabilities $\Pr(C_j = 1)$:
  - Estimate parameters φ_K^{CS} by maximum likelihood as for the ESD modeling approach (Appendix 1)
  - Parameterize cause-specific cumulative incidence function $F_j^{CS}(t)$ (Equation 7) using φ_K^{CS} and estimate event-specific probabilities $\Pr(C_j = 1)$ using Equation 14
- For all competing events $K = \{1,...,k\}$ simultaneously, fit time-to-event distributions D_K
  - For each individual i, define observed censoring indicators c_{ij} that indicates whether individual i experienced event j, $j = 1,...,k$ (c_{ij} =1) or not (c_{ij} =0)
  - Parameterize subdistribution hazards likelihood function $L^{SD}(t_1,...,t_n \mid \varphi_K^{SD})$ according to a specific distribution type (Equation 17)
  - Estimate parameters $\varphi_K^{SD} = \{\varphi_1^{SD},...,\varphi_k^{SD}\}$ that define distributions D_K by maximum likelihood

Simulation:

- Select the competing event to occur:
  - Draw a random number from a Uniform distribution $U(0,1)$
  - Select event j to occur using event probabilities $\Pr(C_j = 1)$, $j = 1,...,k$ and the random number
- Obtain a time-to-event for the selected event:
  - Draw a time t for selected event j by performing a random draw from the corresponding distribution D_j parameterized by φ_j^{SD}
- Simulate the selected event j at time t

APPENDIX 3: pseudo-algorithm for the UDR modeling approach

Data Analysis (see Table 1):

- For all competing events combined, fit a single unimodal time-to-event distribution D:
  - For each individual i, $i=1,...,n$, define observed censoring indicator c_i so that individual i experienced any event (c_i =1) or is censored (c_i =0)
  - Parameterize simplified likelihood function $L^{CS|UDR}(t_1,...,t_n \mid \varphi^{CS})$ (Equation 18) or $L^{SD|UDR}(t_1,...,t_n \mid \varphi^{SD})$ (Equation 19) according to a specific unimodal distribution type
  - Estimate parameters $\varphi_{unimodal}$ that define distribution D by maximum likelihood
- Fit a (multinomial) logistic regression model to predict the competing event to occur:
  - Fit (multinomial) logistic regression model $r(t)$ that predicts probabilities $\Pr(C_j=1 \mid T=t)$ of each competing event j, $j=1,...,k$ to occur (dependent variable) based on time t (independent variable)

Simulation:

- Obtain a time-to-event for an event to occur:
  - Draw a time t for the event to occur by performing a random draw from unimodal distribution D
- Select the competing event to occur:
  - Obtain probabilities $\Pr(C_j=1 \mid T=t)$ for each competing event j, $j=1,...,k$ to occur based on time-to-event t, using (multinomial) logistic regression model $r(t)$
  - Draw a random number from a Uniform distribution $U(0,1)$
  - Select event j to occur using event probabilities $\Pr(C_j=1 \mid T=t)$ and the random number
- Simulate selected event j at time t

APPENDIX 4: pseudo-algorithm for the MDR modeling approach

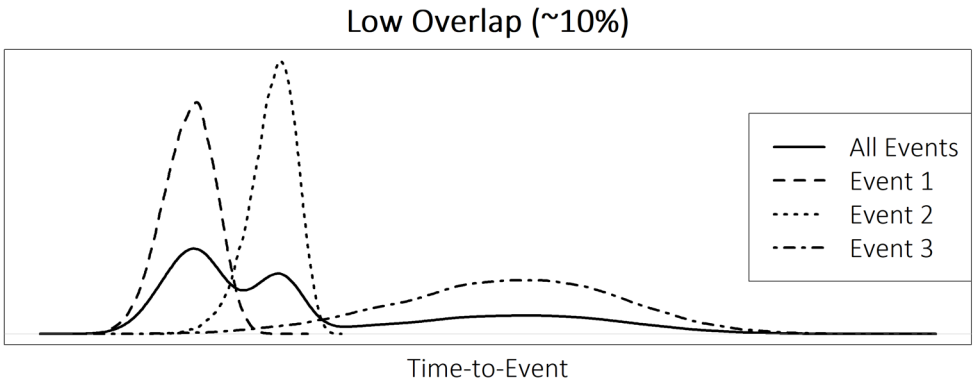
Data Analysis (see Table 1):

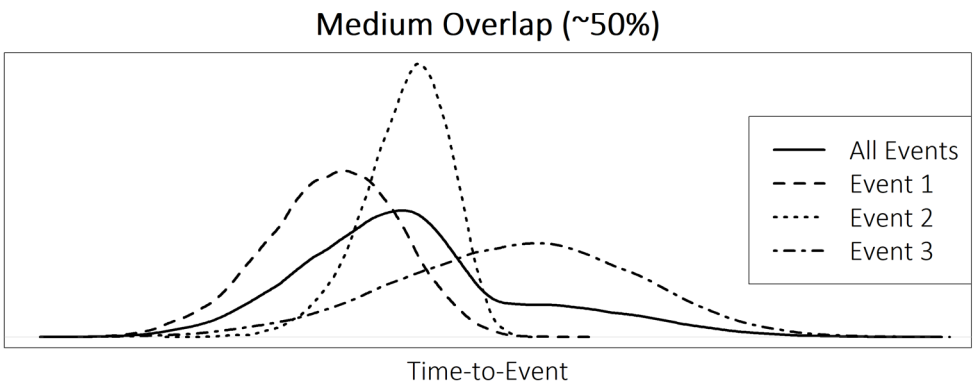
- For all competing events combined, define a single multimodal mixture time-to-event distribution D :
  - Define mixing proportions $\alpha_M = \{\alpha_1,...,\alpha_m\}$ based on event-specific probabilities $\Pr(C_j = 1)$, $j = 1,...,k$ estimated according to the ESPD modeling approach (where $m = k$)
  - Define mixture distribution parameters $\varphi_M = \{\varphi_1,...,\varphi_m\}$ based on event-specific distributions parameters $\varphi_K^{SD} = \{\varphi_1^{SD},...,\varphi_k^{SD}\}$ estimated according to the ESPD modeling approach (where $m = k$)
- Fit a (multinomial) logistic regression model to predict the competing event to occur:
  - Fit (multinomial) logistic regression model $r(t)$ that predicts probabilities $\Pr(C_j = 1 \mid T = t)$ of each competing event j , $j = 1,...,k$ to occur (dependent variable) based on time t (independent variable)

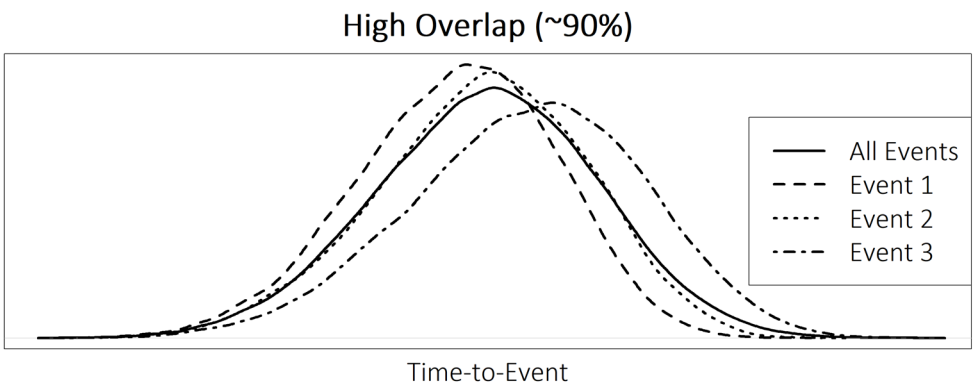
Simulation:

- Obtain a time-to-event for an event to occur:
  - Draw a time t for the event to occur by performing a random draw from multimodal distribution D
- Select the competing event to occur:
  - Obtain probabilities $\Pr(C_j = 1 \mid T = t)$ for each competing event j , $j = 1,...,k$ to occur based on time-to-event t , using (multinomial) logistic regression model $r(t)$
  - Draw a random number from a Uniform distribution $U(0,1)$
  - Select event j to occur using event probabilities $\Pr(C_j = 1 \mid T = t)$ and the random number
- Simulate selected event j at time t

APPENDIX 5: probability density functions illustrating different levels of distribution overlap for three competing events (see Apendix 6 for the methods used to calculate the percentage of overlap)







APPENDIX 6: calculation of the overlap between distributions

No gold standard approach for calculating the overlap between multiple (competing) distributions is available. The approach used in the current study defines the overlap between distributions as the mean overlap percentage over all distribution-specific overlaps, for which the distribution-specific overlap is defined as the percentage of the distribution that is overlapping with any other distribution. More precisely, the calculation can be performed according to the following steps:

1. Define a pre-specified number of equally sized bins that cover the observations of all distributions
2. Calculate the density of all distributions for each bin (1)
3. For each distribution, calculate the "overlap density" in each of the bins (1) by determining how much the density (2) overlaps with the corresponding highest density of the other distributions
4. Calculate the distribution-specific overlap by dividing the area under the "overlap density" curve (3) by the area under overall density curve (2)
5. (Calculate the overall overlap between all distributions as the average of the distribution-specific overlaps (4)

R code used to calculate the overlap between competing time-to-event distributions is provided below.

```
# This function calculates the overlap between competing
# time-to-event distributions based on a 'data' object
# containing two columns, i.e. 'time' and 'event'.

getOverlap <- function(data, n.bins=512, digits=3) {

 # Function inputs:
 # - data:    data.frame with two columns:
 #          + time:    column containing the times
 #          + event:   column containing the events corresponding to the times
 # - n.bins:  the number of bins used for the 'density' function
 # - digits:  the number of digits to round the results to

 # Load package for calculating the Area Under the Curve (AUC)
 library(flux);

 # Get the events in the data
 events <- sort(unique(data$event));
 n.events <- length(events);

 # Get event-specific times
 times <- lapply(events, function(e) {
  subset(data, event==e)$time;
```

```
 });

 # Density settings
 min <- min(data$time);
 max <- max(data$time);
 n <- n.bins;
 x <- seq(from=min, to=max, length.out=n);

 # Get event-specific densities
 y <- sapply(1:n.events, function(i1) {
  density(times[[i1]], from=min, to=max, n=n)$y;
 });

 # Get event-specific density overlap
 yo <- sapply(1:n.events, function(i2) {

  sapply(1:n, function(j){

   own <- y[j,i2];
   other <- max(y[j,which(c(1:n.events)!=i2)]);

   if(own<=0) {
    return(0)
   } else
    return(min(own,other))

  })
 });

 # Calculate event-specific overlap
 o <- sapply(1:n.events, function(i3) {
  (auc(x=x, y=yo[,i3]) / auc(x=x, y=y[,i3])) * 100;
 });

 return(round(mean(o), digits=digits));
}
```

APPENDIX 7: exact definitions of the populations used in the simulation study

	Number of Competing Risks	Distribution Overlap (%)	Event 1			Event 2			Event 3			Event 4		
			Incidence	Shape	Scale	Incidence	Shape	Scale	Incidence	Shape	Scale	Incidence	Shape	Scale
Population 1	2	10	0.61	5	250	0.39	10	418	-	-	-	-	-	-
Population 2	2	50	0.61	5	250	0.39	10	302	-	-	-	-	-	-
Population 3	2	90	0.61	5	250	0.39	5	265	-	-	-	-	-	-
Population 4	3	10	0.41	5	250	0.25	10	418	0.34	4.5	950	-	-	-
Population 5	3	50	0.41	5	250	0.25	10	302	0.34	4.5	409	-	-	-
Population 6	3	90	0.41	5	250	0.25	5	265	0.34	5	291	-	-	-
Population 7	4	10	0.41	5	250	0.20	10	418	0.29	4.5	950	0.10	7	2070
Population 8	4	50	0.41	5	250	0.20	10	302	0.29	4.5	409	0.10	7	637
Population 9	4	90	0.41	5	250	0.20	5	265	0.29	5	291	0.10	5	337

APPENDIX 8: exact definitions of the populations used in the simulation study

Subgroup	Sample Size	Event Observations				
		Progression *(Post-induction)*	Death *(Post-induction)*	Progression *(Re-induction)*	Death *(Re-induction)*	Death *(salvage)*
Subgroup 0: No Subgroups	558	519	39	370	109	410
Observation	279	270	9	200	54	216
Maintenance	279	249	30	170	55	194
Subgroup 1: SD	191	185	6	128	43	142
Observation	95	95	0	68	22	73
Maintenance	96	90	6	60	21	69
Subgroup 2: CR/PR	367	334	33	242	66	268
Observation	184	175	9	132	32	143
Maintenance	183	159	24	110	34	125
Subgroup 3: Synchronous	410	382	28	275	82	300
Observation	191	186	5	143	33	153
Maintenance	219	196	23	132	49	147
Subgroup 4: Metachronous	147	137	10	95	27	110
Observation	88	84	4	57	21	63
Maintenance	59	53	6	38	6	47
Subgroup 5: SD & Synchronous	141	136	5	94	34	102
Observation	67	67	0	49	15	52
Maintenance	74	69	5	45	19	50
Subgroup 6: SD & Metachronous	50	49	1	34	9	40
Observation	28	28	0	19	7	21
Maintenance	22	21	1	15	2	19
Subgroup 7: CR/PR & Synchronous	269	246	23	181	48	198
Observation	124	119	5	94	18	101
Maintenance	145	127	18	87	30	97
Subgroup 8: CR/PR & Metachronous	97	88	9	61	18	70
Observation	60	56	4	38	14	42
Maintenance	37	32	5	23	4	28

APPENDIX 9: detailed simulation study results on performance

See subsequent pages for the detailed simulation study results on performance.

Number of Events	Distribution Overlap (%)	Sample Size	Censored Proportion (%)	Incidence of Events Relative Absolute Difference (%)							
				External Validation					Internal Validation		
				Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial
2	10	50	0	11.0	13.4	13.4	11.9	13.1	7.2	7.2	6.2
			10	11.0	15.3	15.3	16.9	23.0	10.2	10.2	15.8
			30	10.0	25.2	25.2	12.9	31.1	25.4	25.4	15.0
			60	13.8	35.0	35.4	41.9	21.2	41.5	41.9	33.8
			90	12.4	67.0	66.6	92.3	58.6	65.4	65.4	87.0
2	10	150	0	6.4	9.2	9.2	6.9	8.7	6.9	6.9	3.8
			10	6.3	11.1	11.1	15.6	22.3	9.8	9.8	15.6
			30	6.4	26.4	26.4	13.9	33.2	26.6	26.6	14.1
			60	6.6	47.2	47.6	44.2	27.8	49.8	50.3	41.8
			90	6.6	60.3	60.0	102.0	54.7	60.0	59.8	100.2
2	10	500	0	3.5	7.1	7.1	4.3	6.3	6.8	6.8	3.0
			10	3.5	9.8	9.8	15.5	22.3	9.7	9.7	15.5
			30	3.5	26.4	26.4	14.3	33.4	26.5	26.5	14.3
			60	3.5	55.7	55.8	45.1	31.5	56.3	56.5	44.6
			90	3.5	59.4	59.4	103.4	47.4	59.5	59.4	102.8
2	50	50	0	11.1	11.2	11.2	11.1	11.2	1.5	1.5	2.0
			10	11.0	11.7	11.7	11.2	12.8	4.4	4.4	8.2
			30	10.9	15.5	15.6	5.8	12.7	11.4	11.6	8.2
			60	11.0	29.6	29.9	42.2	16.9	29.4	29.7	40.7
			90	12.0	65.5	65.0	88.8	53.7	65.0	64.8	85.2
2	50	150	0	6.3	6.5	6.5	6.5	6.4	1.2	1.2	1.7
			10	6.4	7.0	7.0	8.8	10.5	2.7	2.7	7.9
			30	6.4	9.8	9.8	3.6	11.5	7.7	7.7	4.9
			60	6.3	24.7	24.9	44.0	14.9	24.5	24.6	44.2
			90	6.5	58.8	58.6	101.1	51.0	59.0	58.8	99.9
2	50	500	0	3.5	3.7	3.7	3.7	3.6	1.1	1.1	1.5
			10	3.5	3.9	3.9	7.9	9.8	1.8	1.8	7.9
			30	3.5	6.6	6.6	2.5	11.2	5.8	5.8	3.1
			60	3.5	19.1	19.1	44.2	9.7	18.9	19.0	44.4
			90	3.5	55.4	55.3	103.3	46.0	55.5	55.5	103.1

	Time-to-Event Distributions								
	Kullback-Leibler Divergence								
	External Validation					Internal Validation			
	MDR vs. Trial	Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
	6.5	0.072	0.239	0.043	0.120	0.057	0.225	0.041	0.115	0.046
	22.6	0.071	0.246	0.054	0.202	0.132	0.225	0.045	0.194	0.114
	33.6	0.071	0.395	0.261	0.211	0.301	0.357	0.217	0.199	0.256
	26.0	0.122	0.388	1.278	0.316	1.368	0.377	0.730	0.358	0.914
	55.2	0.117	1.251	2.240	1.253	2.305	1.472	2.035	1.315	2.125
	6.0	0.027	0.221	0.013	0.099	0.017	0.210	0.018	0.090	0.020
	22.4	0.027	0.223	0.017	0.195	0.099	0.209	0.020	0.177	0.092
	33.5	0.028	0.307	0.125	0.171	0.163	0.289	0.114	0.154	0.146
	29.7	0.076	0.329	1.353	0.471	1.347	0.307	0.825	0.507	0.997
	54.2	0.081	0.775	1.952	1.454	2.684	0.938	1.399	1.498	2.206
	5.8	0.010	0.216	0.004	0.093	0.006	0.207	0.008	0.085	0.009
	22.4	0.010	0.215	0.006	0.192	0.089	0.206	0.009	0.177	0.085
	33.4	0.010	0.248	0.036	0.160	0.073	0.240	0.039	0.145	0.071
	32.1	0.045	0.296	1.418	0.574	1.283	0.270	0.968	0.594	1.039
	47.3	0.051	0.649	1.924	1.750	2.602	0.726	1.410	1.758	2.235
	1.1	0.073	0.036	0.044	0.047	0.040	0.055	0.042	0.062	0.047
	9.9	0.074	0.045	0.069	0.051	0.051	0.058	0.067	0.064	0.054
	12.6	0.072	0.093	0.200	0.070	0.129	0.105	0.205	0.081	0.128
	17.6	0.082	0.299	1.039	0.265	1.000	0.296	0.841	0.278	0.807
	51.6	0.112	1.248	2.284	0.990	2.316	1.248	1.956	1.018	2.042
	0.9	0.028	0.017	0.013	0.026	0.014	0.026	0.019	0.035	0.022
	9.8	0.028	0.020	0.026	0.034	0.024	0.029	0.034	0.037	0.027
	11.5	0.028	0.039	0.123	0.044	0.063	0.050	0.137	0.046	0.066
	15.1	0.050	0.171	1.113	0.268	1.203	0.178	0.825	0.274	0.911
	50.8	0.079	0.694	2.127	1.285	2.791	0.736	1.498	1.287	2.262
	0.8	0.011	0.011	0.005	0.020	0.007	0.014	0.008	0.024	0.011
	9.9	0.011	0.011	0.014	0.029	0.016	0.016	0.019	0.028	0.016
	11.3	0.011	0.018	0.095	0.035	0.038	0.024	0.105	0.034	0.039
	9.9	0.029	0.073	1.069	0.242	1.208	0.076	0.809	0.245	0.940
	46.1	0.050	0.517	2.127	1.591	2.848	0.537	1.564	1.565	2.395

Number of Events	Distribution Overlap (%)	Sample Size	Censored Proportion (%)	Incidence of Events Relative Absolute Difference (%)							
				External Validation					Internal Validation		
				Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial
2	90	50	0	11.0	11.0	11.0	11.1	11.0	0.9	0.9	0.8
			10	11.0	12.0	12.0	9.5	9.6	5.3	5.3	4.4
			30	11.1	15.3	15.3	9.4	9.1	11.6	11.6	10.4
			60	10.9	26.3	26.3	27.8	16.4	25.3	25.3	30.1
			90	10.9	60.3	60.3	61.9	36.6	61.6	61.6	65.0
2	90	150	0	6.4	6.4	6.5	6.4	6.5	0.8	0.8	0.8
			10	6.4	7.3	7.3	5.6	5.6	3.5	3.6	2.6
			30	6.5	9.7	9.7	8.2	6.7	7.5	7.6	8.4
			60	6.4	17.1	17.1	27.6	16.1	16.2	16.2	28.0
			90	6.4	48.6	48.6	66.5	36.1	48.8	48.8	67.8
2	90	500	0	3.5	3.5	3.6	3.6	3.6	0.8	0.8	0.8
			10	3.5	4.4	4.4	3.1	3.1	2.8	2.8	1.6
			30	3.5	6.5	6.5	7.9	6.0	5.7	5.7	7.9
			60	3.5	11.1	11.1	27.3	16.3	10.6	10.6	27.5
			90	3.5	31.8	31.8	68.7	34.4	31.8	31.8	68.8
3	10	50	0	15.8	19.1	19.1	22.1	20.7	9.8	9.8	22.1
			10	15.8	21.4	21.4	24.3	30.3	13.8	13.8	22.9
			30	14.1	36.2	36.2	24.4	40.1	34.2	34.2	25.9
			60	12.1	41.1	41.2	35.4	36.9	44.2	44.4	37.9
			90	11.5	63.8	63.5	80.0	50.1	66.6	67.0	76.4
3	10	150	0	9.2	12.9	12.9	19.9	13.8	9.4	9.4	20.3
			10	9.1	15.9	15.9	22.3	28.7	13.4	13.4	22.2
			30	8.9	43.4	43.4	22.2	45.2	43.6	43.7	23.0
			60	6.8	48.1	48.1	34.9	43.7	49.3	49.4	36.0
			90	6.3	66.2	66.0	93.8	46.6	67.5	67.5	92.9
3	10	500	0	5.0	10.0	10.0	18.6	10.6	9.3	9.3	18.8
			10	5.0	13.7	13.7	22.0	28.3	13.3	13.3	22.0
			30	5.0	47.0	47.0	21.8	46.8	47.3	47.3	22.1
			60	4.3	60.0	60.0	36.4	55.6	59.9	60.0	37.0
			90	3.5	72.2	72.1	95.9	47.2	72.7	72.7	95.7

	Time-to-Event Distributions								
	Kullback-Leibler Divergence								
	External Validation					Internal Validation			
MDR vs. Trial	Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
0.8	0.070	0.033	0.043	0.033	0.034	0.049	0.038	0.047	0.046
4.4	0.069	0.039	0.087	0.032	0.039	0.050	0.079	0.047	0.051
8.6	0.070	0.064	0.294	0.051	0.114	0.073	0.277	0.065	0.115
17.4	0.068	0.181	0.948	0.154	0.508	0.188	0.877	0.157	0.476
38.9	0.086	1.032	2.074	0.801	1.825	1.011	1.807	0.777	1.642
0.8	0.026	0.013	0.013	0.012	0.012	0.022	0.017	0.021	0.021
2.6	0.026	0.016	0.041	0.012	0.017	0.024	0.047	0.022	0.024
6.6	0.026	0.023	0.227	0.019	0.073	0.032	0.232	0.029	0.074
16.4	0.026	0.057	0.855	0.061	0.377	0.068	0.827	0.068	0.356
36.7	0.043	0.563	1.945	0.541	1.665	0.562	1.719	0.533	1.494
0.8	0.010	0.007	0.005	0.006	0.005	0.011	0.008	0.010	0.010
1.6	0.010	0.008	0.028	0.006	0.009	0.012	0.033	0.010	0.013
6.0	0.010	0.011	0.212	0.009	0.061	0.015	0.217	0.014	0.060
16.5	0.010	0.021	0.832	0.028	0.332	0.025	0.819	0.033	0.318
34.4	0.020	0.171	1.872	0.252	1.460	0.176	1.724	0.256	1.338
11.9	0.093	0.618	0.086	0.223	0.086	0.612	0.075	0.228	0.068
28.9	0.090	0.604	0.094	0.566	0.420	0.591	0.083	0.562	0.397
40.3	0.089	0.581	0.300	0.477	0.462	0.566	0.244	0.475	0.410
44.6	0.084	0.376	0.524	0.228	0.540	0.364	0.339	0.222	0.369
51.0	0.101	0.648	1.266	0.528	1.263	0.666	1.082	0.529	1.127
10.2	0.033	0.606	0.039	0.201	0.035	0.586	0.041	0.192	0.033
28.5	0.032	0.591	0.048	0.559	0.388	0.570	0.050	0.536	0.369
45.2	0.042	0.608	0.341	0.452	0.457	0.585	0.252	0.438	0.390
46.0	0.058	0.298	0.734	0.132	0.683	0.302	0.413	0.131	0.416
47.7	0.059	0.367	1.171	0.607	1.440	0.376	0.801	0.617	1.212
9.8	0.012	0.602	0.028	0.196	0.022	0.586	0.030	0.187	0.022
28.3	0.012	0.587	0.036	0.557	0.379	0.570	0.037	0.537	0.366
46.9	0.021	0.633	0.318	0.447	0.445	0.614	0.241	0.434	0.382
55.9	0.045	0.301	0.850	0.175	0.720	0.309	0.523	0.190	0.503
47.7	0.037	0.316	1.174	0.899	1.434	0.311	0.827	0.896	1.278

Number of Events	Distribution Overlap (%)	Sample Size	Censored Proportion (%)	Incidence of Events Relative Absolute Difference (%)							
				External Validation					Internal Validation		
				Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial
3	50	50	0	15.9	17.0	17.0	15.8	16.9	4.1	4.1	7.3
			10	15.9	21.1	21.1	21.5	24.2	13.2	13.2	18.9
			30	15.8	40.1	40.1	20.2	30.0	38.4	38.4	22.5
			60	14.9	49.1	49.3	36.3	35.7	49.0	49.2	39.5
			90	12.8	73.8	72.9	85.7	57.7	74.1	73.8	82.5
3	50	150	0	9.2	10.1	10.2	10.2	9.9	3.8	3.8	6.7
			10	9.0	15.1	15.1	18.9	21.5	12.6	12.6	18.5
			30	9.1	41.4	41.4	20.3	28.8	41.2	41.2	20.8
			60	9.1	50.4	50.5	35.9	39.2	50.3	50.3	37.7
			90	8.3	78.8	77.9	97.9	66.7	78.8	78.2	97.3
3	50	500	0	5.0	6.2	6.2	7.5	5.7	3.8	3.7	6.5
			10	5.0	12.9	12.9	18.4	20.9	12.4	12.5	18.4
			30	5.0	42.2	42.2	20.6	28.0	42.2	42.2	20.6
			60	5.0	50.4	50.4	35.5	39.1	50.4	50.4	36.0
			90	5.0	82.3	81.8	114.2	77.3	82.5	81.9	114.1
3	90	50	0	15.9	16.0	16.0	15.9	16.0	1.4	1.4	1.2
			10	15.7	17.7	17.7	15.0	15.5	8.2	8.2	9.9
			30	16.0	23.1	23.1	12.6	14.8	18.4	18.4	15.2
			60	15.5	38.5	38.5	26.3	18.5	38.4	38.4	34.0
			90	13.4	75.0	74.7	68.4	46.0	77.3	76.9	72.8
3	90	150	0	9.0	9.1	9.2	9.1	9.1	1.2	1.3	1.1
			10	9.2	11.3	11.3	10.7	11.0	6.7	6.7	8.7
			30	9.1	16.7	16.7	10.8	11.8	14.3	14.3	11.2
			60	9.1	27.4	27.4	21.9	13.7	26.4	26.4	24.8
			90	8.9	69.1	68.9	74.9	42.8	69.9	69.7	77.7
3	90	500	0	4.9	5.1	5.1	5.1	5.1	1.2	1.2	1.1
			10	5.0	7.7	7.8	8.9	9.2	6.2	6.3	8.5
			30	5.0	13.7	13.7	10.0	10.6	12.8	12.8	10.1
			60	5.0	21.6	21.6	19.7	11.8	21.1	21.1	20.9
			90	4.9	48.4	48.4	64.9	32.7	48.5	48.5	66.1

	Time-to-Event Distributions								
	Kullback-Leibler Divergence								
	External Validation					Internal Validation			
MDR vs. Trial	Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
3.7	0.108	0.303	0.083	0.151	0.106	0.313	0.070	0.171	0.110
21.5	0.107	0.324	0.121	0.254	0.208	0.315	0.103	0.259	0.207
31.2	0.107	0.477	0.370	0.300	0.358	0.455	0.319	0.306	0.341
40.0	0.108	0.464	0.991	0.349	0.854	0.456	0.789	0.354	0.723
57.2	0.110	0.883	1.488	0.614	1.481	0.877	1.274	0.627	1.307
2.7	0.037	0.289	0.029	0.127	0.069	0.280	0.031	0.132	0.072
21.0	0.037	0.298	0.044	0.232	0.171	0.285	0.045	0.224	0.163
29.3	0.039	0.478	0.196	0.284	0.290	0.456	0.188	0.280	0.280
40.9	0.059	0.446	0.947	0.352	0.900	0.441	0.760	0.354	0.778
67.1	0.079	0.787	1.518	0.931	1.935	0.822	1.087	0.934	1.614
2.4	0.014	0.285	0.014	0.120	0.058	0.274	0.017	0.120	0.059
20.8	0.014	0.289	0.022	0.224	0.158	0.279	0.024	0.215	0.151
28.2	0.016	0.480	0.144	0.277	0.261	0.464	0.143	0.273	0.254
39.6	0.032	0.442	0.814	0.338	0.851	0.430	0.701	0.334	0.772
77.4	0.053	0.782	1.675	1.356	2.270	0.794	1.239	1.329	1.960
1.4	0.107	0.058	0.078	0.054	0.057	0.084	0.062	0.075	0.073
10.2	0.106	0.072	0.157	0.055	0.066	0.089	0.132	0.084	0.091
15.0	0.106	0.116	0.427	0.081	0.180	0.133	0.377	0.108	0.193
22.6	0.102	0.304	1.118	0.205	0.739	0.313	0.970	0.215	0.664
49.9	0.103	0.967	1.580	0.618	1.460	0.961	1.300	0.606	1.270
1.2	0.037	0.032	0.025	0.024	0.024	0.043	0.027	0.035	0.034
9.0	0.036	0.039	0.076	0.032	0.035	0.049	0.076	0.045	0.047
11.5	0.036	0.058	0.280	0.040	0.108	0.069	0.270	0.054	0.115
14.8	0.038	0.120	0.947	0.088	0.553	0.135	0.877	0.100	0.517
44.5	0.069	0.726	1.575	0.625	1.691	0.730	1.217	0.619	1.442
1.2	0.014	0.024	0.010	0.015	0.014	0.028	0.013	0.020	0.019
8.8	0.013	0.028	0.045	0.024	0.024	0.033	0.048	0.029	0.030
10.4	0.014	0.039	0.206	0.027	0.075	0.045	0.207	0.034	0.079
12.0	0.015	0.060	0.833	0.043	0.447	0.068	0.798	0.052	0.426
33.3	0.041	0.244	1.549	0.301	1.600	0.254	1.256	0.311	1.385

Number of Events	Distribution Overlap (%)	Sample Size	Censored Proportion (%)	Incidence of Events Relative Absolute Difference (%)							
				External Validation					Internal Validation		
				Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial
4	10	50	0	18.4	22.9	22.9	29.9	29.0	12.2	12.2	30.5
			10	17.1	20.8	20.8	20.1	25.6	11.4	11.4	16.5
			30	14.5	28.3	28.3	19.9	34.2	23.2	23.2	22.0
			60	11.6	43.9	44.0	40.1	40.2	47.6	47.8	42.8
			90	12.2	74.9	73.6	95.1	68.3	72.8	73.2	86.7
4	10	150	0	10.7	16.4	16.4	27.7	22.3	11.9	11.9	27.4
			10	9.9	16.2	16.2	17.3	25.0	12.8	12.8	16.8
			30	8.8	25.0	25.0	18.5	35.1	23.7	23.7	19.0
			60	6.6	48.0	48.1	38.8	44.3	49.6	49.6	39.9
			90	6.4	70.4	68.7	101.2	70.7	69.1	68.5	97.9
4	10	500	0	5.8	13.1	13.1	26.1	19.5	11.7	11.7	25.9
			10	5.7	16.5	16.5	18.3	28.8	15.8	15.8	18.3
			30	4.8	23.1	23.1	19.2	35.1	22.9	22.9	19.2
			60	4.0	56.6	56.7	39.2	52.4	56.8	56.8	39.7
			90	3.4	65.2	63.8	102.1	65.8	65.1	64.1	101.2
4	50	50	0	18.2	21.2	21.2	21.7	23.6	9.5	9.5	16.4
			10	17.1	20.9	20.9	19.1	24.7	11.6	11.6	15.6
			30	15.4	34.8	34.8	20.2	29.2	32.5	32.5	23.1
			60	14.2	48.1	48.3	39.1	35.1	48.6	48.6	42.8
			90	12.1	75.5	74.3	90.5	63.0	75.7	75.1	86.8
4	50	150	0	10.7	14.4	14.4	16.9	16.7	9.0	9.0	14.9
			10	10.4	17.5	17.5	17.0	25.6	13.7	13.8	16.2
			30	9.1	36.9	36.9	22.2	32.3	36.8	36.8	22.4
			60	8.8	49.7	49.8	39.5	39.0	49.9	50.0	41.6
			90	7.8	78.3	77.2	100.7	68.7	78.4	77.4	99.9
4	50	500	0	5.8	10.7	10.7	15.0	12.3	8.8	8.8	14.5
			10	5.8	15.7	15.7	16.7	26.3	14.8	14.8	16.6
			30	5.2	40.8	40.8	26.6	39.8	40.9	40.9	26.5
			60	4.9	51.9	51.9	42.3	41.5	52.0	52.0	43.0
			90	4.8	80.8	79.8	114.1	77.5	81.1	80.2	114.0

	Time-to-Event Distributions								
	Kullback-Leibler Divergence								
	External Validation					Internal Validation			
MDR vs. Trial	Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
21.1	0.105	0.646	0.146	0.219	0.108	0.636	0.140	0.233	0.097
22.7	0.089	0.575	0.173	0.581	0.469	0.568	0.158	0.577	0.447
35.0	0.088	0.556	0.222	0.502	0.418	0.546	0.186	0.501	0.382
48.4	0.080	0.350	0.423	0.223	0.452	0.341	0.281	0.220	0.320
65.5	0.107	0.662	1.130	0.530	1.134	0.726	1.095	0.582	1.101
19.1	0.036	0.641	0.098	0.204	0.063	0.616	0.094	0.196	0.057
24.8	0.033	0.579	0.209	0.582	0.486	0.557	0.187	0.558	0.453
35.2	0.034	0.556	0.174	0.499	0.392	0.534	0.146	0.478	0.353
46.9	0.049	0.281	0.592	0.124	0.554	0.284	0.342	0.120	0.346
69.7	0.050	0.469	0.992	0.750	1.446	0.571	0.760	0.812	1.249
18.4	0.013	0.641	0.083	0.202	0.048	0.619	0.080	0.192	0.045
28.8	0.015	0.586	0.304	0.581	0.515	0.564	0.269	0.559	0.481
35.0	0.013	0.545	0.147	0.504	0.373	0.527	0.131	0.484	0.347
53.0	0.037	0.278	0.707	0.136	0.604	0.287	0.439	0.149	0.422
65.8	0.031	0.404	0.999	0.953	1.585	0.477	0.735	0.970	1.351
13.8	0.141	0.429	0.188	0.199	0.151	0.428	0.179	0.214	0.150
21.4	0.122	0.369	0.185	0.305	0.271	0.371	0.165	0.321	0.266
30.6	0.111	0.379	0.303	0.252	0.278	0.362	0.260	0.262	0.267
40.2	0.102	0.372	0.813	0.275	0.709	0.367	0.644	0.285	0.600
62.0	0.103	0.749	1.292	0.535	1.309	0.749	1.132	0.558	1.179
12.2	0.046	0.418	0.126	0.170	0.104	0.398	0.121	0.171	0.104
25.0	0.046	0.359	0.199	0.289	0.284	0.341	0.156	0.278	0.243
32.5	0.040	0.361	0.210	0.227	0.244	0.343	0.184	0.225	0.229
41.0	0.055	0.362	0.804	0.290	0.781	0.357	0.629	0.296	0.668
69.1	0.067	0.686	1.284	0.866	1.730	0.719	0.940	0.875	1.473
10.6	0.017	0.418	0.094	0.163	0.086	0.399	0.091	0.160	0.085
26.2	0.020	0.351	0.197	0.281	0.283	0.334	0.148	0.267	0.235
39.8	0.019	0.357	0.303	0.256	0.314	0.345	0.256	0.257	0.296
42.1	0.031	0.364	0.720	0.314	0.774	0.354	0.598	0.315	0.699
77.7	0.046	0.691	1.424	1.252	2.020	0.706	1.067	1.232	1.770

Number of Events	Distribution Overlap (%)	Sample Size	Censored Proportion (%)	Incidence of Events Relative Absolute Difference (%)							
				External Validation					Internal Validation		
				Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial
4	90	50	0	18.3	18.6	18.6	18.4	19.2	2.4	2.5	1.8
			10	18.0	20.1	20.1	16.1	18.5	10.0	10.0	11.5
			30	17.8	25.8	25.8	23.4	25.3	20.9	20.9	25.4
			60	16.6	41.1	41.0	43.7	29.9	40.9	40.8	50.1
			90	12.8	74.7	74.4	71.1	50.5	77.1	76.7	75.4
4	90	150	0	10.7	11.0	11.0	10.8	12.1	2.0	2.0	1.4
			10	10.7	13.4	13.4	10.9	13.2	8.5	8.5	8.9
			30	10.7	19.9	19.9	25.1	26.3	17.6	17.6	26.7
			60	10.4	32.4	32.4	64.1	37.6	32.0	31.9	66.8
			90	9.1	75.2	74.7	78.0	48.2	76.2	75.6	80.5
4	90	500	0	5.8	6.1	6.1	6.0	7.9	1.9	1.9	1.4
			10	5.8	9.7	9.7	8.4	10.6	7.9	7.9	7.7
			30	5.9	16.9	16.9	24.1	24.7	16.0	16.0	24.6
			60	5.8	25.8	25.8	72.1	37.4	25.4	25.4	73.1
			90	5.6	63.8	63.4	86.1	51.0	64.1	63.7	86.9

	Time-to-Event Distributions								
	Kullback-Leibler Divergence								
	External Validation					Internal Validation			
MDR vs. Trial	Trial vs. Population	ESD vs. Population	ESPD vs. Population	UDR vs. Population	MDR vs. Population	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
5.4	0.146	0.086	0.166	0.075	0.100	0.113	0.159	0.096	0.112
13.8	0.137	0.097	0.219	0.065	0.118	0.115	0.204	0.108	0.151
24.6	0.126	0.134	0.506	0.102	0.340	0.150	0.457	0.145	0.344
32.9	0.110	0.296	1.030	0.241	0.833	0.303	0.885	0.259	0.747
54.5	0.098	0.863	1.404	0.554	1.307	0.852	1.163	0.540	1.135
4.8	0.046	0.056	0.107	0.037	0.058	0.065	0.106	0.047	0.065
11.3	0.047	0.062	0.145	0.040	0.086	0.070	0.143	0.059	0.100
26.6	0.047	0.074	0.515	0.082	0.414	0.085	0.451	0.113	0.396
38.4	0.049	0.154	0.960	0.350	1.058	0.170	0.841	0.369	0.986
49.7	0.076	0.699	1.278	0.598	1.514	0.704	0.971	0.592	1.257
4.9	0.016	0.047	0.088	0.025	0.045	0.049	0.087	0.030	0.049
10.0	0.018	0.050	0.115	0.032	0.071	0.054	0.117	0.041	0.077
24.8	0.022	0.048	0.503	0.057	0.406	0.053	0.433	0.078	0.382
37.7	0.024	0.072	0.940	0.370	1.164	0.080	0.828	0.390	1.105
51.4	0.055	0.401	1.092	0.697	1.629	0.405	0.864	0.688	1.380

APPENDIX 10: detailed ase study results on performance

Please note that the table has been split into two parts. The first part concerns the performance ragerding the incidence of events and the second part concerns the time-to-event distributions.

[1]Treatment Strategy: 0) observation strategy, 1) maintenance treatment. SD, stable disease; CR/PR, complete or partial response; Meta., metachronous; Synch., synchronous. [2]A hyphen (-) indicates that the corresponding probabilities were either zero or one, i.e. one of the competing events was not observed, making a comparison of the relative incidence irrelevant. [3]A hyphen (-) indicates that insufficient observations were available in de original trial data to calculate the Kullback-Leibler divergence.

						Incidence of Events[2] (Relative Difference (%))							
						Post-Induction: Progression				Reintroduction: Progression			
Subgroup Number	Treatment Response	Stage of Disease	Censored Proportion	Treatment Strategy[1]	Subgroup Size	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
0	-	-	0.0	0	279	1.9	1.9	1.4	1.7	2.7	2.7	2.7	2.7
				1	279	2.0	2.0	1.7	1.7	3.1	3.1	3.1	3.1
0	-	-	0.1	0	279	2.1	2.1	1.9	1.8	2.7	2.7	2.6	2.6
				1	279	2.3	2.3	1.8	1.9	3.6	3.6	3.0	3.0
0	-	-	0.3	0	279	2.3	2.3	1.8	1.8	3.0	3.0	3.5	3.5
				1	279	4.7	4.7	2.8	1.7	3.5	3.5	6.0	6.0
0	-	-	0.6	0	279	2.4	2.7	1.1	1.8	3.6	3.6	16.5	16.7
				1	279	4.2	4.2	16.6	3.0	5.7	5.7	11.9	12.0
1	SD	-	0.0	0	95	-	-	0.0	0.0	4.8	4.8	4.8	4.8
				1	96	2.1	2.1	2.1	2.1	5.2	5.2	5.2	5.2
1	SD	-	0.1	0	95	-	-	0.0	0.0	4.9	4.8	4.7	4.7
				1	96	2.1	2.1	2.2	2.2	5.5	5.5	5.1	5.1
1	SD	-	0.3	0	95	-	-	0.0	0.0	6.0	6.0	5.1	5.1
				1	96	2.3	2.3	7.1	4.4	5.7	5.7	5.7	5.7
1	SD	-	0.6	0	95	-	-	0.5	0.5	5.7	5.7	19.7	19.6
				1	96	6.0	5.9	15.3	9.9	8.9	8.9	11.4	11.5
2	CR/PR	-	0.0	0	184	2.6	2.6	2.0	2.4	3.1	3.1	3.1	3.1
				1	183	2.8	2.8	2.4	2.4	3.7	3.7	3.7	3.7
2	CR/PR	-	0.1	0	184	3.0	3.0	2.6	2.5	3.1	3.1	3.1	3.1
				1	183	3.4	3.4	2.7	3.0	4.2	4.2	3.6	3.6

Subgroup Number	Treatment Response	Stage of Disease	Censored Proportion	Treatment Strategy[1]	Subgroup Size	Incidence of Events[2] (Relative Difference (%))							
						Post-Induction: Progression				Reintroduction: Progression			
						ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
2	CR/PR	-	0.3	0	184	3.5	3.5	2.6	2.7	3.5	3.5	4.0	4.1
				1	183	7.2	7.2	2.2	3.2	4.2	4.2	7.1	7.1
2	CR/PR	-	0.6	0	184	3.2	3.2	1.7	2.6	3.7	3.7	13.1	13.2
				1	183	8.2	8.2	16.9	3.2	5.4	5.4	12.7	12.7
3	-	Synch	0.0	0	191	1.2	1.2	1.3	1.2	2.9	2.9	2.9	2.9
				1	219	2.1	2.1	1.9	1.9	3.7	3.7	3.7	3.6
3	-	Synch	0.1	0	191	1.3	1.3	1.2	1.2	2.9	2.9	2.9	2.9
				1	219	2.9	2.9	2.0	2.1	4.4	4.4	3.4	3.4
3	-	Synch	0.3	0	191	1.5	1.5	1.2	1.2	3.1	3.1	3.3	3.3
				1	219	4.7	4.7	3.5	1.9	5.9	5.9	4.2	4.2
3	-	Synch	0.6	0	191	2.2	2.2	1.3	1.3	4.1	4.1	17.0	17.2
				1	219	4.3	4.3	18.9	4.0	7.9	7.9	10.8	10.8
4	-	Meta	0.0	0	88	4.1	4.1	2.9	4.1	5.6	5.6	5.6	5.6
				1	59	3.5	3.5	3.4	3.3	4.7	4.7	4.7	4.7
4	-	Meta	0.1	0	88	4.2	4.2	3.6	4.2	5.6	5.6	5.6	5.6
				1	59	3.6	3.6	3.3	3.3	4.8	4.8	4.8	4.8
4	-	Meta	0.3	0	88	4.8	4.8	3.8	3.8	6.6	6.6	5.3	5.2
				1	59	4.7	4.7	3.2	3.0	6.2	6.2	13.6	12.9
4	-	Meta	0.6	0	88	4.8	4.8	2.3	2.3	7.7	7.7	11.7	11.7
				1	59	5.3	5.3	15.8	5.8	5.8	5.9	9.9	8.2
5	SD	Synch	0.0	0	67	-	-	0.0	0.0	5.6	5.6	5.5	5.5
				1	74	2.4	2.4	2.4	2.4	6.6	6.6	6.6	6.6
5	SD	Synch	0.1	0	67	-	-	0.0	0.0	5.6	5.6	5.4	5.4
				1	74	2.5	2.6	2.9	2.8	7.8	7.8	7.2	7.2
5	SD	Synch	0.3	0	67	-	-	0.0	0.0	7.0	7.0	5.3	5.3
				1	74	2.6	2.6	5.4	4.1	7.5	7.5	6.6	6.7
5	SD	Synch	0.6	0	67	-	-	0.5	0.5	7.8	7.8	14.3	14.4
				1	74	4.9	4.9	19.1	10.3	9.1	9.1	14.0	14.0
6	SD	Meta	0.0	0	28	-	-	0.0	0.0	9.4	9.4	9.9	9.4
				1	22	3.7	3.7	3.8	3.8	7.5	7.5	7.8	7.7
6	SD	Meta	0.1	0	28	-	-	0.0	0.0	9.6	9.6	10.2	9.6
				1	22	3.7	3.7	3.7	3.7	9.8	9.8	17.1	13.8

Subgroup Number	Treatment Response	Stage of Disease	Censored Proportion	Treatment Strategy[1]	Subgroup Size	Incidence of Events[2] (Relative Difference (%))							
						Post-Induction: Progression				Reintroduction: Progression			
						ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
6	SD	Meta	0.3	0	28	-	-	0.0	0.0	9.6	9.6	10.1	9.6
				1	22	3.9	3.9	5.5	5.0	9.4	9.4	17.4	16.2
6	SD	Meta	0.6	0	28	-	-	1.5	1.5	11.2	11.2	18.2	18.2
				1	22	9.6	9.6	15.4	12.3	10.6	11.5	7.0	7.6
7	CR/PR	Synch	0.0	0	124	1.8	1.8	2.0	1.7	3.3	3.3	3.3	3.3
				1	145	3.0	3.0	2.7	2.6	4.4	4.4	4.4	4.4
7	CR/PR	Synch	0.1	0	124	2.0	2.0	1.8	1.7	3.4	3.4	3.3	3.3
				1	145	3.5	3.5	2.8	3.1	5.0	5.0	4.1	4.1
7	CR/PR	Synch	0.3	0	124	2.2	2.2	1.8	1.8	3.4	3.4	4.6	4.7
				1	145	5.8	5.8	2.8	2.7	6.0	6.0	6.7	6.8
7	CR/PR	Synch	0.6	0	124	2.6	2.6	1.8	2.1	3.6	3.6	13.3	13.2
				1	145	8.0	8.0	18.8	3.5	8.6	8.6	8.5	8.6
8	CR/PR	Meta	0.0	0	61	5.5	5.5	3.3	5.5	6.9	6.9	6.8	6.8
				1	37	5.3	5.3	5.1	5.0	6.1	6.1	5.9	5.9
8	CR/PR	Meta	0.1	0	61	5.7	5.7	4.7	5.8	6.8	6.8	6.8	6.8
				1	37	5.6	5.6	4.9	5.1	6.2	6.2	6.0	6.0
8	CR/PR	Meta	0.3	0	61	7.1	7.1	5.2	5.2	8.4	8.4	6.2	6.2
				1	37	5.7	5.8	4.7	4.8	7.3	7.3	9.3	9.0
8	CR/PR	Meta	0.6	0	61	7.1	7.1	4.1	4.1	9.6	9.6	6.5	6.4
				1	37	9.7	9.7	11.7	6.5	8.6	8.6	17.2	15.7

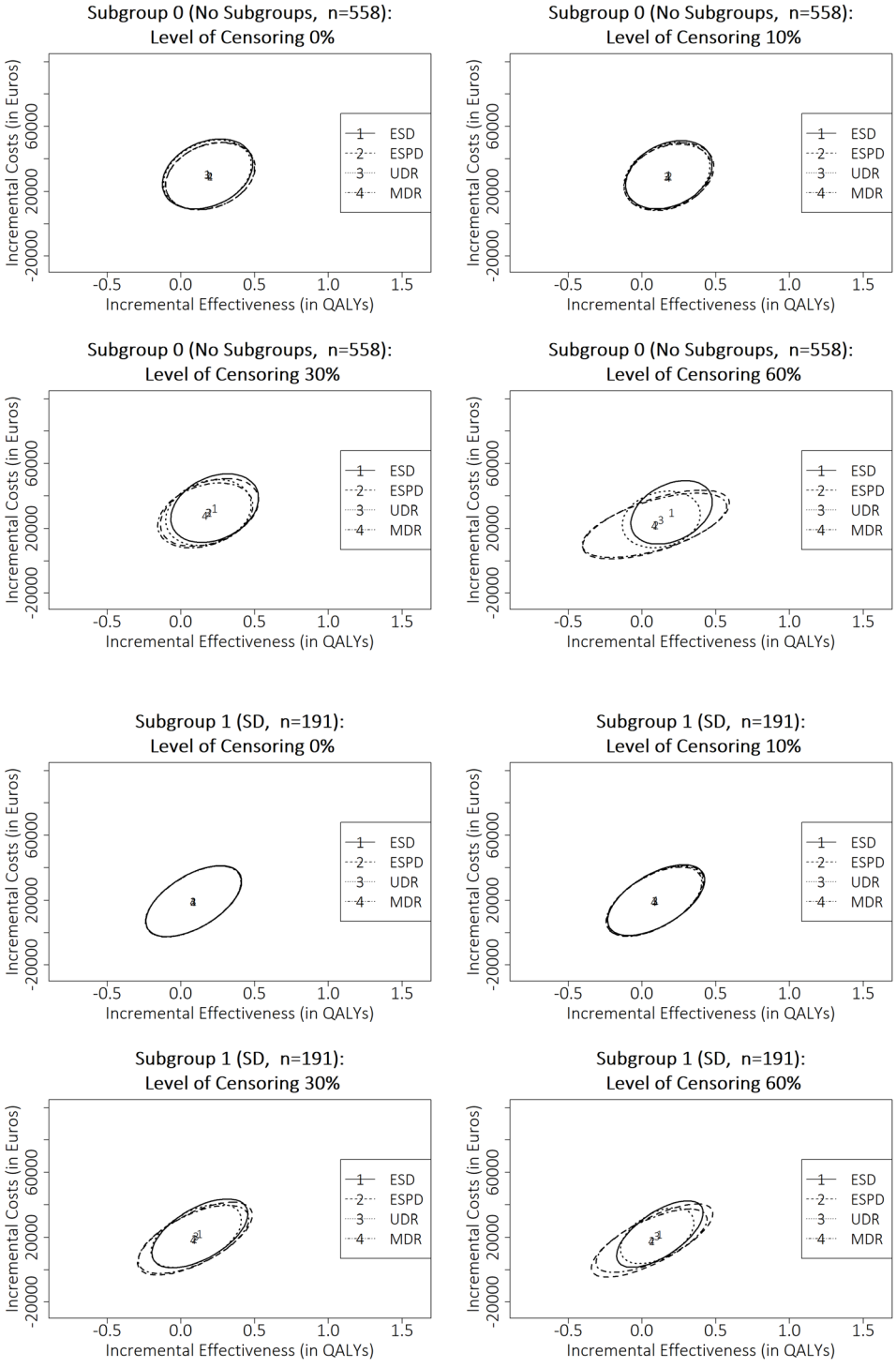
Subgroup Number	Censored Proportion	Time-to-Event Distributions[3] (Kullback-Leibler Divergence)															
		Post-Induction: Progression				Post-Induction: Death				Reintroduction: Progression				Reintroduction: Death			
		ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
0	0.0	0.071	0.013	0.063	0.013	0.735	0.145	0.621	0.183	0.025	0.013	0.028	0.021	0.038	0.032	0.053	0.056
		0.009	0.010	0.007	0.010	0.102	0.036	0.042	0.037	0.063	0.035	0.059	0.046	0.049	0.020	0.035	0.046
0	0.1	0.060	0.013	0.057	0.013	0.611	0.116	0.499	0.229	0.023	0.013	0.026	0.021	0.040	0.033	0.052	0.055
		0.008	0.010	0.008	0.009	0.138	0.047	0.047	0.048	0.047	0.016	0.046	0.026	0.082	0.020	0.051	0.085
0	0.3	0.044	0.015	0.041	0.015	0.583	0.165	0.466	0.191	0.014	0.014	0.017	0.010	0.048	0.033	0.058	0.068
		0.009	0.027	0.010	0.034	0.133	0.060	0.044	0.038	0.032	0.018	0.034	0.021	0.118	0.020	0.079	0.129
0	0.6	0.024	0.017	0.020	0.018	0.444	0.077	0.386	0.154	0.013	0.119	0.016	0.079	0.064	0.096	0.087	0.093
		0.012	0.146	0.080	0.188	0.239	0.113	0.230	0.111	0.016	0.086	0.021	0.044	0.249	0.027	0.190	0.271
1	0.0	0.018	0.018	0.018	0.018	-	-	-	-	0.018	0.027	0.020	0.025	0.119	0.084	0.153	0.166
		0.020	0.021	0.021	0.020	0.249	0.600	0.255	0.252	0.017	0.025	0.015	0.016	0.098	0.068	0.105	0.107
1	0.1	0.018	0.018	0.018	0.018	-	-	-	-	0.018	0.026	0.019	0.023	0.115	0.086	0.144	0.159
		0.020	0.023	0.021	0.022	0.236	0.350	0.248	0.238	0.017	0.025	0.016	0.019	0.142	0.118	0.139	0.139
1	0.3	0.018	0.018	0.018	0.018	-	-	-	-	0.018	0.058	0.020	0.038	0.098	0.102	0.137	0.173
		0.021	0.065	0.030	0.072	0.158	0.404	0.312	0.191	0.016	0.051	0.021	0.033	0.137	0.086	0.118	0.112
1	0.6	0.018	0.018	0.018	0.018	-	-	-	-	0.059	0.312	0.056	0.176	0.146	0.162	0.215	0.186
		0.046	0.130	0.101	0.169	0.472	0.667	0.620	0.509	0.025	0.084	0.022	0.041	0.164	0.123	0.122	0.139
2	0.0	0.090	0.016	0.076	0.016	0.638	0.143	0.572	0.170	0.046	0.028	0.045	0.035	0.074	0.090	0.092	0.084
		0.013	0.012	0.008	0.013	0.128	0.081	0.086	0.081	0.085	0.057	0.081	0.068	0.077	0.050	0.049	0.059

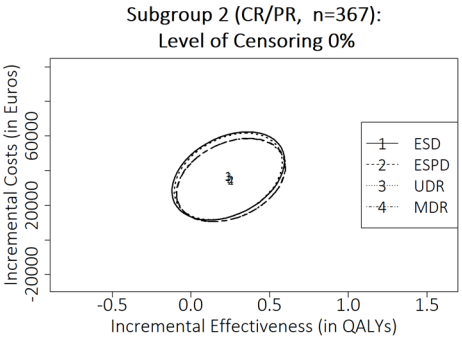
Subgroup Number	Censored Proportion	Time-to-Event Distributions[3] (Kullback-Leibler Divergence)															
		Post-Induction: Progression				Post-Induction: Death				Reintroduction: Progression				Reintroduction: Death			
		Trial vs. ESD	Trial vs. ESPD	Trial vs. UDR	Trial vs. MDR	Trial vs. ESD	Trial vs. ESPD	Trial vs. UDR	Trial vs. MDR	Trial vs. ESD	Trial vs. ESPD	Trial vs. UDR	Trial vs. MDR	Trial vs. ESD	Trial vs. ESPD	Trial vs. UDR	Trial vs. MDR
2	0.1	0.080	0.016	0.076	0.016	0.504	0.127	0.390	0.209	0.044	0.028	0.044	0.035	0.075	0.090	0.088	0.081
		0.011	0.012	0.008	0.010	0.156	0.091	0.088	0.101	0.076	0.042	0.073	0.054	0.106	0.048	0.057	0.080
2	0.3	0.062	0.019	0.058	0.019	0.495	0.212	0.379	0.162	0.029	0.015	0.028	0.012	0.068	0.093	0.081	0.065
		0.012	0.023	0.008	0.030	0.173	0.093	0.067	0.073	0.055	0.026	0.056	0.036	0.152	0.045	0.093	0.147
2	0.6	0.041	0.020	0.033	0.021	0.870	0.562	0.725	0.568	0.016	0.068	0.018	0.050	0.082	0.126	0.094	0.078
		0.012	0.163	0.076	0.200	0.172	0.137	0.128	0.117	0.025	0.116	0.036	0.070	0.313	0.056	0.203	0.353
3	0.0	0.066	0.014	0.059	0.014	0.522	0.499	0.408	0.231	0.022	0.016	0.023	0.021	0.060	0.057	0.082	0.082
		0.008	0.007	0.005	0.007	0.122	0.039	0.048	0.042	0.055	0.025	0.052	0.038	0.045	0.019	0.036	0.048
3	0.1	0.050	0.014	0.047	0.014	0.562	0.540	0.441	0.292	0.019	0.016	0.021	0.020	0.062	0.063	0.079	0.079
		0.007	0.008	0.005	0.007	0.155	0.035	0.048	0.050	0.030	0.012	0.031	0.015	0.094	0.029	0.070	0.107
3	0.3	0.039	0.012	0.036	0.011	0.527	0.248	0.378	0.169	0.011	0.012	0.012	0.008	0.061	0.057	0.089	0.096
		0.007	0.029	0.007	0.039	0.155	0.056	0.064	0.045	0.019	0.018	0.024	0.015	0.142	0.036	0.097	0.143
3	0.6	0.020	0.015	0.017	0.016	0.261	0.078	0.168	0.076	0.013	0.151	0.016	0.112	0.060	0.153	0.107	0.094
		0.009	0.183	0.099	0.238	0.258	0.107	0.245	0.132	0.017	0.147	0.025	0.077	0.228	0.043	0.164	0.247
4	0.0	0.073	0.013	0.062	0.013	0.468	0.374	0.987	0.352	0.032	0.025	0.040	0.031	0.068	0.064	0.078	0.076
		0.044	0.053	0.048	0.053	0.145	0.437	0.141	0.142	0.078	0.073	0.073	0.077	0.176	0.438	0.148	0.147
4	0.1	0.067	0.017	0.061	0.016	0.640	0.451	0.683	0.587	0.032	0.025	0.040	0.032	0.068	0.063	0.077	0.075
		0.044	0.053	0.047	0.050	0.165	0.428	0.107	0.123	0.078	0.073	0.073	0.078	0.175	0.457	0.144	0.144

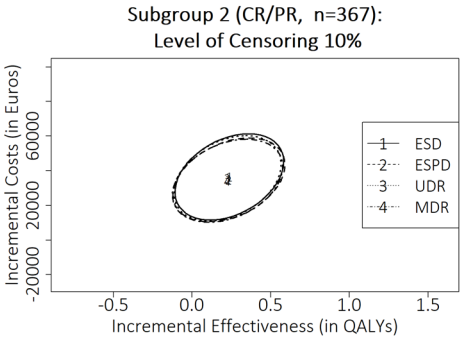
Subgroup Number	Censored Proportion	Time-to-Event Distributions[3] (Kullback-Leibler Divergence)															
		Post-Induction: Progression				Post-Induction: Death				Reintroduction: Progression				Reintroduction: Death			
		ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial	ESD vs. Trial	ESPD vs. Trial	UDR vs. Trial	MDR vs. Trial
4	0.3	0.044	0.044	0.044	0.044	-	-	-	-	0.026	0.030	0.034	0.025	0.074	0.051	0.057	0.059
		0.045	0.060	0.048	0.061	0.192	0.255	0.119	0.131	0.079	0.069	0.070	0.064	0.177	0.262	0.125	0.193
4	0.6	0.026	0.026	0.023	0.023	-	-	-	-	0.027	0.136	0.028	0.080	0.083	0.077	0.079	0.111
		0.052	0.186	0.133	0.204	0.141	0.176	0.153	0.086	0.056	0.149	0.056	0.121	0.443	0.282	0.229	0.279
5	0.0	0.024	0.024	0.024	0.024	-	-	-	-	0.016	0.028	0.019	0.023	0.309	0.239	0.359	0.396
		0.014	0.015	0.014	0.015	0.276	0.751	0.259	0.258	0.019	0.031	0.019	0.020	0.164	0.149	0.181	0.187
5	0.1	0.024	0.024	0.024	0.024	-	-	-	-	0.017	0.027	0.019	0.022	0.281	0.232	0.329	0.367
		0.015	0.023	0.015	0.023	0.226	0.315	0.236	0.206	0.025	0.032	0.021	0.025	0.182	0.192	0.180	0.181
5	0.3	0.024	0.024	0.024	0.024	-	-	-	-	0.020	0.027	0.018	0.025	0.279	0.366	0.338	0.338
		0.015	0.050	0.020	0.058	0.189	0.334	0.301	0.193	0.028	0.056	0.025	0.035	0.179	0.193	0.172	0.168
5	0.6	0.023	0.023	0.023	0.023	-	-	-	-	0.106	0.249	0.072	0.164	0.237	0.335	0.378	0.333
		0.031	0.183	0.134	0.262	0.377	0.600	0.531	0.352	0.044	0.190	0.045	0.090	0.179	0.195	0.149	0.135
6	0.0	0.053	0.053	0.053	0.053	-	-	-	-	0.102	0.104	0.085	0.110	0.114	0.292	0.232	0.208
		0.098	0.094	0.098	0.103	-	-	-	-	0.051	0.100	0.099	0.112	0.346	2.480	0.207	0.245
6	0.1	0.053	0.053	0.053	0.053	-	-	-	-	0.102	0.103	0.084	0.109	0.113	0.278	0.227	0.203
		0.099	0.096	0.099	0.105	-	-	-	-	0.058	0.160	0.124	0.179	0.381	1.958	0.325	0.363
6	0.3	0.053	0.053	0.053	0.053	-	-	-	-	0.102	0.103	0.084	0.110	0.111	0.290	0.223	0.200
		0.089	0.103	0.101	0.114	-	-	-	-	0.066	0.175	0.095	0.151	0.373	1.885	0.644	0.684

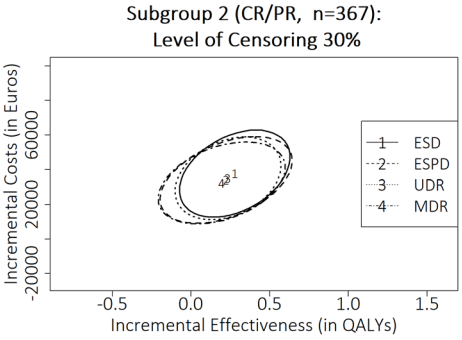
Subgroup Number	Censored Proportion	Time-to-Event Distributions[3] (Kullback-Leibler Divergence)															
		Post-Induction: Progression				Post-Induction: Death				Reintroduction: Progression				Reintroduction: Death			
		Trial ESD vs.	Trial ESPD vs.	Trial UDR vs.	Trial MDR vs.	Trial ESD vs.	Trial ESPD vs.	Trial UDR vs.	Trial MDR vs.	Trial ESD vs.	Trial ESPD vs.	Trial UDR vs.	Trial MDR vs.	Trial ESD vs.	Trial ESPD vs.	Trial UDR vs.	Trial MDR vs.
6	0.6	0.051	0.051	0.049	0.049	-	-	-	-	0.242	0.337	0.172	0.238	0.191	0.377	0.222	0.228
		0.148	0.202	0.187	0.221	-	-	-	-	0.123	0.138	0.137	0.194	1.946	6.668	0.215	1.193
7	0.0	0.089	0.020	0.077	0.019	0.439	0.509	0.349	0.216	0.044	0.042	0.045	0.042	0.101	0.131	0.117	0.113
		0.011	0.013	0.007	0.014	0.131	0.065	0.077	0.067	0.075	0.042	0.070	0.054	0.078	0.051	0.049	0.062
7	0.1	0.074	0.020	0.068	0.020	0.473	0.520	0.351	0.258	0.039	0.044	0.041	0.039	0.099	0.128	0.106	0.105
		0.009	0.013	0.007	0.012	0.178	0.084	0.078	0.084	0.055	0.015	0.052	0.019	0.121	0.045	0.063	0.123
7	0.3	0.059	0.021	0.054	0.021	0.509	0.511	0.362	0.235	0.025	0.026	0.025	0.018	0.080	0.159	0.102	0.078
		0.010	0.033	0.011	0.041	0.207	0.085	0.087	0.083	0.039	0.029	0.041	0.020	0.175	0.064	0.098	0.201
7	0.6	0.038	0.017	0.032	0.017	0.638	0.422	0.410	0.331	0.022	0.078	0.024	0.063	0.082	0.238	0.128	0.094
		0.012	0.201	0.089	0.245	0.168	0.117	0.124	0.105	0.019	0.096	0.027	0.062	0.364	0.223	0.267	0.366
8	0.0	0.088	0.023	0.070	0.023	0.400	0.350	0.957	0.344	0.043	0.026	0.042	0.026	0.122	0.078	0.115	0.119
		0.049	0.051	0.047	0.052	0.346	0.766	0.334	0.357	0.113	0.114	0.114	0.117	0.208	1.049	0.290	0.298
8	0.1	0.081	0.026	0.071	0.026	0.404	0.306	0.579	0.451	0.043	0.026	0.043	0.026	0.122	0.078	0.114	0.118
		0.047	0.051	0.046	0.049	0.354	0.880	0.318	0.351	0.113	0.114	0.114	0.117	0.209	1.060	0.295	0.302
8	0.3	0.054	0.054	0.053	0.053	-	-	-	-	0.042	0.041	0.042	0.026	0.104	0.046	0.070	0.093
		0.048	0.065	0.047	0.060	0.461	0.504	0.327	0.403	0.109	0.125	0.106	0.109	0.218	0.531	0.182	0.221
8	0.6	0.036	0.036	0.032	0.032	-	-	-	-	0.025	0.065	0.024	0.037	0.180	0.065	0.127	0.157
		0.050	0.113	0.093	0.146	0.346	0.314	0.284	0.324	0.107	0.306	0.103	0.215	0.299	0.346	0.200	0.361

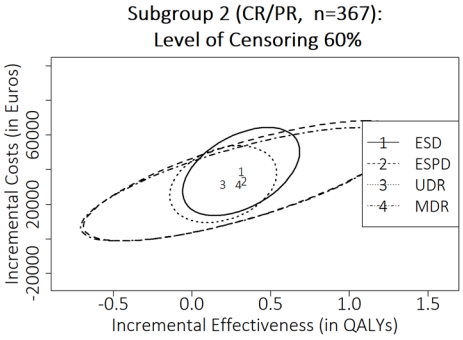
APPENDIX 11: incremental cost-effectiveness planes for all case study subgroup analyses

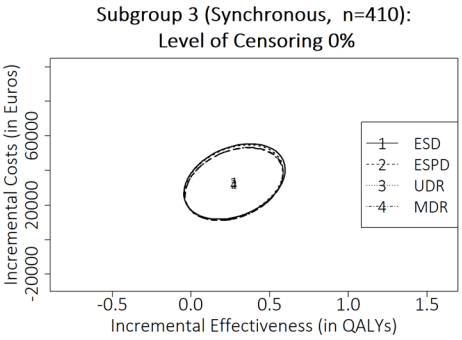


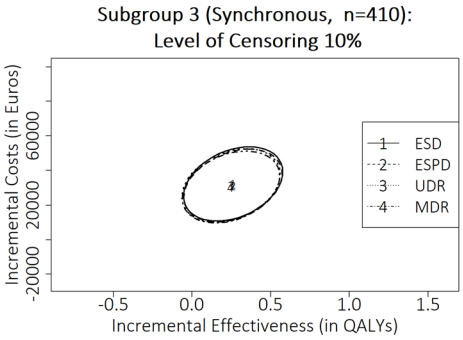

Subgroup 2 (CR/PR, n=367):
Level of Censoring 0%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

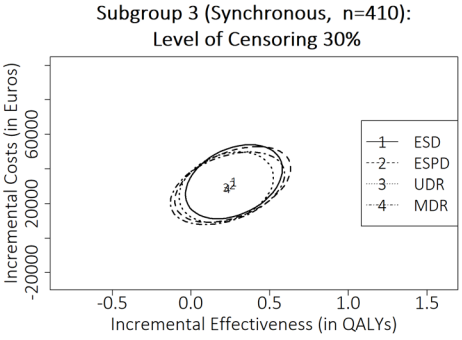

Subgroup 2 (CR/PR, n=367):
Level of Censoring 10%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

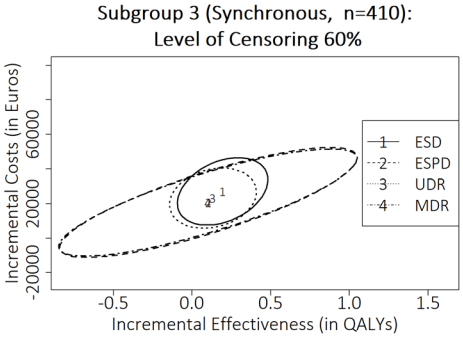

Subgroup 2 (CR/PR, n=367):
Level of Censoring 30%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

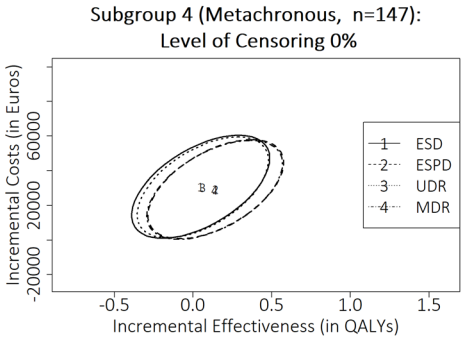

Subgroup 2 (CR/PR, n=367):
Level of Censoring 60%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

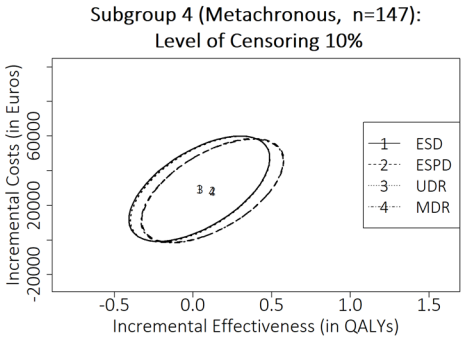

Subgroup 3 (Synchronous, n=410):
Level of Censoring 0%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

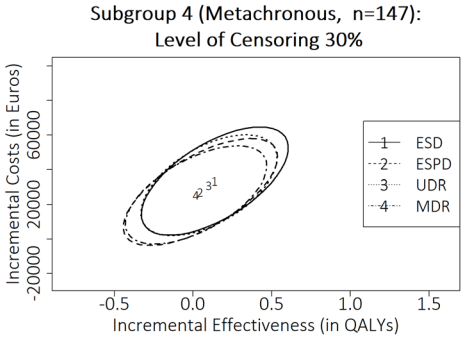

Subgroup 3 (Synchronous, n=410):
Level of Censoring 10%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

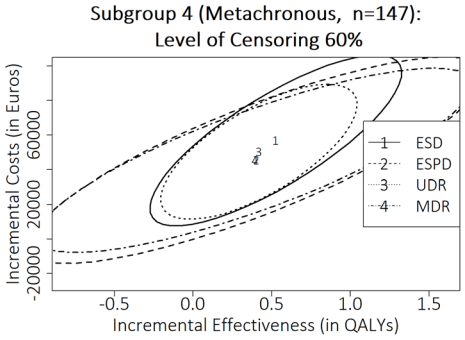

Subgroup 3 (Synchronous, n=410):
Level of Censoring 30%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

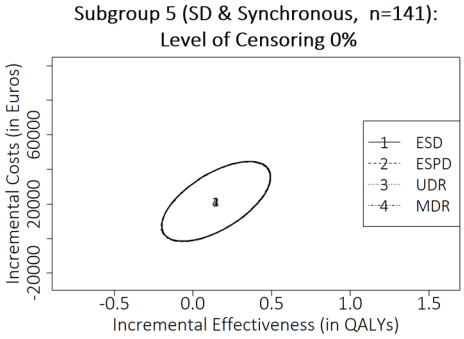

Subgroup 3 (Synchronous, n=410):
Level of Censoring 60%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

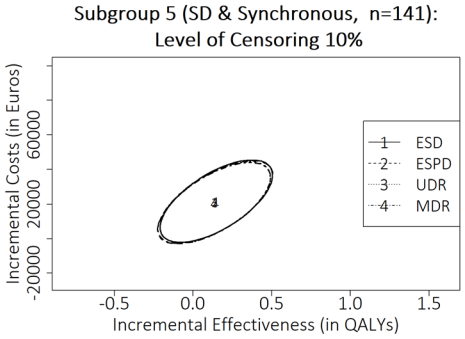

Subgroup 4 (Metachronous, n=147):
Level of Censoring 0%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

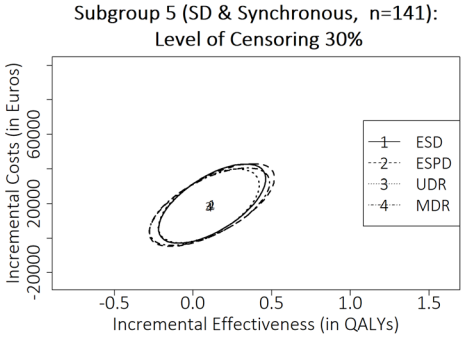

Subgroup 4 (Metachronous, n=147):
Level of Censoring 10%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

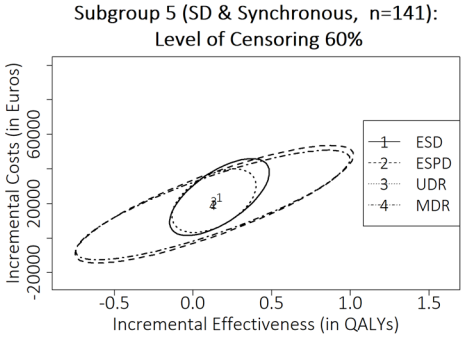

Subgroup 4 (Metachronous, n=147):
Level of Censoring 30%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

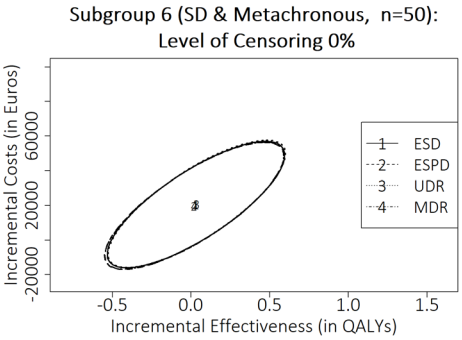

Subgroup 4 (Metachronous, n=147):
Level of Censoring 60%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

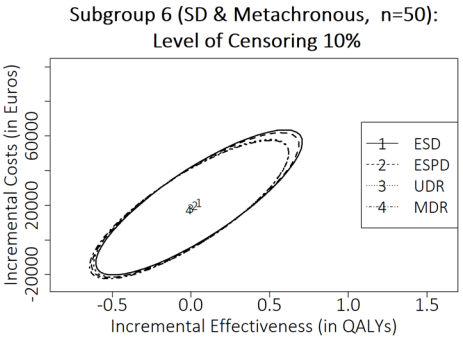

Subgroup 5 (SD & Synchronous, n=141):
Level of Censoring 0%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

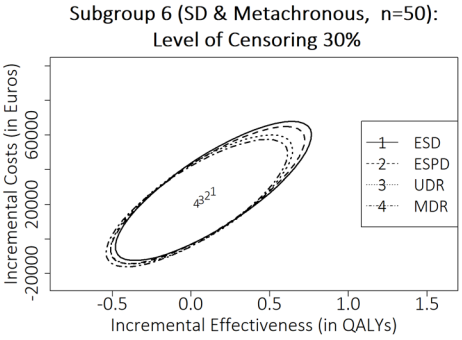

Subgroup 5 (SD & Synchronous, n=141):
Level of Censoring 10%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

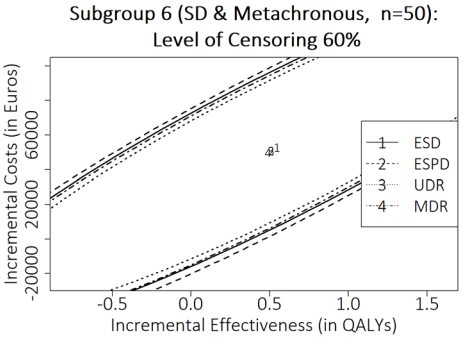

Subgroup 5 (SD & Synchronous, n=141):
Level of Censoring 30%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

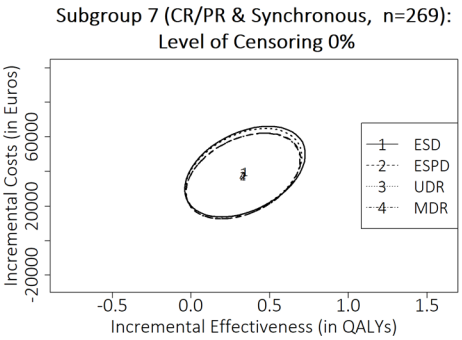

Subgroup 5 (SD & Synchronous, n=141):
Level of Censoring 60%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

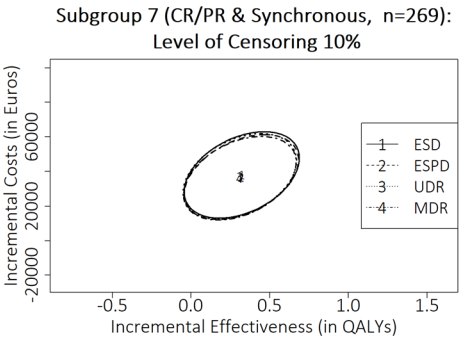

Subgroup 6 (SD & Metachronous, n=50):
Level of Censoring 0%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

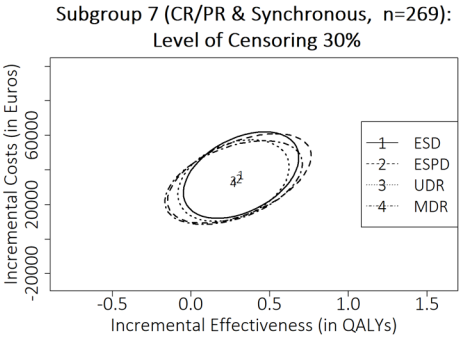

Subgroup 6 (SD & Metachronous, n=50):
Level of Censoring 10%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

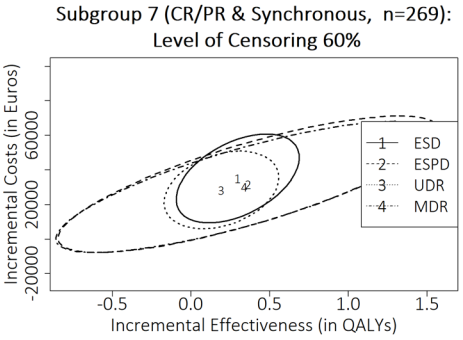

Subgroup 6 (SD & Metachronous, n=50):
Level of Censoring 30%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

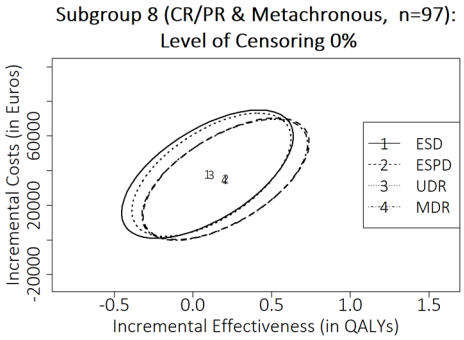

Subgroup 6 (SD & Metachronous, n=50):
Level of Censoring 60%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

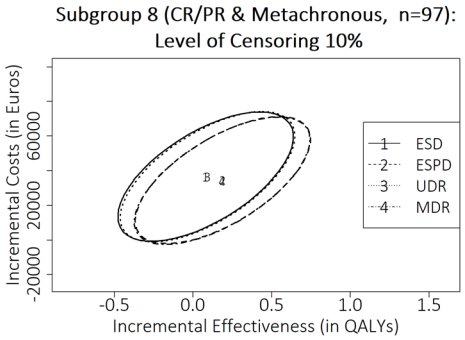

Subgroup 7 (CR/PR & Synchronous, n=269):
Level of Censoring 0%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

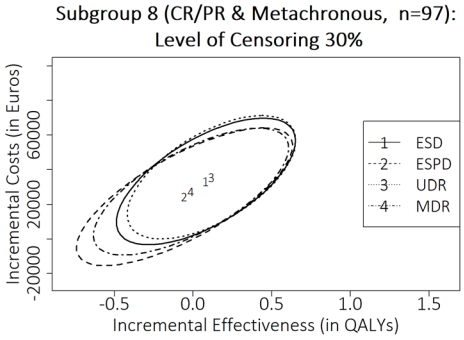

Subgroup 7 (CR/PR & Synchronous, n=269):
Level of Censoring 10%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

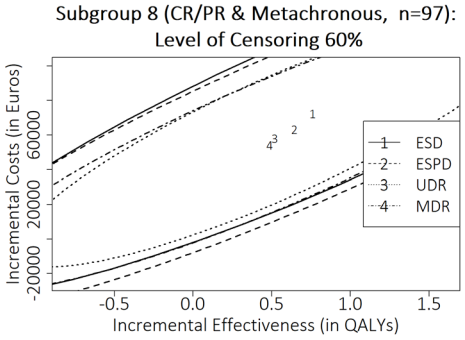

Subgroup 7 (CR/PR & Synchronous, n=269):
Level of Censoring 30%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR


Subgroup 7 (CR/PR & Synchronous, n=269):
Level of Censoring 60%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR


Subgroup 8 (CR/PR & Metachronous, n=97):
Level of Censoring 0%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR


Subgroup 8 (CR/PR & Metachronous, n=97):
Level of Censoring 10%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR


Subgroup 8 (CR/PR & Metachronous, n=97):
Level of Censoring 30%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR


Subgroup 8 (CR/PR & Metachronous, n=97):
Level of Censoring 60%
Incremental Costs (in Euros)
-20000 20000 60000
-0.5 0.0 0.5 1.0 1.5
Incremental Effectiveness (in QALYs)
1 ESD
2 ESPD
3 UDR
4 MDR

PART 3

Metamodeling to reduce computational burden of analyzing simulation models

CHAPTER 8

A scoping review of metamodeling applications and opportunities for advanced health economic analyses

Koen Degeling
Maarten J. IJzerman
Hendrik Koffijberg

This chapter has been published as:
Degeling K, IJzerman MJ, Koffijberg H. A scoping review of metamodeling applications and opportunities for advanced health economic analyses. Expert Review of Pharmacoeconomics & Outcomes Research. 2019;19(2):181-7.

ABSTRACT

Introduction: Metamodels, also known as meta-models, surrogate models, or emulators, are used in several fields of research to negate runtime issues with analyzing computational demanding simulation models. This study introduces metamodeling and presents results of a review on metamodeling applications in health economics.

Areas covered: A scoping review was performed to identify studies that applied metamodeling methods in a health economic context. After search and selection, 13 publications were found to employ metamodeling methods in health economics. Metamodels were used to perform value of information analysis (n=5, 38%), deterministic sensitivity analysis (n=4, 31%), model calibration (n=1, 8%), probabilistic sensitivity analysis (n=1), or optimization (n=1, 8%). One study was found to extrapolate a simulation model to other countries (n=1, 8%). Applied metamodeling techniques varied considerably between studies, with linear regression being most frequently applied (n=7, 54%).

Expert commentary: Although it has great potential to enable computational demanding analyses of health economic models, metamodeling in health economics is still in its infancy, as illustrated by the limited number of applications and the relatively simple metamodeling methods applied. Comprehensive guidance specific to health economics is needed to provide modelers with the information and tools needed to utilize the full potential of metamodels.

INTRODUCTION

Health economic decision modeling is common practice in evaluations of health care interventions informing policy decisions on reimbursement, resource allocation, and research prioritization. However, ongoing personalization within health care creates additional complexities for commonly used health economic modeling methods, such as cohort state-transition models, with multiple clinical pathways based on individual risk profiles and preferences [1-3]. Consequently, there seems a trend towards the use of microsimulation methods, such as patient-level state-transition modeling [4] and discrete event simulation [5]. Although such modeling methods allow individual patients to be traced throughout simulations, their development may be more complex, and their analysis requires more computational power compared to traditional cohort simulation models [6].

Long runtimes required for simulating sufficient hypothetical patients to obtain stable outcomes has been a major drawback of the analysis of patient-level simulation models [7]. Despite advances in computational power, there are still scenarios in which available computer resources are insufficient to perform computationally demanding analyses within reasonable time horizons. Depending on the analysis that needs to be performed, runtime may increase in magnitude and become infeasible. For example, a runtime of a single (deterministic) model simulation of one minute can result in a runtime of several days, if a probabilistic sensitivity analysis of 10,000 samples is performed using that model. Even if this runtime would be considered feasible, repeatedly performing this analysis, for example to apply optimization algorithms or to perform value of information analysis using an inner and outer simulation loop [8, 9], would no longer be feasible.

Metamodels, also referred to as meta-models, surrogate models, or emulators, can be used to negate runtime issues of complex models and analyses, by approximating the outcome of computationally demanding models within feasible time [5, 10]. More specifically, a metamodel is defined as a function that approximates the output of a simulator (i.e., original model) based on input provided to that simulator [11]. Since such an approximation function can typically be evaluated almost instantaneously, it can be used to reduce runtime issues with computationally demanding analyses as simulator substitute. For example, regression techniques can be used to estimate relations between model inputs and outputs based on results of a probabilistic sensitivity analysis. This regression model can subsequently be used to provide a fast approximation of model outcomes based on model inputs and, thereby, perform other computationally demanding analyses. This application of metamodels as substitutes to perform additional simulation-based analyses, can be classified as simulation modeling. Alternatively, metamodeling applications can be classified as statistical modeling if their objective and use is to obtaining insights into relations between simulator inputs and outputs.

Although the field of metamodeling has been studied well in areas such as engineering [12, 13] and environmental modeling [14, 15], implementation in health economics appears to be limited so far. Besides that modelers might not feel a need to apply metamodeling methods, this may also be due to unfamiliarity or a lack of guidance on the development of metamodels, which involves design choices such as selecting the method to be used, structuring the

model, structuring the dataset used to fit the metamodel, and fitting the metamodel [16-18]. Although development of metamodels may seem challenging, when developed, metamodels of complex (patient-level) health economic models have great potential to substantially decrease model runtime and, thereby, enable complex analyses to be performed within feasible time horizons.

This study aims to provide an overview of previous and potential applications of metamodeling methods in health economics. To this end, a scoping review is performed to identify prior metamodeling studies and, thereby, potential metamodeling applications in a health economic context. Although metamodeling is primarily introduced here as a means towards reducing runtime issues with computationally demanding analyses, the scope of this review is not restricted to simulation modeling, as applications classifiable as statistical modeling are also considered. Results of this study should create awareness of, and provide an introduction to, the application of metamodeling methods in health economics.

SCOPING REVIEW ON METAMODELING APPLICATIONS IN HEALTH ECONOMICS

A literature review was performed to identify possible applications of metamodeling in health economics. The search was performed in PubMed on 19 February 2018, employing primary search terms on metamodeling, meta-modeling, surrogate modeling, and emulators, combined with well-known and previously used key words on modeling, simulation, and health economics in title or abstract [19]. Refer to Appendix 1 for the exact search strategy. PubMed was consulted as single database, as result of an iterative strategy in which additional potentially relevant databases were identified by cross-referencing from the initial search results. Since studies that apply metamodeling methods in a health economic context were expected to be published in the general health economic literature, and PubMed has good coverage of this literature, this database was selected first. Because no publications from journals that are not indexed in PubMed were obtained by cross-referencing, no additional databases were searched. Health economic MeSH terms were not included in the search strategy, since a recommended search strategy including MeSH terms [20] was considered too narrow. Hence, a previously used broader strategy for identifying health economic publications was adopted. No exclusion criteria were applied during the search in PubMed, i.e. all results were included in an initial sample for screening based on title and abstract.

Publications were assessed based on title and abstract by one reviewer (KD) and excluded if no modeling methods were applied or if a modeling study did not relate to: health economics, health logistics, or epidemiology, as the objective of the review was to identify previous and potential applications of metamodeling methods in health economics. Consequently, methodological studies discussing a theoretical framework for metamodeling in a health economic context without an (illustrative) application, were excluded. Next, publications in the refined sample were assessed for inclusion based on full-text and the same exclusion criteria. In this process, a proposition whether to include or exclude a publication was made by KD. For each publication in the refined sample, the final decision to include a publication was based on consensus between KD and HK by either adopting or rejecting the proposed inclusion or

exclusion. Excluded publications were categorized according to the reason for exclusion. The included sample was enriched by cross-referencing based on full-text [21], which resulted in the final sample used for analysis. Inclusion of cross-references was also based on consensus between KD and HK following a proposition by KD.

The final sample was analyzed to identify general study characteristics, metamodeling specific information, and cross-references. More specifically, the following general study characteristics were identified: year of publication, journal of publication, clinical context, type of study, and method used for developing the original model (i.e., simulator). Identified information specifically on metamodeling included: metamodeling technique used, analysis performed using the metamodel, design of experiments used for generating input data for fitting metamodels, type of input parameters, type of output parameters, and software used for metamodel development. These aspects were selected to identify why a metamodel was developed, which methods were used to develop this metamodel, and how these metamodeling methods were applied. Study characteristics and metamodeling information were extracted by KD following a full-text review of all publications, the results of which were discussed with HK for validation.

RESULTS

The literature search yielded 478 publications, which were all included in the initial sample for title and abstract screening (Figure 1). A total of 452 publications (95%) were excluded based on title and abstract. These publications predominantly related to in vitro and in vivo research (n = 186, 41%), engineering research (n = 68, 15%), environmental research (n = 57, 13%), and biomedical research (n = 45, 10%). From the resulting 26 publications in the refined sample, 16 publications (62%) were excluded based on full-text screening. These publications were excluded, because the referred-to metamodels were: 1) models developed using an identical modeling technique yet with different model parameters (n = 3, 19%), 2) meta-analyses for defining prediction or classification models (n = 3, 19%), 3) (prediction) models developed for use in non-health economic contexts (n = 3, 19%), 4) ensembles or comparisons of previously developed models (n = 2, 13%), 5) related to biomedical research (n = 2, 13%), 6) related to environmental research (n = 2, 13%), or 7) related to information systems research (n = 1, 6%). One publication (6%) was excluded because the full-text was not available, and could also not be obtained from the publication authors [22]. The included sample [23-32] was enriched by three cross-references [33-35], resulting in a final sample of 13 publications for analysis, representing a total of 14 case studies, as one publication includes two different case studies [27].

Characteristics of publications included in the final sample are presented in Table 1 and results following extraction of key features with respect to the applied metamodeling methods are presented in Table 2. As presented in Table 1, clinical contexts of the studies included: inflammatory arthritis, human immunodeficiency virus, excessive alcohol consumption, influenza, varicella, breast cancer, renal transplantation, multiple sclerosis, osteoporosis, and deep vein thrombophlebitis. One study did not focus on a specific disease area but mod-

eled an emergency department. Three studies employed fictitious models, representing a hypothetical case study. Most publications represented traditional health economic studies that compare cost and effectiveness outcomes for different treatment strategies (n = 10 / 13, 77%), two represented infectious disease modeling studies (15%), and one represented a health care logistics study (8%). Applied modeling techniques for the original model (i.e., simulator) were: cohort state-transition modeling (n = 6 / 14, 43%), agent based modeling (n = 3 / 14, 21%), decision tree analysis (n = 2 / 14, 14%), microsimulation state-transition modeling (n = 2 / 14, 14%), and hybrid microsimulation modeling (n = 1 / 14, 7%). Finally, an upward trend in applications of metamodeling methods over time can be observed from Tables 1 and 2.

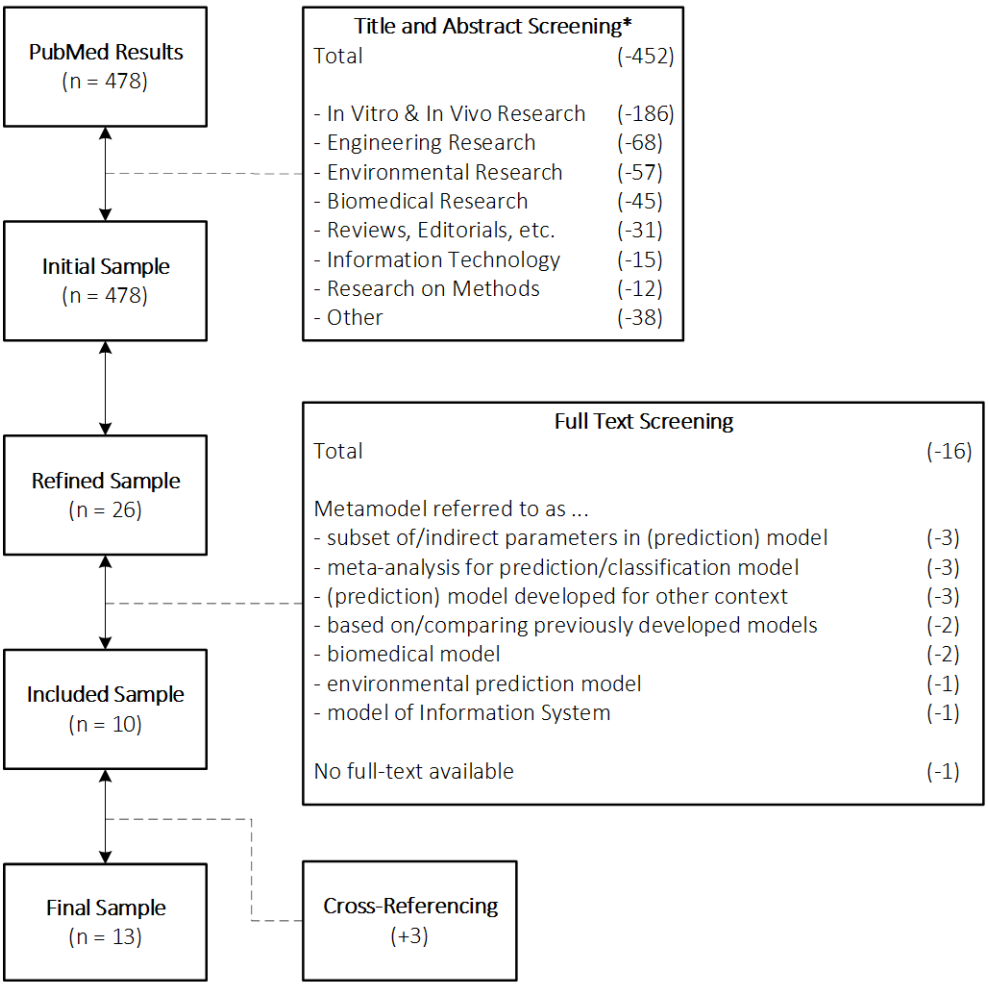


Figure 1. *Graphical representation of the search and selection process.*
**Exclusion categories not mutually exclusive.*

Metamodels were most often developed and applied to perform value of information analyses (n = 5 / 13, 38%), followed by deterministic sensitivity analyses (n = 4 / 13, 31%). With regard to the latter, metamodels were either used to obtain stable outcomes over multiple runs with the same input values, or to obtain insights into parameter importance directly. Single studies employed metamodeling for model calibration, optimization, probabilistic sensitivity analysis, and extrapolation to other countries. The metamodel used for extrapolation to other countries was developed based on a parameter subset of a country-specific simulator, and applied to estimate outcomes for other countries after gathering data on the included parameters for these countries. Of these applications, obtaining stable outcomes over multiple runs, performing model calibration, probabilistic sensitivity analysis, value of information analyses, and optimization, directly relate to negating runtime issues as reason for employing metamodeling methods. Most publications included metamodels that were based on linear (logistic) regression (n = 8 / 13, 62%) or Gaussian Process modeling (n = 4 / 13, 31 %). Other applied metamodeling techniques were (ensembles of) neural networks, symbolic regression, and regression splines. Methods used to generate samples for fitting metamodels, i.e. design of experiments, included: sampling using deterministic ranges for one parameter at a time, random sampling using either uniform or parameter-specific distributions (i.e., Monte Carlo sampling), Latin Hypercube sampling, and fractional and full factorial designs. Eight studies (62%) defined their metamodel using the same parameters that defined the simulator, whereas others (also) used subsets of simulator inputs (n = 7 / 13, 54%). Outcomes of interest, i.e. dependent variables or response variables, included (aggregated) health economic outcomes (e.g., costs, quality adjusted life-years, and opportunity loss), as well as epidemiological outcomes (e.g., disease population size and clinical attack rate) and process outcomes (e.g., average length of stay).

CONCLUSIONS

Applications of metamodeling methods in health economics are scarce. Studies applying metamodeling methods to negate runtime issues with computational demanding analyses of health economic models illustrate its potential to serve several objectives, such as obtaining stable outcomes over multiple runs with the same input values, performing model calibration, probabilistic sensitivity analyses, value of information analyses, and optimization. Additionally, not directly relating to runtime issues, metamodels have been used to obtain insights into parameter importance (i.e., deterministic sensitivity analyses) and for extrapolation to other countries. Linear (logistic) regression, symbolic regression, regression splines, Gaussian Processes, and neural networks were techniques used to develop metamodels. Most original health economic models, i.e. simulators, were cohort-based, rather than microsimulation, i.e. patient-level, simulation models. Different design of experiments methods were used to generate input data for fitting metamodels, including: random designs, full factorial designs, fractional factorial designs, and Latin Hypercube designs. Predicted outcomes, i.e. response variables or dependent outcomes, were mostly health economic outcomes, such as costs and quality adjusted life-years.

Table 1. *Study characteristics of publications included in the final sample for analysis. NA, not applicable.*

First Author	Year of Publication	Journal of Publication	Clinical Context	Type of Study	Simulator Type
Merz	1992	Medical Decision Making	Deep Vein Thrombophlebitis	Health Economic Modeling	Decision Tree Model
Tappenden	2004	Health Technology Assessment	Multiple Sclerosis	Health Economic Modeling	Cohort State-transition Model
Stevenson	2004	Medical Decision Making	Osteoporosis	Health Economic Modeling	Microsimulation State-transition Model
Woodroffe	2005	Health Technology Assessment	Renal Transplantation	Health Economic Modeling	Microsimulation State-transition Model
Rojnik	2008	Value in Health	Breast Cancer	Health Economic Modeling	Cohort State-transition Model
Jalal	2013	Medical Decision Making	NA (Fictitious)	Health Economic Modeling	Cohort State-transition Model
Willem	2014	PLoS Computational Biology	Influenza & Varicella	Infectious Disease Modeling	Agent-Based Model
Jalal	2015	Medical Decision Making	NA (Fictitious)	Health Economic Modeling	Decision Tree Model & Cohort State-transition Model
Andrianakis	2015	PLoS Computational Biology	HIV	Infectious Disease Modeling	Agent-Based Model
Angus	2016	European Journal of Public Health	Excessive Alcohol Consumption	Health Economic Modeling	Hybrid Microsimulation Model
Jutkowitz	2017	Pharmacoeconomics	Gout, Inflammatory Arthritis	Health Economic Modeling	Cohort State-transition Model
Yousefi	2018	Artificial Intelligence in Medicine	Emergency Department	Health Care Logistics	Agent-Based Model
Jalal	2018	Medical Decision Making	NA (Fictitious)	Health Economic Modeling	Cohort State-transition Model

EXPERT COMMENTARY

Metamodeling methods have traditionally been applied to address runtime issues with computationally expensive simulation models and their corresponding analysis in different fields of research, such as engineering and environmental modeling. Although computational burden in health economics may typically be perceived as matter of concern only for microsimulations, results show that runtime issues are not only applicable to analyses of patient-level simulation models, but also to analyses of cohort-based simulation models. The growth in metamodeling applications over time may, therefore, not only be due to an increase in health economic publications or the use of microsimulation methods, such as patient-level state-transition modeling and discrete event simulation, but also due to an increase in applications of advanced model analyses, such as value of information analyses [36].

According to its aim of reducing runtime issues associated with computational demanding analyses [5, 10], most identified applications of metamodeling in health economics were aimed at addressing runtime issues with analyzing cohort-based or patient-level simulation models, for example to perform model calibration, probabilistic sensitivity analyses, value of information analyses, or optimization. Interestingly, applications were not limited to those addressing computational issues, as metamodels have also been developed to obtain insights into parameter importance [26, 27, 33] and to extrapolate simulation models to other countries [30]. One publication developed and used a metamodel to obtain stable outcomes over multiple simulator runs [24], though this indirectly relates to addressing runtime issues, because running this simulator for a sufficiently large number of hypothetical patients would also stabilize outcomes. The required number of simulated individuals might be so high, however, that it becomes infeasible to perform the required analyses, i.e. runtime is an issue.

The different identified reasons for applying metamodels show these methods' potential to address several challenges in the field of health economic modeling, not only from a simulation modeling perspective but also from a statistical modeling perspective. Especially the use of metamodeling methods to obtain insights in characteristics of simulators, i.e. original models, appears from the review results. Such analyses can be classified as statistical modeling, and general methods to this end have been described previously [37]. Although straightforward structures to interpret, such as simple linear regression models, can be used as a first step in performing comprehensive sensitivity analyses [38], no useful information regarding simulator characteristics can be obtained from more complex metamodeling structures, such as symbolic regression or Gaussian processes, as these structures are hard to interpret. Use of such complex structures in a simulation modeling context to perform fast simulations to approximate simulator outcomes, however, is straightforward, which may be an explanation of why most identified applications used metamodeling as a toolbox to this end.

Despite computational challenges, successful applications, and ample utilization opportunities, metamodeling in health economics appears to be in its infancy, as only a small number of applications could be identified. Reasons for this limited number of applications may include the absence of a need to perform computer resource demanding analyses, for example because these are not required by decision makers, or that only few health economic model-

Table 2. *Results of analyzing the final sample to identify metamodeling specific information. MC, Monte Carlo; GLIMS, Generalized Linear Interactive Modeling System; ENB, Expected Net Benefit; QALYs, Quality Adjusted Life-Years; GEMSA, Gaussian Emulation Machine for Sensitivity Analysis; NHB, Net Health Benefit.*

First Author	Year of Publication	Metamodeling Technique	Analysis with Metamodel
Merz	1992	Logistic Regression	Deterministic Sensitivity Analysis (Parameter Influence)
Tappenden	2004	Linear Regression & Gaussian Process	Value of Information
Stevenson	2004	Gaussian Process	Probabilistic Sensitivity Analysis
Woodroffe	2005	Linear Regression	Deterministic Sensitivity Analysis (Stable Outcomes)
Rojnik	2008	Linear Regression & Gaussian Process	Value of Information
Jalal	2013	Linear Regression	Deterministic Sensitivity Analysis (Parameter Influence)
Willem	2014	Symbolic Regression	Deterministic Sensitivity Analysis (Parameter Influence)
Jalal	2015	Linear Regression	Value of Information
Andrianakis	2015	Gaussian Process	Model Calibration
Angus	2016	Linear Regression	Extrapolation to other countries
Jutkowitz	2017	Linear Regression	Value of Information
Yousefi	2018	(Ensembles of) Neural Networks	Optimization
Jalal	2018	Spline Regression	Value of Information

ers are aware of these methods and their potential. Another reason for the limited number of applications may be the lack of comprehensive guidance, specifically in a health economic context. Some guidance has been identified in the reviewed sample [23, 25, 27, 32], though this guidance is not coherent, nor comprehensive. Furthermore, guidance available for metamodeling in other fields of research [16-18, 39-41] is not completely transferrable to, and comprehensive for, application in health economic analyses.

The limited uptake of, and guidance for, metamodeling in health economics is also illustrated by the diversity of metamodeling methods identified in the review. Linear regression was the technique applied for developing the majority of metamodels, and although linear regression is well known due to its prominent role in statistics, it is rarely used for metamodeling in other fields of research, because more accurate and flexible methods are available. These more ad-

Design of Experiments	Simulator Parameters	Metamodeling Outcome	Metamodeling Software
Random Sampling (MC)	All	Preferred Treatment Strategy	GLIMS
Random Sampling (MC) & Unknown Efficient Design	All & Subset	ENB	Matlab
Random Sampling (MC)	Subset	Costs, QALYs & Years on Pharmaceutical Treatment	-
Deterministic Ranges	Subset	Costs & QALYs	-
Latin Hypercube Sampling & Random Sampling (Uniform)	All	Costs & QALYs	SAS & GEMSA
Random Sampling (MC), One-Factor-at-a-Time & Full Factorial Design	All	Costs, QALYs & NHB	Stata
Latin Hypercube Sampling	Subset & All	Clinical Attack Rate, Moment of Epidemic Peak & QALYs	Mathematica
Random Sampling (MC)	All	Incremental Net Benefit	R & MS Excel
Latin Hypercube Sampling	Subset	Population size, Incidence & Prevalence	-
Fractional Factorial Design	Subset	Costs & QALYs	Stata
Random Sampling (MC)	All	Opportunity Loss	R
Latin Hypercube Sampling	Subset	Average Length of Stay	Matlab
Random Sampling (MC)	All	Opportunity Loss	R

vanced methods, however, are more complex to fit, interpret, and apply compared to linear regression models. Furthermore, the variety of design of experiments methods used, and the lack of reasoning why these were applied, suggests that comprehensive guidance in health economic literature as a central source of information would be valuable to guide such design choices.

Our study has certain limitations. First of all, only one database was searched, which may have partly caused a low number of publications to be included in the final sample. This limitation might have caused publications just outside the standard health economic context, e.g. in operations research journals that are not indexed in PubMed, to be excluded from our results. Additionally, cross-referencing was applied to minimalize the potential impact of both these limitations. The three publications identified by cross-referencing do not compromise

our search strategy. One cross-reference used the term emulator, though this was not identified in the full-text by the search machine [34]. The two remaining cross-references did not contain any health economic terms in their title or abstract [33, 35]. Finally, results of literature reviews of modeling studies are generally prone to underreporting, which also relates to the limited space available for publications in peer-reviewed journals, potentially causing authors to exclude details about metamodeling methods from their publications.

FIVE-YEAR VIEW

In order for health economic models to persist in informing decision making by meaningfully representing increasingly complex clinical pathways, these models will continue becoming more and more sophisticated, and their corresponding analysis will thus become progressively computationally demanding. Additionally, new consensus guidance on advanced analyses in health economics, such as for value of information analyses and optimization [42], is expected to become available on the short term, and will boost the application of these analyses. Since these analyses may place a large burden on computational resources, it may not be possible to carry them out within feasible time horizons using standard health economic modeling methods. Metamodeling methods provide opportunities to address these challenges by providing (almost) instant approximations of simulator outcomes. Consequently, the use of metamodels in health economics is expected to rise with an increase in applications of demanding analyses.

Currently, comprehensive guidance on the application of metamodeling methods in health economics is being developed as part of a successive study. When this guidance becomes available, and health economic modelers become aware of the potential of these methods, applications of metamodeling methods in health economics may increase substantially. Health economic modelers will be provided with an extensive set of tools to perform advanced computationally demanding analyses, which were largely infeasible to perform beforehand. This is likely to work both ways, as not only more studies will utilize the value of metamodeling methods in health economics, but more studies will also perform advanced analyses with their simulation models. Eventually, when results of analyses that are currently not being performed due to unacceptable runtime become available, decision makers may be better informed and, thereby, health economics outcomes, both in terms of patient outcomes and health care costs, may improve.

KEY ISSUES

- Metamodels estimate the output of a simulator (i.e., original model) using an approximation function or structure that can be evaluated (almost) instantaneously.
- Metamodeling methods can be used to negate runtime issues associated with advanced analyses of health economic models, such as model calibration, probabilistic sensitivity analyses, value of information analyses, and optimization.
- Metamodeling in health economics is in its infancy, both in terms of the number of applications and in the methods that are being applied.

- Comprehensive guidance on how to apply metamodeling methods in health economics is needed to provide modelers the information and tools needed to utilize the full potential of these methods.

DECLARATION OF INTEREST

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

CONTRIBUTION STATEMENT

The research design, including the search strategy and inclusion criteria, was developed as a combined effort of all authors. Performing the search and primary analysis was done by KD in close cooperation with, and under supervision of HK. Interpretation of the results was a combined effort of all authors. The initial manuscript was drafted by KD and revised by HK and MIJ. The overall guarantor of this study is HK.

REFERENCES

1. Annemans L, et al. Current methodological issues in the economic assessment of personalized medicine. Value Health. 2013;16(S6):S20–6.

2. Miller JD, et al. Current challenges in health economic modeling of cancer therapies: a research inquiry. Am Health Drug Benefits. 2014;7(3):153–162.

3. Phillips KA, et al. Challenges to the translation of genomic information into clinical practice and health policy: utilization, preferences, and economic value. Curr Opin Mol Ther. 2008;10(3):260–266.

4. Siebert U, et al. State-transition modeling: a report of the ISPOR-SMDM modeling good research practices task force-3. Value Health. 2012;15(6):812–820.

5. Karnon J, et al. Modeling using discrete event simulation: a report of the ISPOR-SMDM modeling good research practices task force-4. Value Health. 2012;15(6):821–827.

6. Brennan A, et al. A taxonomy of model structures for economic evaluation of health technologies. Health Econ. 2006;15(12):1295–1310.

7. Karnon J, Haji Ali Afzali H. When to use discrete event simulation (DES) for the economic evaluation of health technologies? A review and critique of the costs and benefits of DES. Pharmacoeconomics. 2014;32(6):547–558.

8. Heath A, et al. A review of methods for analysis of the expected value of information. Med Decis Mak. 2017;37(7):747–758.

9. Madan J, et al. Strategies for efficient computation of the expected value of partial perfect information. Med Decis Mak. 2014;34(3):327–342.

10. Sacks J, et al. Design and analysis of computer experiments. Statist Sci. 1989;4(4):409–423.

11. Law AM. Simulation modeling and analysis. Singapore: McGraw-HillHigher Education; 2007.

12. Li H, Gu J, et al. Multi-objective optimization of coronary stent using kriging surrogate model. Biomed Eng Online. 2016;15(2):148.

13. Xu G, Liang X, et al. Multi-objective aerodynamic optimization of the streamlined shape of high-speed trains based on the kriging model. PloS one. 2017;12(1):e0170803.

14. Butcher JB, et al. Estimating future temperature maxima in lakes across the United States using a surrogate modeling approach. PloS one. 2017;12(11):e0183499.

15. Ouyang Q, et al. Conservative strategy-based ensemble surrogate model for optimal groundwater remediation design at DNAPLs-contaminated sites. J Contam Hydrol. 2017;203:1–8.

16. Barton RR. Simulation metamodels. Proceedings of the 30th conference on winter simulation. Washington, D.C., USA: IEEE Computer Society Press; 1998:167–176.

17. Levy S, Steinberg DM. Computer experiments: a review. Adv StatAnal. 2010;94(4):311–324.

18. Simpson TW, et al. Metamodels for computer-based engineering design: survey and recommendations. Eng Comput. 2001;17(2):129–150.

19. Degeling K, et al. A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models. Expert Rev Pharmacoecon Outcomes Res. 2017;17(1):17–25.

20. Strings attached: CADTH database search filters [Internet]. Ottawa: CADTH; 2016. Available from: https://www. cadth.ca/resources/finding-evidence/strings-attached-cadthsdatabase-search-filters#eco [Accessed Apr 12, 2018]

21. Wolfswinkel JF, et al. Using grounded theory as a method for rigorously reviewing literature. Eur J Inf Syst. 2011;22(1):45–55.

22. Kilmer RA, et al. An emergency department simulation and a neural network metamodel. J Soc Health Syst. 1997;5(3):63–79.

23. Tappenden P, et al. Methods for expected value of information analysis in complex health economic models: developments on the health economics of interferon-beta and glatiramer acetate for multiple sclerosis. Health Technol Assess. 2004;8(27):iii,1–78.

24. Woodroffe R, et al. Clinical and cost-effectiveness of newer immunosuppressive regimens in renal transplantation: a systematic review and modelling study. Health Technol Assess. 2005;9(21):1–179, iii-iv.

25. Rojnik K, Naversnik K. Gaussian process metamodeling in Bayesian value of information analysis: a case of the complex health economic model for breast cancer screening. Value Health. 2008;11 (2):240–250.

26. Jalal H, et al. Linear regression metamodeling as a tool to summarize and present simulation model results. Med Decis Mak. 2013;33(7):880–890.

27. Willem L, et al. Active learning to understand infectious disease models and improve policy making. PLoS Comput Biol. 2014;10(4):e1003563.

28. Andrianakis I, et al. Bayesian history matching of complex infectious disease models using emulation: a tutorial and a case study on HIV in Uganda. PLoS Comput Biol. 2015;11(1):e1003968.

29. Jalal H, et al. Computing expected value of partial sample information from probabilistic sensitivity analysis using linear regression metamodeling. Med Decis Mak. 2015;35(5):584–595.

30. Angus C, et al. Estimating the cost-effectiveness of brief interventions for heavy drinking in primary health care across Europe. Eur J Public Health. 2017;27(2):345–351.

31. Jutkowitz E, et al. Prioritizing future research on allopurinol and febuxostat for the management of gout: value of information analysis. Pharmacoeconomics. 2017;35 (10):1073–1085.

32. Yousefi M, et al. Chaotic genetic algorithm and adaboost ensemble metamodeling approach for optimum resource planning in emergency departments. Artif Intell Med. 2018;84:23–33.

33. Merz JF, et al. Measuring decision sensitivity: a combined monte carlo-logistic regression approach. Med Decis Mak. 1992;12(3):189–196.

34. Stevenson MD, et al. Gaussian process modeling in conjunction with individual patient simulation modeling: a case study describing the calculation of cost-effectiveness ratios for the treatment of established osteoporosis. Med Decis Mak. 2004;24(1):89–100.

35. Jalal H, Alarid-Escudero F. A gaussian approximation approach for value of information analysis. Med Decis Mak. 2018;38(2):174–188.

36. Steuten L, et al. Critical review of the evolving methods and applications of value of information in academia and practice. PharmacoEconomics. 2013;31(1):25–48.

37. Kleijnen JPC. An overview of the design and analysis of simulation experiments for sensitivity analysis. Eur J Oper Res. 2005;164(2):287–300.

38. Saltelli A, et al. Global sensitivity analysis: the primer. West Sussex: Wiley; 2008.

39. Cozad A, et al. Learning surrogate models for simulation-based optimization. AIChE J. 2014;60(6):2211–2227.

40. Kleijnen JPC, Sargent RG. A methodology for fitting and validating metamodels in simulation. Eur J Oper Res. 2000;120(1):14-29.

41. Li YF, et al. A systematic comparison of metamodeling techniques for simulation optimization in decision support systems. Appl Soft Comput. 2010;10(4):1257–1273.

42. Crown W, et al. Constrained optimization methods in health services research- an introduction: report 1 of the ISPOR optimization methods emerging good practices task force. Value Health. 2017;20(3):310–319.

APPENDIX 1: search strategy

We performed a search in PubMed using the search strategy specified below.
Primary search terms on metamodel, meta-modeling, surrogate modeling, and emulators:

- "emulator*" referring to emulator, emulators, etc.
- "meta-model*" referring to meta-model, meta-models, meta-modeling, etc.
- "metamodel*" referring to metamodel, metamodels, metamodeling, etc.
- "surrogate model*" referring to surrogate model, surrogate models, surrogate modeling, etc.

Secondary health economic search terms in the title or abstract:

- "*effect*" referring to effect, effects, effectiveness, cost-effectiveness, etc.
- "*util*" referring to utility, cost-utility, risk-utility, etc.
- "*benefit*" referring to benefit, benefits, risk-benefit, etc.
- "*risk*" referring to risk, risks, benefit-risk, etc.
- "*conseq*" referring to consequence, consequences, etc.
- "cost"
- "costs"
- "decision-anal*" referring to decision-analytic, decision-analytical, etc.

Combined, the exact used search algorithm was: *(((((emulator*) OR meta model*) OR metamodel*) OR surrogate model*)) AND (*effect*[tiab] OR *util*[tiab] OR *benefit*[tiab] OR *risk*[tiab] OR *conseq*[tiab] OR cost[tiab] OR costs[tiab] OR decision-anal*[tiab])*

8

CHAPTER 9

How to use metamodeling for reducing computational burden of advanced analyses with health economic models: a structured overview of metamodeling methods in a six-step application process

Koen Degeling
Maarten J. IJzerman
Mariel S. Lavieri
Mark Strong
Hendrik Koffijberg

ABSTRACT

Background: Metamodels can be used to reduce computational burden associated with computationally demanding analyses of health economic models. Challenges remain regarding the selection of appropriate metamodeling methods for use in health economics due to a lack of comprehensive guidance, which may restrict application of these methods. This study aimed to address these challenges by identifying a process for applying metamodeling in health economics, including suitable methods and directions for their selection and application.

Methods: General (i.e., non-health economic specific) metamodeling literature, clinical prediction modeling literature, and a previously published systematic literature search were exploited to consolidate a process, and to identify candidate metamodeling methods. Methods were considered applicable to health economics if they are able to account for mixed (i.e., continuous and discrete) input parameters and continuous outcomes (i.e., response variables).

Results: A process was structured based on six steps that were identified as relevant for applying metamodeling methods in health economics, i.e. 1) identification of suitable techniques, 2) data simulation, 3) metamodel fitting, 4) performance assessment, 5) analysis execution, and 6) results verification. Different methods were discussed to support each step, including their characteristics, directions for use, key references, and available R packages.

Conclusions: To address challenges regarding metamodeling methods selection, a first guide towards using metamodels to reduce runtime issues with the analysis of health economic models was developed. This guidance may increase applications of metamodeling in health economics, enabling increased use of state-of-the-art analyses, e.g. a full probabilistic uncertainty and sensitivity analysis, with computationally burdensome simulation models.

INTRODUCTION

Traditional health economic modeling methods are being challenged due to ongoing personalization within health care [1-3]. Patient-level simulation methods, such as microsimulation state-transition modeling [4] and discrete event simulation [5], have gained attention to complement more common cohort-level simulation models, such as decision trees and Markov models, due to their ability to encode complex care processes and behavioral processes more easily [6, 7]. Furthermore, computationally demanding model analyses, such as value of information analysis [8], are increasingly being performed in addition to traditional model analyses, such as probabilistic sensitivity analysis [9]. Additionally, optimization algorithms are applied to, or within, health economic models to identify screening of treatment strategies that maximize patient outcomes, while accounting for certain constraints [10]. Although these advanced analyses and patient-level simulation models provide valuable information and insights, the time required to evaluate those patient-level models is higher compared to cohort-level simulation models [6, 11], and large numbers of model evaluations may be required for advanced analyses, which induces runtime issues and may hinder application of these methods and analyses within feasible timescales.

Metamodeling methods can be applied to reduce computational burden of computationally demanding analyses with patient-level (or cohort-level) simulation models [5, 12]. Metamodels, also known as surrogate models or emulators, in general can be thought of as functions that approximate outcomes (i.e., response variables or dependent variables) of simulators (i.e., the original simulation model) based on input provided to simulators [13]. These functions are computationally cheap to evaluate, though they may themselves take some time to build and validate. Metamodeling methods are used widely across different fields of science and engineering, for example to optimize designs of coronary stents [14], high speed trains [15], and groundwater remediation [16], as well as to estimate future water temperatures [17].

Runtime issues associated with analyses of health economic models can also be reduced by applying metamodeling methods, as was shown by a recent review of metamodeling applications in health economics [18]. This review identified a limited number of studies, where in most cases, metamodels were developed to facilitate value of information analysis, which typically imposes a considerable additional computational burden beyond more standard health economic analyses. Other analyses performed using metamodels included deterministic sensitivity analyses, probabilistic sensitivity analyses, model calibration, and optimization, all of which can increase computational burden to the extent that using the original simulation model itself becomes infeasible. Interestingly, the metamodeling methods that were applied were relatively basic compared to those used in other fields of research. They also varied considerably between studies, and were often not discussed in detail. Metamodeling in health economics appears still to be in its infancy, in part because most health economic models and applied analyses have, until recently, been relatively simple.

The limited number of examples of metamodeling being used in computationally burdensome health economic evaluation studies may be either due to a lack of awareness of the

potential of metamodeling methods to reduce runtime, or due to a lack of guidance on how to apply these methods in health economics [18]. This need for guidance may also explain the diversity in methods applied and lack of clear communication on key metamodeling aspects in corresponding publications. Guidance on selecting and applying metamodeling methods in a health economics context may therefore enhance their adoption.

This study aims to provide a structured overview of metamodeling methods that can be used to reduce computational burden of analyses with health economic models, based on previous metamodeling applications in health economics and guidance available from other fields of research. Additionally, points of consideration for selecting and applying these methods are discussed, including directions specific to health economics. This study should increase awareness amongst health economic modelers of opportunities associated with metamodeling, as well as methods available and instructions for their use.

METHODS

Several publications that provide information on steps to be taken when applying metamodeling in health economics [19-22] were identified in a recent review of metamodeling applications in health economics [18] and used as basis for this study. Since steps included and methods applied differed substantially between these publications, they were complemented with guidance available from general, i.e. not health economic-specific, metamodeling literature. Additionally, guidance on prediction modeling in medical research, the process of which shows similarities to that of metamodeling, was used to identify steps to be taken and methods available.

More specifically, metamodeling methods and steps to be taken when applying them, were identified by a scoping literature search that was performed by KD and involved online searches, searches in Scopus and PubMed, and cross-referencing. By using the previously mentioned publications on metamodeling in health economics as a starting point [19-22], method-specific information, other candidate metamodeling methods, and potentially relevant process steps were identified by iterative searches on methods and process steps introduced in these publications and by cross-referencing. For example, if the impact of different experimental designs on metamodel performance was discussed in a paper found from a search on a specific metamodeling technique (i.e., structure of the metamodel), additional searches on these designs of experiments were performed to identify further information on these experimental designs and other designs of experiments. Search terms for a specific method were initially based on its name as referred to in the publication in which it was encountered. The search terms were updated to include synonyms when those were identified. The iterative search process terminated when additionally found literature did not result in further inclusion of methods, i.e. until theoretical saturation was reached.

Metamodeling methods were only included if they are considered appropriate for use in health economics and have been commonly used in other fields of research, in line with the objective of the study. Methods were considered applicable to health economics if they are

able to account for mixed (i.e., continuous and discrete) input parameters and continuous outcomes (i.e., response variables). Relevant steps to be taken when applying metamodeling methods in health economics were not prespecified, but extracted from the literature as described above, and structured in a process. Metamodeling methods and their characteristics were described according to this process, and presented in a table or graphs when appropriate. For example, characteristics of different metamodeling techniques were presented in an overview table. Additionally, examples of packages available to implement methods in R Statistical Software [23] were identified via an online search, and introduced along with the corresponding methods.

RESULTS

A six-step process for metamodeling in health economics was structured, covering methods from selecting suitable metamodeling techniques up to validating metamodel outputs against simulator outputs (Figure 1). As for any type of modeling study, the process is iterative, since new insights may question prior decisions. Depending on the analysis to be performed, the sixth step is facultative, as will be discussed. Next, each process step will be described, including an overview of corresponding methods.

Step 1: identifying candidate metamodeling techniques

Identification of theoretically suitable metamodeling techniques is based on study characteristics, including the analysis to be performed, type of input parameters (continuous, categorical, or mixed), number of input parameters, and type of outcome (continuous or categorical). Here, focus is on techniques capable of handling mixed input parameters and continuous outcomes, as these are typically inputs for, and obtained from health economic models. Ideally, one should apply all techniques considered appropriate and use the best-performing model for analysis. However, in presence of time or budget constraints for metamodel development, one might alternatively select one technique and, if the resulting model yields acceptable performance (see Step 4), use this model for analysis.

Tappenden et al. [20] identified five metamodeling techniques for application in value of information analysis: linear regression, response surface methodology, multivariate adaptive regression splines, Gaussian processes, and neural networks. These techniques are complemented with symbolic regression, which was also identified from the review [21], and generalized additive models, which have been used previously for performing value of information analysis [24, 25]. In Table 1, an overview of techniques and their characteristics is provided. For each metamodeling technique, this overview includes the typically required number of observations (which we have defined as low: $n < 500$, or high: $n \geq 500$), number of input parameters it allows (which we have defined as low: $n < 20$, or high: $n \geq 20$), interpretability of the resulting metamodel structure (which we have classified as low: not or barely possible to understand relations between inputs and outputs, moderate: input-output relations can be understood to some extent, or high: input-output relations can be understood), and the description of any R packages available to apply the technique.

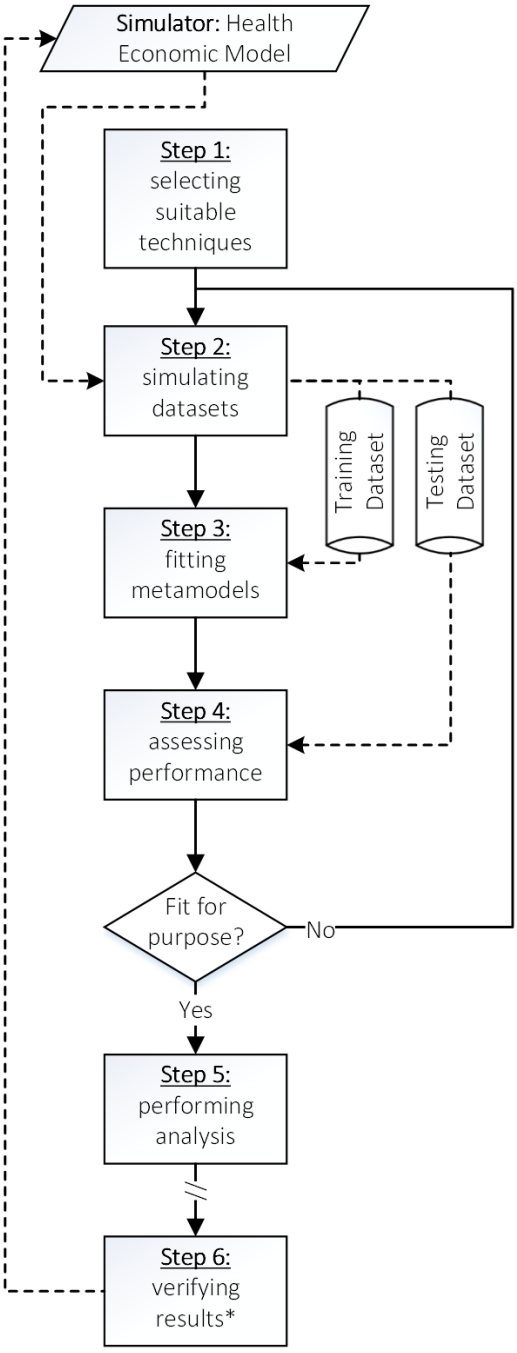


Figure 1. *Process for applying metamodeling methods in health economics.*

Simple linear regression assumes a linear relationship between independent variables (i.e., input parameters) and the dependent variable (i.e., outcome of interest) and is linear in the regression model parameters [26]. These models can easily be fitted to datasets of all sizes, including datasets with large numbers of observations and input parameters, while allowing for both continuous and categorical input parameters. Although fitting linear regression models and interpreting results can be considered relatively easy, more complex techniques, allowing for more flexible model structures, may be more suitable for representing health economic models, which typically induce complex parameter interactions.

Response surface methodology is also linear in the regression model parameters, but does not assume a linear input-output relationship, and fits polynomial regression models to predict responses, i.e. outcomes [27-29]. Both continuous and categorical input parameters can be considered in response surface models, and datasets including large numbers of observations and input parameters can be used. However, high non-linearity will require higher order polynomials, which will require larger numbers of observations and, hence, it will require larger up-front simulator runtime. Although polynomial models are more difficult to interpret compared to linear models, if desired, general trends on model parameter influence can still be extracted from their model structures.

Symbolic regression uses genetic programming to construct a mathematical expression from elementary operators (e.g., '+' and '×') and elementary functions (e.g., 'log') accurately describing the relation between input parameters and the outcome of interest, without making any priori assumption about this relationship [30, 31]. Fitting an accurate symbolic regression model may take substantial time, due to a potentially large number of candidate metamodels, i.e. large solution space. However, symbolic regression is capable of handling large datasets including a large number of mixed input parameters. Symbolic regression models can be difficult to interpret unless the final expression is relatively simple or is simplified.

Multivariate adaptive regression splines were developed to model input-outcome relations that may not be constant across input space [27, 32, 33]. Regression spline modeling divides the outcome domain into intervals, and then estimates an equation, typically a low-order polynomial, for each interval. Different types of splines can be distinguished, based on how the number of intervals and level of smoothness are defined. Fitting multivariate adaptive regression splines includes an automated input parameter importance analysis (see Step 2). Although capable of handling large datasets of mixed input parameters, regression splines are difficult to interpret and prone to overfitting. In contrast to the previously discussed metamodeling techniques, the interpretability of multivariate adaptive regression splines is limited.

Generalized additive models assume that the dependent variable is a smooth, but unknown, function of the independent variables [34, 35]. This unknown underlying smooth function is usually represented using splines, with the cubic spline as a common choice. In its simplest case, a univariate cubic spline represents an arbitrary smooth single-input function as a series of short cubic polynomials joined piecewise such that the function is twice-differentiable at

the ''knots'' (i.e., join points). The same spline can also be represented as the weighted sum of a series of predetermined ''basis functions'' that extend over the whole range of the function input. Simple univariate cubic splines have natural extensions to higher dimensions and to a metamodeling framework, where the spline parameters (i.e., the basis function weights) are estimated from noisy data. Generalized additive models can handle large datasets and high numbers of input parameters, but their structure is difficult to interpret.

Gaussian process regression is a nonparametric regression method also known as Kriging [27, 36]. Gaussian processes use information on neighbor observations for new predictions, while directly providing information on the uncertainty in these predictions. This is unique for metamodeling techniques, since other techniques require additional effort to obtain information on prediction uncertainty. Although Gaussian processes are capable of considering mixed input parameters [37], Treed Gaussian processes have been developed specifically for this type of data [38]. A disadvantage of Gaussian processes is that their interpretability is low, and that computational burden, both in terms of fitting and predicting, increases dramatically with increasing numbers of observations and parameters, limiting their applicability. Hence, Gaussian processes are often well suitable for optimization problems, which are typically defined by limited numbers of decision parameters. Furthermore, input parameter importance analysis can be performed to reduce the number of parameters (see Step 2) and, thereby, computational burden.

Table 1. *Overview of candidate metamodeling techniques for application in health economics, which are all able to account for mixed input parameters and continuous outcomes. * No longer maintained by authors.*

Metamodeling Technique	Required Observations	Number of Inputs	Interpretability	R Package	References
Linear Regression	High	Large	High	*stats* [23]	[26]
Response Surface Methodology	High	Large	Moderate	*rsm* [67]	[27-29]
Symbolic Regression	High	Large	Moderate	*RGP** [68]	[30, 31]
Multivariate Adaptive Regression Splines	High	Large	Low	*earth* [69], *mda* [70]	[27, 32, 33]
Generalized Additive Models	High	Large	Low	*gam* [71], *mgcv* [35, 72]	[34, 35]
Gaussian Processes	Low	Low	Low	*GPfit* [73], *tgp* [74, 75]	[27, 36]
Neural Networks	High	Large	Low	*Neuralnet* [76], *nnet* [77]	[27, 33, 39]

Neural networks are non-parametric models that are commonly found in machine learning applications. These models exist of networks of nodes (called neurons) and layers, which learn about relationships between inputs, either continuous or categorical, and outputs, typically using large datasets [27, 33, 39]. Although neural networks are commonly used for classification, they are also able to predict continuous outcomes [40]. Since no assumption regarding simulator structure is made, neural networks may well represent complex, i.e. non-linear, health economic models. Developing large neural networks typically requires large numbers of observations, which may pose challenges regarding obtaining sufficient simulator samples. Similar to multivariate adaptive regression splines, generalized additive models, and Gaussian processes, neural networks are "black boxes" that are hard to interpret.

In conclusion, simple linear regression may be used for exploratory analyses or when reasonably confident about a linear relationship between model input and output. However, in practice, this relation will often be non-linear. Gaussian processes are particularly useful when obtaining sufficient simulator samples to apply the other techniques is infeasible. Response surface methodology, symbolic regression, multivariate adaptive regression splines, generalized additive models, and neural networks are typically useful when sufficient samples can be obtained from the simulator, i.e. original health economic model. If metamodel interpretation is important, simple linear regression model structures are most easily interpreted, followed by response surface methodology and symbolic regression.

Step 2: simulating datasets

Simulating data from the simulator is crucial in metamodeling studies, as metamodel performance is highly dependent on the data used for fitting [13]. Modelers control the number and definition of experiments (i.e., observations) used for fitting metamodels, which is fundamentally different from prediction modeling studies, for which data is typically observed from clinical studies or registries [26]. Furthermore, challenges regarding handling missing data, reversed causality, omitted variables, and measurement error are not applicable to metamodeling. There are five key aspects to simulating datasets for metamodeling: 1) the number of datasets, 2) sampling method, 3) design of experiments, 4) number of experiments, and 5) analysis used for obtaining simulator outcomes.

As in prediction modeling, two distinct datasets are preferred for metamodeling studies: one for fitting (i.e., training or development dataset) and one for validation (i.e., testing or validation dataset). In prediction modeling studies, validation datasets would typically be obtained by isolating a proportion of the data from a single cohort for internal validation, or by gathering additional data from another 'plausibly related' cohort for external validation [26]. In metamodeling, however, it is preferable to obtain two separate datasets from the simulator, each having a prespecified design with comprehensive coverage. Obtaining one large dataset and separating it in two datasets for training and validation, may compromise the coverage of these datasets: either dataset may lack the structure and properties induced by the design of experiments used to generate the single large dataset. By obtaining two separate datasets, their structure and properties according to the design of experiments used will be maintained, as will be discussed.

Methods for obtaining training datasets can be classified in two categories: single pass methods and iterative methods, also known as adaptive sampling or active learning methods [41, 42]. Single pass methods first define experiments according to a design of experiments method, which will be discussed subsequently, and then evaluate these experiments using the simulator, the results of which are used for fitting the metamodel [43]. However, metamodels fitted to datasets generated based on single-pass methods may lack fidelity (i.e., have low accuracy) if these datasets would contain only a limited number of data points in the region later determined to be of interest (see below). In this case, the use of iterative methods may be desirable. Iterative methods use an initial dataset for fitting an initial metamodel, which is subsequently used in an iterative process to identify additional experiments to be added to the dataset, to update the initial metamodel and check the updated metamodel performance, until this performance is according to a pre-defined threshold [44]. Here, focus is on single pass methods, as these methods are simpler to implement, and do not require simulators to be available in the exact same environment used for generating experiments and fitting the metamodel.

Design of experiments methods determine how experiments, i.e. combinations of parameter values, are selected, to be evaluated from the simulator in order to obtain datasets for fitting and validation [45]. The objective of these methods is to cover parameter spaces and parameter interactions as effectively and efficiently as possible, i.e. with the least number of experiments. Failing to represent the full parameter spaces and parameter interactions will decrease metamodel performance. Commonly used designs are: random designs, full factorial designs, and Latin Hypercube designs.

Random designs, also known as Monte Carlo Sampling methods, obtain n sets of experiments by generating n draws from the joint probability distribution for the input parameters [45]. These designs require a large number of experiments to sufficiently cover the parameter space. Input distributions may be designed to cover a pre-specified range with equal probability (i.e. uniform distributions), or may represent judgements about the true unknown value of some population quantity, for example using Gamma distributions for parameters with a positive range [46]. When a random design is used, one large dataset can be separated in two datasets for fitting and validation, while maintaining its random properties.

Full factorial designs fully enumerate possible combinations of discrete parameter values [43]. More specifically, for n values of k parameters, a full factorial design represents all n^k combinations of these parameter values. Although full factorial designs are able to cover the full parameter space and interactions, the number of experiments exponentially increases with the number of parameters and they are, therefore, often infeasible to use. Fractional factorial designs have been introduced to address challenges regarding high numbers of experiments when using factorial designs, and consist of subsets of full factorial designs [47].

Latin Hypercube designs have been used often for designing computer experiments, as they efficiently cover the full parameter space [45, 48, 49]. In its simplest form, Latin Hypercube samples represent random combinations of values for each parameter, which are equally

spaced between their minimum and maximum value for each parameter. More often, Latin Hypercube samples represent random combinations of random values from equally sized bins that cover the parameters' domains. Over the years, more advanced versions have been developed, such as the maximin Latin Hypercube design [50], which maximizes the minimum distance between design points, and orthogonal Latin Hypercube designs [51].

Figure 2 illustrates how random, full factorial, and maximin Latin Hypercube designs may define nine experiments for two continuous parameters x_1 and x_2. This figure illustrates how parameter spaces are most effectively covered by maximin Latin Hypercube sampling. From this figure, it can also easily be seen why a separate dataset needs to be simulated for model validation when a non-random design is used, as isolating a (random) proportion from a dataset generated according to a full factorial or Latin Hypercube design, for example, will result in a training dataset that no longer covers the full parameter space consistently. If a proportion is being isolated from datasets, an experiment will no longer be defined for every bin in datasets generated according to a Latin Hypercube design, and datasets generated according to a full factorial design will no longer represent all parameter value combinations.

In general, Latin Hypercube designs are preferable for both training and testing datasets, especially when only a limited number of experiments can be evaluated from the simulator in the available time. Optimized Latin Hypercube designs can easily be generated in most software environments, for example using the *lhs* package in R [52]. However, these designs are challenging to apply when constraints on combinations of parameters are applicable. Although some work has been done on conditioned Latin Hypercube designs, accounting for inequality constraints [53], this might not enable all constraints to be accounted for. In such situations, factorial designs can be used if the resulting number of experiments is considered feasible. However, when using factorial designs for continuous variables, a finite set of discrete values within the continuous parameter range needs to be defined, which may result in an infeasible number of experiments to cover those parameters' ranges at the desired level of detail. If using a factorial design is considered infeasible, random designs allow constraints to be accounted for easily. However, random designs are likely less efficient, which may result in suboptimal solution space coverage and, consequently, lower metamodel performance, especially when a limited number of experiments can be evaluated from the simulator.

How many experiments are required, i.e. how large n should be, heavily depends on desired metamodel accuracy, which will be discussed in Step 4. Additionally, the design of experiments method used, and how well the metamodeling techniques match the unknown relation between inputs and outputs, influence the number of experiments required [54, 55]. A general rule of thumb is to start with $n = 10d$, where d refers to input space dimensionality, i.e. number of input parameters [55, 56]. After evaluating model performance for the initial set of experiments (Step 4), n may be increased until the desired level of overall accuracy is achieved (see Step 4 for details). Alternatively, adaptive sampling methods may be applied to improve accuracy in local regions of the parameter space [44], but these methods are outside the scope of this study. If the desired model accuracy cannot be achieved with a feasible number of experiments, importance analysis methods may be applied to reduce input space

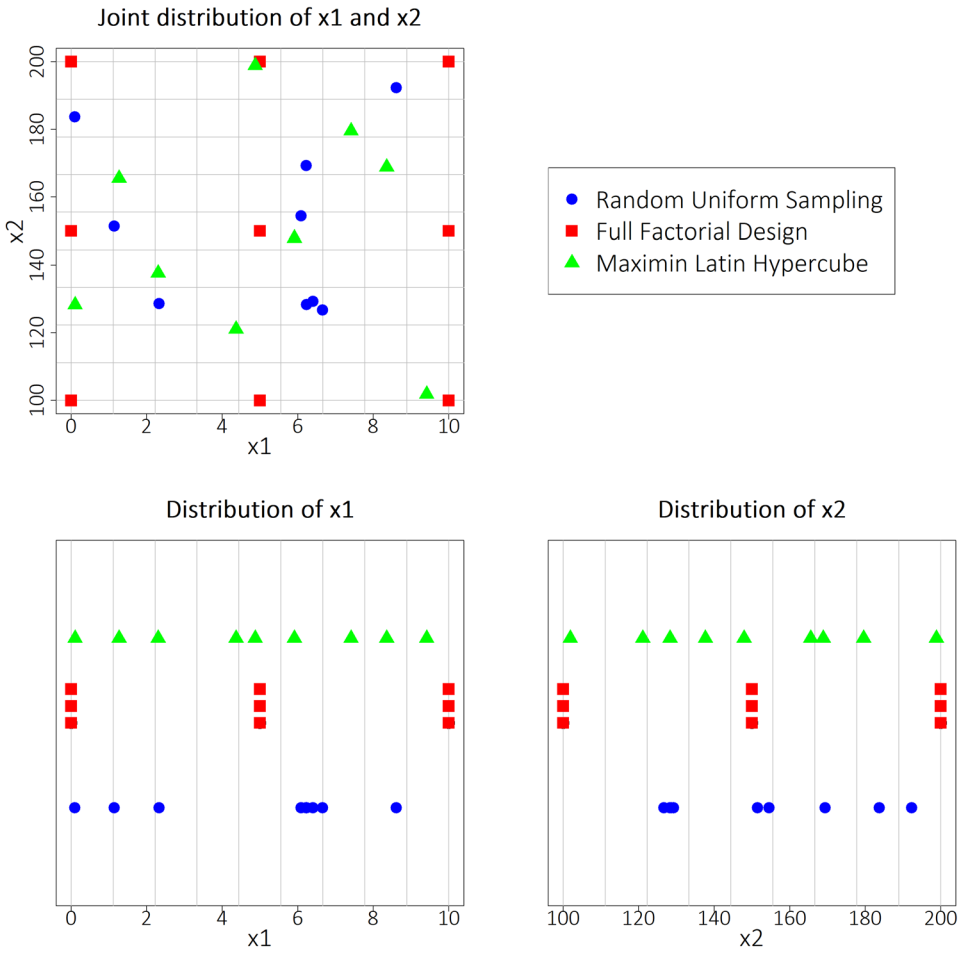


Figure 2. *Illustration of how a random uniform sample, full factorial design, and maximin Latin Hypercube sample may define nine experiments for two continuous parameters* x_1 *and* x_2.

dimensionality d [20, 57, 58]. Reducing dimensionality might result in less complex metamodel structures, and if redundant input parameters can be removed, metamodel accuracy may improve as overfitting is reduced.

Which analysis needs to be performed to evaluate experiments from the simulator, depends on the analysis to be performed with the metamodel. As health economic models tend to be non-linear, outcomes of a base-case analysis based on the expected values of parameters need not be equal to the expected outcome values of a probabilistic sensitivity analysis [59]. Consequently, since expected values of probabilistic sensitivity analyses are the values expected to be observed in reality given the current information, these should be used to fit metamodels used for optimization problems. However, performing a probabilistic sensitivity

analysis as part of each experiment might not be feasible because of the required simulator runtime. In practice, first-order (i.e., stochastic) and second-order (i.e., parameter) uncertainty may not be completely removed if an insufficient number of patients is simulated in stochastic (i.e., patient-level) simulation models or if an insufficient number of model runs is performed for probabilistic sensitivity analyses, which may lead to different outcomes for both base-case analyses and probabilistic sensitivity analyses when these analyses are re-run. The subsequent noise in the data used for fitting has a pernicious effect on metamodel performance and stability of outcomes is, therefore, an important issue. Outcomes can be considered stable, if outcomes obtained from simulations with different random numbers, but with the same input parameter values, are sufficiently similar. What defines "sufficiently similar", differs over case studies and should be discussed with all relevant stakeholders, e.g. care providers, decision makers, and modelers. Obtaining stable outcomes may require a substantial number of patients to be simulated, or simulation runs to be performed, and may not be feasible in practice. However, to reduce the number of patients to be simulated, or number of runs to be performed, variance reduction techniques may be applied, such as using common random numbers when comparing strategies [60, 61] .

Step 3: fitting metamodels

After evaluating an initial set of experiments from the simulator, this training dataset can be used to fit selected metamodeling techniques. Steps involved in fitting processes differ between techniques, as well as any settings to be provided. We refer to the corresponding literature and software documentation to learn about steps to be taken and settings to be provided. As a basic example, some metamodeling techniques or software packages require input parameters to be rescaled. Fitting meta-models is an iterative process, in which settings may be adapted, or more experiments may be evaluated from the simulator, after assessing model performance (Step 4).

Step 4: assessing metamodel performance

Assessing performance of fitted metamodels is essential to further improve that performance, by iteratively improving (extending) the design of the training dataset used, or adapting the settings for fitting these models. Additionally, an initially selected metamodeling technique may be deemed inappropriate if performance does not reach an acceptable level, resulting in exclusion of this technique from the list of potential candidates (Step 1). Performance can be assessed using the testing dataset evaluated from the simulator in Step 2. Since metamodels of health economics models will typically predict continuous scale outcomes, of main interest is to quantify how close predictions are to actual simulator outcomes. Assessing performance and comparing different metamodels can be done graphically and using quantitative performance criteria. A validation plot, with predicted values on the x-axis and observed values from the simulator on the y-axis, is fundamental in assessing model performance, and presents information on systematic trends, as well as general performance. Several quantitative performance criteria are available, including mean or maximum values of the absolute error, absolute relative error, squared error, which may all be normalized using the sample range or standard deviation, and summarized by their mean or maximum values, and R^2 [33, 55, 62, 63].

It is important to be aware of performance criteria characteristics when selecting one, or several, to compare metamodels or to assess whether model performance is acceptable. For example, compared to mean absolute errors, mean squared errors places more weight on outliers. Additionally, compared to squared errors, setting a desired level of accuracy is more straightforward for absolute (relative) errors, as these can be set by answering questions such as "what is the maximum mean deviation in predicted life-years the metamodel is allowed to have compared to the simulator outcomes?" What performance can be considered acceptable for deciding to apply a metamodel for performing analyses (Step 5), differs over case studies and should be based on input from all stakeholders. Since different performance criteria and definitions of acceptable performance may yield alternative conclusions, these should be decided upon prior to metamodel development.

Step 5: applying metamodel

Once a metamodel has been selected, it can be used to perform analyses that could not be performed in a feasible time period with the original health economic model. Previous applications of metamodels in health economics include value of information analysis, model calibration, optimization, probabilistic sensitivity analysis, and obtaining stable outcomes over multiple runs with the same input values [18]. Additionally, metamodels can be used in online tools for which limited computer resources are available.

Step 6: verifying results (optional)

Depending on which analysis is performed in Step 5, it may be preferable to verify results with those that can be obtained from the simulator. In optimization problems for example, it is recommended to re-evaluate a certain number of best strategies identified by the metamodel using the original health economic model, to assess whether their outcome and ordering meaningfully differ. By providing these results, decision makers are better informed about the expected impact of choosing a good but not optimal candidate strategy for implementation, which may be favored over the optimal strategy for practical reasons. For other types of analyses, such as probabilistic sensitivity analyses, additional verification will not add to the validation of Step 4.

DISCUSSION

This review provides an overview of metamodeling methods that may be suitable for application in health economics, and addresses challenges regarding the selection and application of these methods. Similar to ordinary statistical regression modeling, different methods are available with their own advantages, disadvantages, and underlying assumptions, which are discussed and directions for selecting and implementing these methods are provided. Selected methods are structured in a comprehensive six-step process that can be followed to assure essential modeling steps are covered, as it includes all relevant design choices. Additionally, the process discussed can be used as a structure to effectively and efficiently communicate metamodeling studies, to increase modeling transparency and reproducibility.

Given that tools and packages are available to generate experiments according to specific designs and to fit different types of metamodels, for example in R, applying metamodeling methods is feasible for health economic analysts. Currently available software and results from this study enable analysts to perform computationally demanding analyses with their models, such as sensitivity analyses, including value of information analyses, model calibration, and optimization. Benefits of developing metamodels are relevant to analyses using patient-level simulation methods, such as microsimulation state-transition modeling and discrete event simulation, but also to cohort models used to perform analyses that require a large number of model evaluations, such as value of information analysis and optimization.

Applying metamodeling methods can reduce computational burden, but this usually comes at the price of introducing additional uncertainty in the model outputs. Consequently, checking whether underlying assumptions are met and checking metamodel performance, are crucial to success and essential to build confidence in the metamodel. Since modelers typically have access to the health economic model, validation is often not a problem, though likely to be more demanding in terms of effort compared to developing the metamodel itself. The starting point for building any metamodel, however, should be a realistic and validated health economic model, since metamodels can theoretically be as accurate as their corresponding simulators, but will not compensate for inaccuracies in these simulators. Moreover, when metamodels are used for optimization, the strategies considered, and possibly identified as optimal, may not be supported by (the data underlying) the simulator. Caution is required when such extrapolation is (automatically) performed, and such optimization results should only serve to initiate discussion on the appropriateness and validity of the simulator and the data supporting it. In addition, application of metamodeling methods requires communication of metamodeling design choices made in publications, for which space typically already is limited. Hence, metamodeling studies may be published separately from their simulator to assure the metamodeling process can be appropriately described. Furthermore, there is a 'sweet spot' for metamodeling: sufficient experiments need to be evaluated using the simulator to develop an accurate metamodel, but evaluating all experiments of interest should not be feasible.

Several technical challenges regarding the application of metamodeling methods in health economics remain. Simulators in health economics may include complex behavior, such as rigid cutoffs due to clinical decision rules, which may be complex for metamodeling techniques to capture. Additionally, (combinations of) model input parameters may be subject to constraints, which are difficult to incorporate in efficient designs of experiments, such as Latin Hypercube sampling. If sufficient samples can be evaluated from the simulator to use random or full factorial designs in which constraints can be accounted for more easily, however, this might not be an issue. Alternatively, more advanced adaptive sampling strategies may need to be applied.

Focusing on generally employed metamodeling methods is a limitation to this study. Consequently, not all metamodeling methods theoretically suitable for application in health economics are discussed, for example, techniques that can be used for categorical outcomes

(i.e., classification) are not presented. These techniques can be applied in health economics, however, after discretizing continuous outcomes, which has been done previously to use logistic regression in this context [58]. Additionally, alternative performance criteria for metamodel validation can be found, or may be developed, based on the study-specific needs. With regard to sampling methods, adaptive methods have not been discussed and additional designs of experiments are available, such as D-optimal designs, which are efficient and can account for constraints, but for which a linear or quadratic model simulator model structure should be known [64], or so called Sobol sequences, which may be more efficient compared to Latin Hypercube designs for low dimension (i.e., number of input parameters) problems [65, 66].

Future metamodeling applications should further illustrate the potential and use of these research methods, and identify common challenges. Once the field of metamodeling in health economics has evolved, good research practices (i.e., consensus guidance) can be identified to further improve the quality of metamodeling studies.

REFERENCES

1. Annemans L, et al. Current methodological issues in the economic assessment of personalized medicine. Value Health 2013;16(S6): S20–6.

2. Miller JD, et al. Current challenges in health economic modeling of cancer therapies: a research inquiry. Am Health Drug Benefits 2014;7(3):153–62.

3. Phillips KA, et al. Challenges to the translation of genomic information into clinical practice and health policy: utilization, preferences, and economic value. Curr Opin Mol Ther 2008;10(3):260–6.

4. Siebert U, et al. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. Med Decis Making. 2012;32(5):690-700.

5. Karnon J, et al. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. Med Decis Making. 2012;32(5):701-11.

6. Brennan A, et al. A taxonomy of model structures for economic evaluation of health technologies. Health Econ. 2006;15(12):1295-310.

7. Stahl JE. Modelling methods for pharmacoeconomics and health technology assessment. Pharmacoeconomics. 2008;26(2):131-48.

8. Heath A, et al. A review of methods for analysis of the expected value of information. Med Decis Making. 2017;37(7):747-58.

9. Steuten L, et al. A systematic and critical review of the evolving methods and applications of value of information in academia and practice. Pharmacorconomics. 2013;31(1):25-48.

10. Crown W, et al. Constrained optimization methods in health services research - an introduction: eport 1 of the ISPOR Optimization Methods Emerging Good Practices Task Force. Value Health. 2017;20(3):310-9.

11. Karnon J, Haji Ali Afzali H. When to use discrete event simulation (DES) for the economic evaluation of health technologies? A review and critique of the costs and benefits of DES. Pharmacoeconomics. 2014;32(6):547-58.

12. Law AM. Simulation modeling and analysis. Singapore: McGraw-Hill Higher Education, 2007.

13. Sacks J, et al. Design and analysis of computer experiments. Statist Sci. 1989;4(4):409-23.

14. Li H, et al. Multi-objective optimization of coronary stent using Kriging surrogate model. Biomed Eng Online. 2016;15(2):148.

15. Xu G, et al. Multi-objective aerodynamic optimization of the streamlined shape of high-speed trains based on the Kriging model. PloS one. 2017;12(1):e0170803.

16. Ouyang Q, et al. Conservative strategy-based ensemble surrogate model for optimal groundwater remediation design at DNAPLs-contaminated sites. J Contam Hydrol. 2017;203:1-8.

17. Butcher JB, et al. Estimating future temperature maxima in lakes across the United States using a surrogate modeling approach. PloS one. 2017;12(11):e0183499.

18. Degeling K, et al. A scoping review of metamodeling applications and opportunities for advanced health economic analyses. Expert Rev Pharmacoecon Outcomes Res. 2019;19(2):181-7.

19. Rojnik K, Naversnik K. Gaussian process metamodeling in Bayesian value of information analysis: a case of the complex health economic model for breast cancer screening. Value health. 2008;11(2):240-50.

20. Tappenden P, et al. Methods for expected value of information analysis in complex health economic models: developments on the health economics of interferon-beta and glatiramer acetate for multiple sclerosis. Health Technol Assess. 2004;8(27), 1-78.

21. Willem L, et al. Active learning to understand infectious disease models and improve policy making. PLoS Comput Biol. 2014;10(4):e1003563.

22. Yousefi M, et al. Chaotic genetic algorithm and Adaboost ensemble metamodeling approach for optimum resource planning in emergency departments. Artif Intell Med. 2018;84:23-33.

23. R Core Team. R: a language and environment for statistical computing [software]. Available from: https://www.r-project.org/

24. Strong M, et al. Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: a nonparametric regression approach. Med Decis Making. 2014;34(3):311-26.

25. Strong M, et al. Estimating the expected value of sample information using the probabilistic sensitivity analysis sample: a fast, nonparametric regression-based method. Med Decis Making. 2015;35(5):570-83.

26. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer, 2009.

27. Barton RR. Simulation metamodels. Proceedings of the 1998 Winter Simulation Conference. 1998:167-76.

28. Carson Y, Maria A. Simulation optimization: methods and applications. Proceedings of the 1997 Winter Simulation Conference. 1997:118-26.

29. Myers RH, et al. Response surface methodology: process and product optimization using designed experiments. Hoboken: Wiley, 2016.

30. Gaucel S, et al. Learning dynamical systems using standard symbolic regression. Berlin: Springer Berlin Heidelberg, 2014.

31. Koza JR. Genetic programming: on the programming of computers by means of natural selection. Cambridge: MIT Press, 1992.

32. Knafl GJ, Ding K. Adaptive regression for modeling nonlinear relationships. Basel: Springer International Publishing, 2016.

33. Li YF, et al. A systematic comparison of metamodeling techniques for simulation optimization in decision support systems. Applied Soft Computing. 2010;10(4):1257-73.

34. Hastie T, Tibshirani R. Generalized additive models. Statist Sci. 1986;1(3):297-310.

35. Wood S. Generalized additive models: an introduction with R. Boca Ralton: Taylor & Francis, 2006.

36. O'Hagan A, Kingman JFC. Curve fitting and optimal design for prediction. J R Stat Soc Ser B Methodol. 1978;40(1):1-42.

37. Swiler LP, et al. Surrogate models for mixed discrete-continuous variables. Cham: Springer International Publishing, 2014:181-202.

38. Gramacy RB, Lee HKH. Bayesian Treed Gaussian Process Models with an application to computer modeling. J Am Stat Assoc. 2008;103(483):1119-30.

39. Dreyfus G. Neural Networks: methodology and applications. Berlin: Springer Berlin Heidelberg, 2005.

40. Hurrion RD. Using a Neural Network to enhance the decision making quality of a visual interactive simulation model. J Oper Res Soc. 1992;43(4):333-41.

41. Alison C, et al. Learning surrogate models for simulation-based optimization. AIChE J. 2014;60(6):2211-27.

42. Wilson ZT, Sahinidis NV. The ALAMO approach to machine learning. Comput Chem Eng. 2017;106:785-95.

43. Simpson TW, et al. Metamodels for computer-based engineering design: survey and recommendations. Eng Comput. 2001;17(2):129-50.

44. Crombecq K, et al. Efficient space-filling and non-collapsing sequential design strategies for simulation-based modeling. Eur J Oper Res. 2011;214(3):683-96.

45. Garud SS, et al. Design of computer experiments: a review. Comput Chem Eng. 2017;106:71-95.

46. Briggs AH, et al. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group–6. Med Decis Making. 2012;32(5):722-32.

47. Box GEP. Statistics for experimenters: design, innovation, and discovery. Hoboken: Wiley, 2005.

48. Husslage BGM, et al. Space-filling Latin hypercube designs for computer experiments. Optim Eng. 2011;12(4):611-30.

49. McKay MD, et al. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics. 1979;21(2):239-45.

50. Johnson ME, et al. Minimax and maximin distance designs. J Stat Plan Inference. 1990;26(2):131-48.

51. Ye KQ. Orthogonal column Latin Hypercubes and their application in computer experiments. J Am Stat Assoc. 1998;93(444):1430-9.

52. Carnell R. lhs: Latin Hypercube samples. 2018 [Available from: https://CRAN.R-project.org/package=lhs] [Accessed 5 Dec 2018].

53. Petelet M, et al. Latin hypercube sampling with inequality constraints. AStA-Adv Stat Anal. 2010;94(4):325-39.

54. Levy S, Steinberg DM. Computer experiments: a review. AStA-Adv Stat Anal. 2010;94(4):311-24.

55. Loeppky JL, et al. Choosing the sample size of a computer experiment: a practical guide. Technometrics. 2009;51(4):366-76.

56. Jones DR, et al. Efficient global optimization of expensive black-box functions. J Global Optim. 1998;13(4):455-92.

57. Jalal H, et al. Linear regression metamodeling as a tool to summarize and present simulation model results. Med Decis Making. 2013;33(7):880-90.

58. Merz JF, et al. Measuring decision sensitivity: a combined Monte Carlo-logistic regression approach. Med Decis Making. 1992;12(3):189-96.

59. Briggs AH, et al. Decision modelling for health economic evaluation. Oxford: Oxford University Press, 2006.

60. Kahn H, Marshall AW. Methods of reducing sample size in Monte Carlo computations. J Oper Res Soc. 1953;1(5):263-78.

61. Murphy DR, et al. Using common random numbers in health care cost-effectiveness simulation modeling. Health Serv Res. 2013;48(4):1508-25.

62. Gano S, et al. Comparison of three surrogate modeling techniques: datascape, Kriging, and second order regression. 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference. American Institute of Aeronautics and Astronautics. 2006.

63. Kleijnen JPC, Sargent R. A methodology for fitting and validating metamodels in simulation. Eur J Oper Res. 2000;120(1):14-29.

64. de Aguiar PF, et al. D-optimal designs. Chemometr Intell Lab Syst. 1995;30(2):199-210.

65. Kucherenko S, et al. Exploring multi-dimensional spaces: a comparison of Latin Hypercube and quasi Monte Carlo sampling techniques. 2015 [Available from: https://arxiv.org/abs/1505.02350] [Accessed 20 June 2018].

66. Sobol' IM. Quasi-Monte Carlo methods. Progress in Nuclear Energy. 1990;24(1):55-61.

67. Lenth RV. Response-Surface methods in R, using rsm. J Stat Soft. 2009;32(7):1-17.

68. Flasch O, et al. rgp: R genetic programming framework. 2014 [Available from: https://cran.r-project.org/src/contrib/Archive/rgp/] [Accessed 5 Dec 2018].

69. Milborrow S. earth: multivariate adaptive regression splines. 2017 [Available from: https://CRAN.R-project.org/package=earth] [Accessed 5 Dec 2018].

70. Hastie T, Tibshirani R. mda: mixture and flexible discriminant analysis. 2017 [Available from: https://CRAN.R-project.org/package=mda] [Accessed 5 Dec 2018].

71. Hastie T. gam: generalized additive models. 2018 [Available from: https://CRAN.R-project.org/package=gam] [Accessed 5 Dec 2018].

72. Wood S. mgcv: mixed GAM computation vehicle with automatic smoothness estimation. 2018 [Available from: https://cran.r-project.org/web/packages/mgcv/mgcv.pdf] [Accessed 5 Dec 2018].

73. MacDonald B, et al. {GPfit}: an {R} package for fitting a Gaussian Process model to deterministic simulator outputs. J Stat Soft. 2015;64(12):1-23.

74. Gramacy RB. tgp: an R package for Bayesian nonstationary, semiparametric nonlinear regression and design by Treed Gaussian Process Models. J Stat Soft. 2007;19(9):1-46.

75. Gramacy RB, Taddy M. Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp Version 2, an R package for Treed Gaussian Process models. J Stat Soft. 2010;33(6):1-48.

76. Fritsch S, Guenther F. neuralnet: training of Neural Networks. 2016 [Available from: https://CRAN.R-project.org/package=neuralnet] [Accessed 5 Dec 2018].

77. Venables WN, Ripley BD. Modern applied statistics with S. New York: Springer; 2002.

CHAPTER 10

Maximizing the benefits of colorectal screening: from evaluation to optimization of diagnostic strategies using a metamodel

Hendrik Koffijberg
Koen Degeling
Maarten J. IJzerman
Veerle M.H. Coupé
Marjolein J.E. Greuter

ABSTRACT

Introduction: In silico decision-analytic studies are increasingly applied to assess the impact of screening and diagnostic strategies when trials are not feasible. Typically, such studies identify the best strategy by evaluating outcomes of only a limited set of pre-defined strategies to manage computational burden, none of which may be optimal. This paper illustrates how computational challenges with complex decision-analytic models can be negated using metamodeling, to allow optimization of diagnostic strategies in a case study on colorectal cancer screening, while accounting for colonoscopy capacity constraints.

Methods: An extensive set of screening strategies was defined, differing in starting age, screening interval, number of rounds, and fecal immunochemical test (FIT) cut off value. Two approaches were compared to determine the best screening strategy in terms of lifeyears gained compared with no screening, based on a limited subset of strategies evaluated with the validated ASCCA model. The traditional approach identified the best strategy within this limited set. The metamodeling approach used Gaussian Process metamodels fitted using this limited set to determine the best screening strategy by evaluating all defined strategies. Colonoscopy capacity was varied between 25-900 colonoscopies (lifelong per 1,000 individuals).

Results: Optimization frontiers from the metamodeling approach indicated that health outcomes could be improved without requiring additional colonoscopy capacity. For a colonoscopy capacity of 550, biennial screening between ages 33-73 with a 150 ng/ml FIT cut-off performed best, yielding 0.092 lifeyears and saving €361 per individual compared to no screening. Lower colonoscopy capacity decreased the screening interval to 1 year, whereas higher capacity decreased the FIT cut-off to 75 ng/ml.

Conclusion: Developing metamodels allowed for optimization rather than evaluation of screening strategies, resulting in the identification of better strategies in terms of health outcomes. The optimal strategies identified support ongoing discussions on decreasing the starting age of colorectal cancer screening.

INTRODUCTION

Health policy recommendations regarding the implementation of screening for a specific condition follow from a thorough assessment of its expected impact in terms of patient outcomes, healthcare outcomes, and costs. Such analyses ensure effective and efficient care to be delivered, and thereby support optimal use of healthcare resources. Although impact assessments have traditionally been performed based on randomized controlled trials (RCTs), such trials are not always feasible given their typically high cost, short duration (i.e., inability to observe long-term outcomes) and limited number of considered strategies (i.e., study arms).

In particular, when considering multiple diagnostic or screening strategies involving combinations of tests, procedures, and their timing, a large number of different strategies can be distinguished. Ideally, all relevant strategies would be compared in order to determine the best strategy. However, this is clearly not feasible in RCTs, as demonstrated by the rarity of even simple test+treatment RCTs [1]. Therefore, in silico decision-analytic impact studies are increasingly applied as a cheaper and more efficient method to evaluate strategies, and also to guide the design of RCTs [2,3]. Such modelling studies can easily incorporate multiple screening or diagnostic strategies for comparison. In doing so, in silico models should adequately represent clinical practice and disease progression. These models should also properly reflect the (indirect) impact of test outcomes on treatment decisions and resulting health outcomes and costs, for incorrect test results, for different test cut-off values, for patient stratification based on test outcomes, and other relevant aspects [4]. To reflect these aspects, patient-level simulation models (i.e., microsimulations) are typically required, which may be more computationally demanding compared to conventional cohort-level models [5].

Even with computationally demanding simulation models, performing traditional model analyses, including probabilistic analysis to reflect uncertainty in the modeling outcomes [6], is typically feasible within acceptable timeframes. Performing more advanced analyses like Value of Information (VOI) analysis, for example, or applying optimization approaches with such models, however, may not be feasible unless simulation code is parallelized and run on high performance computing clusters (see Appendix 1 for a numerical example). Alternatively, approximation methods can sometimes be applied to estimate outcomes of interest within a feasible timeframe, for example in the context of VOI analysis [7,8].

For screening models, which are often complex microsimulation models, it may prove computationally challenging to evaluate a large number of strategies to identify the best screening strategy for a specific condition. This explains why typically only a limited number of screening strategies are evaluated, for example between 2 and 25 strategies, and compared with a reference strategy [9,10]. Although limiting the number of strategies to be evaluated improves the feasibility of the analysis, a direct consequence is that the best screening strategy may not be included in the set of evaluated strategies and, hence, will *not* be identified. Given the typically large number of potential screening strategies, typically tens of thousands to millions of strategies (see Appendix 2 for a numerical example), the likelihood that the best screening strategy is actually identified may be very small.

Estimating the impact of all potential strategies or applying optimization approaches may become feasible, however, when approximating rather than evaluating the outcomes of a computationally demanding simulation model. Metamodels, also known as surrogate models or emulators, can provide such approximations of (complex) simulation models [11,12]. Metamodels have been used in different research contexts, but their application in health economics is still limited [13]. However, interest in metamodeling methods is increasing and recently guidance for their use has been developed [Chapter 9].

In this paper, metamodeling methods are used to address computational challenges associated with a complex simulation model, to enable optimization in a case study on colorectal cancer screening in the Netherlands, while accounting for colonoscopy capacity constraints.

METHODS

Colorectal cancer screening

Colorectal cancer (CRC) accounts for 1.8 million cases and over 850,000 deaths worldwide, making it a major public health issue [14]. CRC originates from colorectal polyps, i.e. adenomas and serrated lesions. These benign precursor lesions and the long preclinical phase make CRC an excellent target for screening. Indeed, several long-term RCTs have shown that CRC screening can considerably decrease CRC mortality [15,16]. As the preventive impact of CRC screening is widely recognized, CRC screening is now implemented in almost 60 countries [17]. These screening efforts vary widely in, amongst others, the type of screening program, i.e. organized population-based screening versus opportunistic screening, type of test used, the age ranges at which screening is offered and the screening interval. This diversity is mainly due to differences in available health care resources, with colonoscopy capacity being the most relevant constraint. In the Netherlands, a population-based CRC screening program was implemented in 2014 [18]. This program consists of biennial fecal immunochemical testing (FIT) in individuals aged 55-75 years. Individuals with a positive FIT result are referred to get a colonoscopy during which all detected colorectal polyps are removed. This program was defined based on a maximum capacity of 550 colonoscopies over the lifetime of 1,000 patients, and is expected to reduce CRC mortality with over 40% compared to no screening in 2040 [19].

The original ASCCA simulation model

The Adenoma and Serrated pathway to Colorectal CAncer (ASCCA) simulation model was designed and implemented to reflect the potential development of adenomas and serrated lesions into CRC, based on the characteristics of colorectal lesions. The natural history model is supplemented with a flexible screening component that can be set up to simulate a range of screening strategies. The model has previously been described in detail [20]. It has been used, amongst others, to evaluate the long-term impact of the Dutch national colorectal screening program in terms of cancer incidence, mortality and colonoscopy capacity requirement [19], and to assess the effectiveness and cost-effectiveness of colonoscopy surveillance in FIT-based screening [21]. Although the ASCCA model is programmed in C++, one of the fastest general programming environments [22], a deterministic analysis (i.e., single simula-

tion run) to evaluate one strategy takes about 40 minutes on a modern desktop computer. This substantial computation burden is due to the fact that all previous screening and surveillance results of an individual are used in the model to simulate future events, as well as due to the large number of individuals that needs to be simulated to obtain stable outcome estimates. Despite the model's valuable ability to evaluate different screening strategies, its computational cost makes it infeasible to identify the optimal screening strategy by evaluating a comprehensive set of thousands or more screening strategies within a reasonable timeframe (e.g., a few days).

Approaches to evaluate screening strategies

Potential FIT-based screening strategies were defined based on a limited set of four parameters: start age (in years, range 30-90), screening interval (in years, range 1-60), number of screening rounds (range 1-31), and FIT cut-off for referral (discrete values: 50, 75, 100, or 150 ng/ml). When limiting the maximum age of screening to 90 years, these combined ranges lead to a set comprising 40,864 unique screening strategies (see Appendix 2 for calculation).

We compared two approaches to identify the best screening strategy from all potential strategies, that is, the strategy with the highest lifeyears gained (LYG) or net monetary benefit (NMB) compared to no screening, requiring less colonoscopy capacity than a predefined maximum colonoscopy capacity. For the NMB a willingness-to-pay of €20,000 per lifeyear gained was used. Colonoscopy capacity was defined as the number of colonoscopies available in the lifetime of 1,000 patients. In the *traditional approach*, the optimal strategy was determined by evaluating a fixed set of strategies using the original ASCCA model and selecting the best strategy in terms of LYG or NMB that does not exceed the maximum colonoscopy capacity.

In the *metamodeling approach*, the exact same set of strategies evaluated in the traditional approach was used to fit Gaussian Process metamodels. Gaussian Process regression is a nonparametric regression method also known as Kriging, which uses information on neighbor observations for new predictions [11,23]. Based on the screening parameters defined earlier, metamodels were fitted to estimate: 1) the number of LYG (primary outcome), 2) incremental costs, and 3) NMB of a screening strategy compared to no screening, and 4) the number of colonoscopies required for a screening strategy. These metamodels were then used to evaluate all possible screening strategies to determine the overall best strategy in terms of predicted LYG or NMB that meets the colonoscopy capacity constraint. As the metamodeling approach can be considered an approximation to the ASCCA model, a feedback step (or iterative loop) to this original model was implemented to check metamodel performance.

Analysis and outcomes of the approaches

We performed three steps to illustrate the use and advantages of metamodels. The first step illustrates the potential benefit of optimization using a metamodel in terms of LYG, across a range of colonoscopy capacity constraints:

1.1 Generate sets of 25, 50, 100, and 150 screening strategies to cover the strategy space, by sampling according to a Latin hypercube design and only including strate-

gies that do not exceed the maximum age of screening of 90 years.

1.2 Evaluate all sets of screening strategies with the ASCCA model to obtain the outcomes in terms of LYG and number of colonoscopies required for each screening strategy.

===== This is where the *traditional approach* stops. =====

1.3 Fit metamodels to link the four screening parameters to the LYG and number of colonoscopies as derived from the ASCCA model for each set of screening strategies separately, resulting in four groups of two metamodels (one group for each set of 25, 50, 100 and 150 screening strategies).

1.4 For each group of metamodels separately, use the metamodels to estimate the LYG and number of colonoscopies required for all 40,864 potential screening strategies.

1.5 For each group of metamodels separately, present the best outcomes in terms of LYG from the traditional and metamodeling approaches in an optimization frontier for different colonoscopy capacities ranging from 25-900 (in steps of 25).

The second step investigates how accurate, and therefore meaningful, outcomes of the LYG metamodels are:

2.1 Based on the outcomes of the 40,864 unique screening strategies estimated in Step 1.4 using the group of metamodels fitted to 150 experiments in Step 1.3, identify 100 strategies with outcomes corresponding to the 100th quantiles of LYG.

2.2 Evaluate this set of 100 strategies with the ASCCA model.

2.3 For all groups of metamodels, compare the outcomes predicted by the LYG metamodel with those from the ASCCA model in calibration plots and report error values.

The third step concerns the identification of the CRC screening strategy expected to be optimal for the Netherlands. For this purpose, we used the group of metamodels fitted based on the largest set of 150 experiments.

3.1 Based on the outcomes of all (remaining) screening strategies estimated using the group of metamodels fitted to 150 experiments in Step 1.3, identify the 150 strategies with the expected highest LYG (for the first iteration, outcomes of all screening strategies from Step 1.4 are used).

3.2 Evaluate these top 150 strategies with the ASCCA model.

3.3 Compare the outcomes predicted by the metamodels with those from the ASCCA model to check for discrepancies. If necessary, fit new metamodels on the ASCCA outcomes of the top 150 strategies and use those metamodels to repeat Steps 3.1 and 3.2. This is an adaptive sampling step, to 'zoom in' on the region with strategies of interest. Hence, the strategies considered in Step 3.1 should be limited to those covered by the parameter ranges as observed in the top 150 strategies used to fit the metamodels.

3.4 When the metamodels are considered sufficiently accurate, they can be used to evaluate all remaining strategies and, thereby, identify the best strategy given a colonoscopy capacity constraint.

Step 3 was also performed with NMB rather than LYG as primary outcome.

RESULTS

Figure 1 shows the LYG optimization frontiers as function of the maximum colonoscopy capacity, for each group of metamodels fitted based on screening strategy sets of different sizes (Step 1). Regardless of the number of included strategies, it is clear that the majority of strategies sampled for the *traditional approach* according to a Latin hypercube design, require limited colonoscopy capacity and yield only small health benefits. Most strategies are located in the lower left corner, yielding between 0 and 0.03 LYG per individual and requiring less than 250 colonoscopies lifelong per 1,000 individuals.

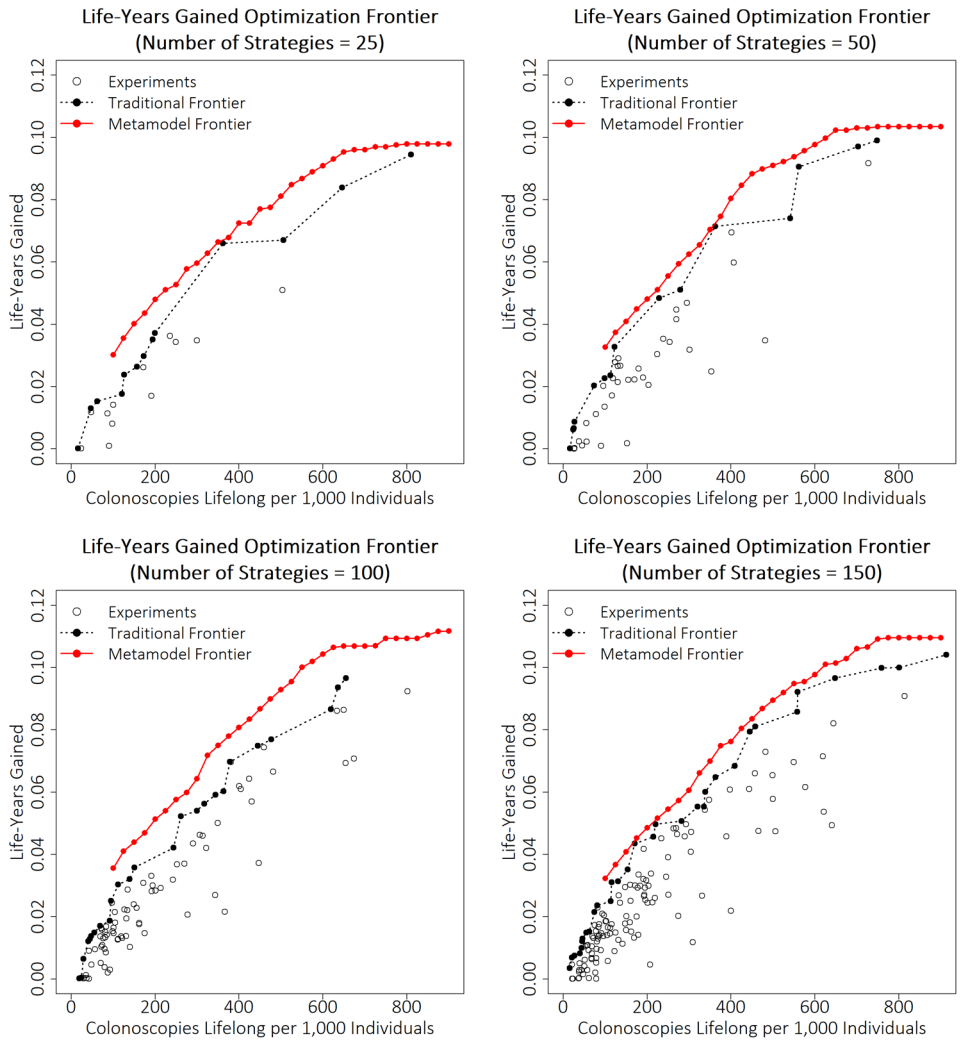


Figure 1. *Optimization frontier for the LYG as function of maximum colonoscopy capacity.*

The potential benefit of using the metamodeling approach is visible as the area between the metamodel frontier and the traditional frontier. The benefit is observed across the entire range of maximum colonoscopy capacity, but varies in size being highest in the region where few strategies have been evaluated (i.e., top right). As expected, the benefit of using metamodels is smallest for a set of 150 strategies. When the number of strategies evaluated with the ASCCA model and used to fit metamodels increases, the best strategy included in this set of strategies approaches to the metamodel frontier and, hence, the added value of the metamodel becomes smaller.

Looking only at the difference between the frontiers, however, would ignore the fact the metamodels are approximations of the ASCCA model and that predicted outcomes from the metamodels may not accurately reflect outcomes of the ASCCA model. Figure 2 shows the results of assessing the accuracy of the LYG metamodel (Step 2): calibration plots for the different number of screening strategies used to fit metamodels. The corresponding error values are presented in Table 1. Clearly, calibration is relatively good for low health benefits and relatively poor for high health benefits, due to the aforementioned distribution of strategies in the lower-left corners of the plots in Figure 1. While calibration improves with an increasing number of strategies considered, calibration for large health benefit values is still poor, even when considering 150 strategies.

To identify the optimal screening strategy (Step 3), the performance of the LYG metamodel developed based on 150 strategies (Step 1.3) with regard to the 150 best strategies (Step 1.4) was assessed. The mean absolute error of 0.0245 and mean relative absolute error of 30.7% were deemed too large to accurately identify the best strategy with the metamodel. Hence, an adaptive sampling step was performed: a new Gaussian Process metamodel was fitted on the ASCCA outcomes of these top 150 strategies. Figure 3a shows the calibration plot for the new metamodel for the 150 best strategies according the new LYG metamodel (Step 3.3, second iteration). Following this adaptive sampling step, performance of this new metamodel was considered sufficient (mean absolute error 0.0002; mean relative absolute error 0.2%).

The new metamodel was used to evaluate all scenarios within the parameter ranges as covered in the top 150 strategies identified in Step 3.1 (first iteration), as these were used to fit the metamodel. These parameter ranges after "zooming in" were start age (range 30-43), screening interval (range 1-3), number of screening rounds (range 18-21), and FIT cut-off values (50, 75, 100, and 150 ng/ml), resulting in 536 remaining feasible strategies. The optimization frontier of these strategies is presented in Figure 3b.

Table 2 shows that for a maximum of 550 colonoscopies lifelong per 1,000 individuals, which reflects current colonoscopy capacity, the optimal screening strategy in terms of LYG yields and expected 0.092 LYG per individual and starts at age 33 years, has a screening interval of 2 years, 21 screening rounds, and a FIT cut-off of 150 ng/ml. In comparison, the best screening strategy identified by the *traditional approach* yields 0.081 LYG and starts screening at age 32, has an interval of 3 years, 14 screening rounds, and also uses FIT cut-off of 150 ng/ml. The definition and outcomes for the top 10 strategies according to the metamodel are shown

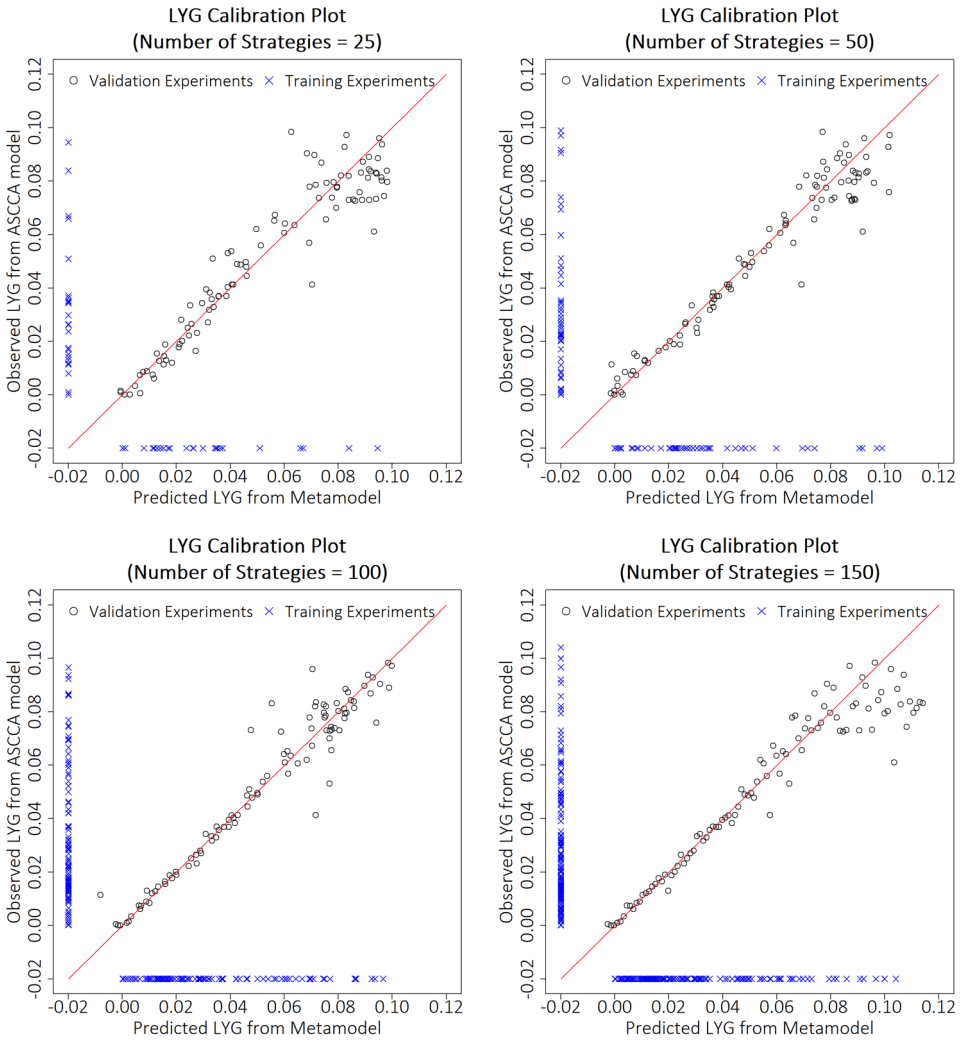


Figure 2. *Calibration plots for the LYG metamodels.*

Table 1. *Mean absolute and mean relative absolute errors for the LYG metamodels. MM, metamodel; ME, mean error; MAE, mean absolute error; MRE, mean relative error; MRAE, mean relative absolute error.*

	100 quantiles (Step 2.3)			
	ME	**MAE**	**MRE (%)**	**MRAE (%)**
MM25	0.0014	0.0072	84.8	101.2
MM50	0.0018	0.0055	48.3	78.9
MM100	-0.0002	0.0044	-53.9	64.5
MM150	0.0039	0.0068	-37.2	55.0

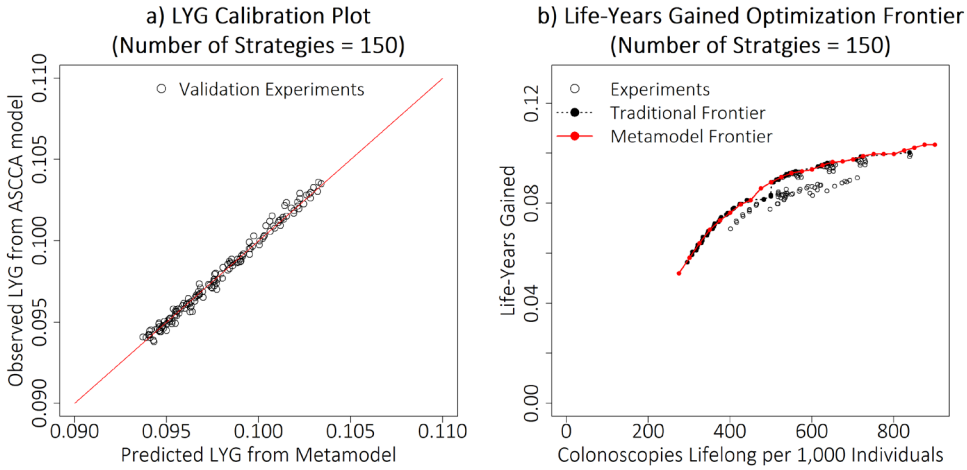


Figure 3. *Calibration plot (a) and optimization frontier (b) for the LYG metamodel after one adaptive sampling step. * Note: range of predicted LYG in this plot has been restricted to higher LYG values, as calibration was based on the top 150 strategies with highest expected LYG.*

Table 2. *Definitions and outcomes of screening strategies expected to be optimal in terms of LYG.*

Constraint	Screening strategy identified as optimal				Predicted strategy outcomes[1]			
Maximum colonoscopy capacity[2]	Start age screening (years)	Screening interval (years)	Number of screening rounds	FIT cut-off[3] for referral (ng/ml)	Number of colonosco-pies[2]	LYG	Incre-mental costs (€)	NMB[4]
300	32	1	18	150	296	0.058	-266	1430
450	40	1	20	150	441	0.081	-316	1941
550	33	2	21	150	546	0.092	-361	2202
650	34	2	21	100	647	0.097	-363	2294
800	35	2	21	75	738	0.100	-380	2374

[1]For the optimal screening strategy compared with no screening. No comparison between screening strategies was made, because their requirements in terms of colonoscopy capacity, which will determine feasibility in clinical practice, varies widely.
[2]Colonoscopy capacity in terms of life long colonoscopies, per 1,000 individuals.
[3]Only discrete cut-off values of 50, 75, 100, and 150 ng/ml were considered in the analysis.
[4]Net monetary benefit determined using a willingness-to-pay to €20,000 per lifeyear gained.

in Table 3. Clearly, both strategy definitions and outcomes are quite similar for these top 10 strategies. When colonoscopy capacity would be substantially lower, the screening interval decreases to 1 year (other aspects remain comparable). Conversely, when colonoscopy capacity would be substantially higher, the FIT cut-off decreases to 75 ng/ml (other aspects remains comparable).

Results found when optimizing the NMB rather than LYG did not provide additional insights, as the impact of costs is limited compared to the impact of LYG (see Appendix 3 - Table S1).

DISCUSSION

The computational burden of using complex, patient-level simulation models for health economic analyses can be problematic. In this case study, metamodeling techniques were used to negate computational challenges with the original ASCCA model, which allowed for identification of the optimal screening strategy for the Netherlands given a maximum colonoscopy capacity.

Table 3. *Top 10 screening strategies in terms of LYG for a colonoscopy capacity of 550.*

Screening strategy identified as optimal				Predicted strategy outcomes[1]			
Start age screening (years)	Screening interval (years)	Number of screening rounds	FIT cut-off[3] for referral (ng/ml)	Number of colonosco-pies[2]	LYG	Incre-mental costs (€)	NMB[4]
33	2	21	150	546	0.0920	-361	2202
34	2	20	150	545	0.0919	-348	2185
36	2	19	150	548	0.0917	-348	2183
35	2	19	150	541	0.0916	-361	2193
33	2	20	150	537	0.0914	-359	2186
32	2	21	150	538	0.0912	-343	2166
34	2	19	150	533	0.0911	-343	2166
37	2	18	150	543	0.0911	-361	2183
36	2	18	150	536	0.0910	-343	2163
38	2	18	150	549	0.0909	-347	2165

[1]For the optimal screening strategy compared with no screening. No comparison between screening strategies was made, because their requirements in terms of colonoscopy capacity, which will determine feasibility in clinical practice, varies widely.
[2]Colonoscopy capacity in terms of life long colonoscopies, per 1,000 individuals.
[3]Only discrete cut-off values of 50, 75, 100, and 150 ng/ml were considered in the analysis.
[4]Net monetary benefit determined using a willingness-to-pay to €20,000 per lifeyear gained.

Although the case study results are not intended to directly influence CRC screening guidelines or screening practice, they warrant further research and discussions on CRC screening in the Netherlands. For example, for the current Dutch colonoscopy capacity of 550 colonoscopies lifelong per 1,000 individuals, the metamodeling approach suggests optimal screening should start already at the age of 33 years and should be performed biennial until age 73, whereas the current Dutch screening program starts at age 55 years. Also the top 10 screening strategies identified suggest that earlier screening may be beneficial. As CRC incidence is rising in young adults [24], there is increasing interest in the optimal start age of screening. This is underpinned by two recent studies evaluating screening strategies starting at age 40 [25,26], which is still slightly later than the 33-38 years found for the top 10 strategies in the current study. This shows that it is unlikely that such early screening strategies would be part of a manually selected set of strategies to evaluate in a traditional approach. On the other hand, there might be reasons why certain strategies are not manually selected. In particular, strategies may be considered infeasible or ineffective in practice, for example, due to low compliance amongst young-adults due to a perceived low risk. Excluding such factors from the original simulation model, may lead to infeasible strategies being identified as optimal, but this is a general issue and not specific to the use of metamodeling.

Important to note is that the best-performing screening strategies identified in this study are extrapolations beyond the scope of the original ASCCA model. Colorectal polyp incidence rates in the ASCCA model were calibrated against polyp prevalence rates as reported by a Dutch study in which 1,420 screening naïve individuals aged 50-75 years underwent colonoscopy [27]. For the ages 20-50 and 75-90, prevalence rates reported by Rutter et al. [28] were used. As the confidence intervals in this latter study were extremely wide, there is considerable uncertainty regarding the ASCCA model predictions for screening at ages <50 years and >75 years. Consequently, this model was not yet used to consider screening strategies starting this early, and may itself need to be validated in this context.

Regarding the use of metamodeling in the case study, the feedback loop with the ASCCA model to ensure accuracy of the predicted outcomes was crucial, as the validation showed an adaptive sampling step was necessary to improve metamodel performance. Clearly, proper validation should be part of every metamodeling study, which is straightforward when access to the original model is available. Without metamodeling, only a potential set of dozens, or hundreds of strategies could have been evaluated with the ASCCA model. It is highly unlikely that the optimal screening would have been identified by such a traditional approach. The additional insight gained from evaluating a much more comprehensive set of strategies, within a single minute, is a clear advantage of metamodeling, and enables better-informed decision making. Beyond optimization based on population-level parameters, there is potential for metamodels to enable further research into more personalized screening in the Netherlands. Such screening strategies may be, for example, gender- and/or age-specific, or strategies in which the frequency of screening is adapted on the basis of previous test results.

Our study has certain limitations. Results shown are, of course, specific to the optimization problem addressed and the original simulation model used. When other CRC screening sim-

ulation models would have been used, results could have been different [29]. Furthermore, we applied a single metamodeling technique, Gaussian Processes to illustrate the potential advantages of metamodeling. Many alternative metamodeling techniques are available, such as for example, linear regression, artificial neural networks and multivariate adaptive regression splines, each with their own (dis)advantages [30,31]. However, rather than comparing these techniques, this study focused on increasing general awareness of the potential added value of metamodeling in a health economic context. For the same reason, validation efforts and performance checks were relatively limited, even though assessing metamodel accuracy is crucial and a range of resampling strategies is available [32]. Finally, the comparison of the traditional and metamodeling approach was based on use of the same 'limited set' of strategies generated using a Latin hypercube design. Using such a design to cover the entire space of strategies for the metamodeling approach is sensible. However, when only a very limited set of diagnostic strategies (e.g., in range of 2-25) is evaluated with the traditional approach, strategies are often not generated by design but defined by researchers. Such a manual selection could potentially either increase or decrease the benefits of metamodeling, depending on how close the strategies selected for evaluation would be to the optimal strategy.

CONCLUSIONS

Metamodels can be used to address computational challenges in complex, model-based health economic analyses and allow for optimization rather than evaluation of strategies. This was illustrated for a colorectal cancer screening context, while accounting for practical constraints regarding the maximum number of colonoscopies to be performed. When applying metamodeling methods, validation showed to be crucial, and an adaptive sampling step was necessary to improve metamodel accuracy. Outcomes from the metamodel indicated potential benefits of starting colorectal cancer screening at a younger age than is currently considered. This raised questions concerning the evidence-base (i.e., generalizability) of the original model, may lead to additional validation efforts, and provides insights beyond those that would have been obtained using only the original simulation model.

REFERENCES

1. Ferrante di Ruffano L, et al. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. J Clin Epidemiol. 2012; 65: 282-7.

2. Buxton MJ, et al. Modelling in economic evaluation: an unavoidable fact of life. Health Econ. 1997; 6: 217-27.

3. McKenna C, Claxton K. Addressing adoption and research design decisions simultaneously: the role of value of sample information analysis. Med Decis Making. 2011; 31: 853-65.

4. Kip MMA, MJ IJ, et al. Toward Alignment in the Reporting of Economic Evaluations of Diagnostic Tests and Biomarkers: The AGREEDT Checklist. Med Decis Making. 2018; 38: 778-88.

5. Degeling K, et al. A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models. Expert Rev Pharmacoecon Outcomes Res. 2017; 17: 17-25.

6. Briggs AH, et al. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. Value Health. 2012; 15: 835-42.

7. Strong M, et al. Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: a nonparametric regression approach. Med Decis Making. 2014; 34: 311-26.

8. Baio G. BCEA: a R package to perform Bayesian Cost-Effectiveness Analysis. Value Health. 2014; 17: A550.

9. Bor AS, et al. Optimal screening strategy for familial intracranial aneurysms: a cost-effectiveness analysis. Neurology. 2010; 74: 1671-9.

10. Greuter MJE, et al. The potential of imaging techniques as a screening tool for colorectal cancer: a cost-effectiveness analysis. Br J Radiol. 2016; 89: 20150910.

11. Barton RR. Simulation metamodels. Proceedings of the 30th conference on Winter simulation. Washington, D.C., USA: IEEE Computer Society Press, 1998.

12. Simpson TW, et al. Metamodels for computer-based engineering design: survey and recommendations. Eng Comput. 2001;17(2):129-50.

13. Degeling K, et al. A scoping review of metamodeling applications and opportunities for advanced health economic analyses. Expert Rev Pharmacoecon Outcomes Res. 2019; 19: 181-87.

14. International Agency for Research on Cancer. Global Cancer Observatory. 2018. Available from: https://gco.iarc.fr/. [Accessed April 2, 2019].

15. Shaukat A, et al. Long-term mortality after screening for colorectal cancer. N Engl J Med. 2013; 369: 1106-14.

16. Scholefield JH, et al. Nottingham trial of faecal occult blood testing for colorectal cancer: a 20-year follow-up. Gut. 2012; 61: 1036-40.

17. Young GP, et al. The global paradigm shift in screening for colorectal cancer. Gastroenterology. 2019; 156: 843-51 e2.

18. Rijksinstituut voor Volksgezondheid en Milieu. Bevolkingsonderzoek darmkanker (In Dutch). 2019. Available from: https://www.rivm.nl/bevolkingsonderzoek-darmkanker. [Accessed April 2, 2019].

19. Greuter MJE, et al. Long-term impact of the dutch colorectal cancer screening program on cancer incidence and mortality-model-based exploration of the serrated pathway. Cancer Epidemiol Biomarkers Prev. 2016; 25: 135-44.

20. Greuter MJE, et al. Modeling the Adenoma and Serrated pathway to Colorectal CAncer (ASCCA). Risk Anal. 2014; 34: 889-910.

21. Greuter MJE, et al. Screening for colorectal cancer with fecal immunochemical testing with and without postpolypectomy surveillance colonoscopy: a cost-effectiveness analysis. Ann Intern Med. 2017; 167: 544-54.

22. Aruoba SB, Fernandez-Villaverde J. A comparison of programming languages in macroeconomics. J Econ Dyn Control. 2015; 58: 265-73.

23. O'Hagan A, Kingman JFC. Curve fitting and optimal design for prediction. J R Stat Soc Ser B Methodol. 1978;40(1):1-42

24. O'Connell JB, et al. Rates of colon and rectal cancers are increasing in young adults. Am Surg. 2003; 69: 866-72.

25. Peterse EFP, et al. The impact of the rising colorectal cancer incidence in young adults on the optimal age to start screening: Microsimulation analysis I to inform the American Cancer Society colorectal cancer screening guideline. Cancer. 2018; 124: 2964-73.

26. Lew JB, et al. Benefits, harms, and cost-effectiveness of potential age extensions to the national bowel cancer screening program in Australia. Cancer Epidemiol Biomarkers Prev. 2018; 27: 1450-61.

27. Stoop EM, et al. Participation and yield of colonoscopy versus non-cathartic CT colonography in population-based screening for colorectal cancer: a randomised controlled trial. Lancet Oncol. 2012; 13: 55-64.

28. Rutter CM, et al. A hierarchical non-homogenous Poisson model for meta-analysis of adenoma counts. Stat Med. 2007; 26: 98-109.

29. Silva-Illanes N, Espinoza M. Critical analysis of markov models used for the economic evaluation of colorectal cancer screening: a systematic review. Value Health. 2018; 21: 858-73.

30. Tappenden P, et al. Methods for expected value of information analysis in complex health economic models: developments on the health economics of interferon-beta and glatiramer acetate for multiple sclerosis. Health Technol Assess. 2004; 8: iii, 1-78.

31. Li YF, et al. A systematic comparison of metamodeling techniques for simulation optimization in Decision Support Systems. Appl Soft Comput. 2010; 10: 1257-73.

32. Bischl B, et al. Resampling methods for meta-model validation with recommendations for evolutionary computation. Evol Comput. 2012; 20: 249-75.

APPENDIX 1: numerical example of computational challenges

With modern computing power, analysis of simulation models may often be performed in limited timeframes. However, performing complex analyses with a patient level (i.e., micro-simulation) model may still lead to computational challenges. For example, suppose a patient-level discrete event simulation model has been developed in R, to estimate the cost-effectiveness of a screening strategy compared to no screening, based on 50 unique input parameters, using a lifetime time horizon. Now assume that running this model, for 1,000 probabilistic analysis samples (using Monte Carlo simulation) and a cohort of 5,000 individuals takes around 1 minute. Accurately answering the question "which of these 50 parameters, if any, are most relevant to collect additional evidence on?", that is, performing an expected value of perfect parameter information (EVPPI) calculation in a brute force analysis could then require:

- Simulating ≥20,000 patients to obtain stable cohort results
- Simulating ≥2,500 Monte Carlo samples, for a proper probabilistic analysis
- Running ≥2,500 outer loop simulations, for an accurate EVPPI estimate
- Repeating this EVPPI analysis for each of the 50 parameters

That would require (20,000/5,000) * (2,500/1,000) * 2,500 * 50 / (60 * 24) ~ 868 days. Clearly, such computation times are infeasible, and approximations to either the EVPPI analysis, or simulation model evaluation in general, are required (and exist) to make such an analysis feasible in acceptable time.

APPENDIX 2: example of an exhaustive set of screening strategies

The most basic set of parameters to define a screening strategy would typically consist of the start age of screening, the screening frequency, the number of screening rounds, and some risk or test cut-off value used to refer screened individuals for further testing, for example, in a hospital. For the case study, the following parameters were defined with corresponding ranges.

Parameter	Unit	Range/Values	# Possible values
Start age	Years	30-90	61
Screening interval	Years	1-60	61
Number of screening rounds	-	1-31	31
Cut-off of the FIT for referral	Ng/ml	50,75,100,150	4

In total, these parameter values define 61*61*31*4 = 461,404 unique screening strategies. However, as parameter ranges were chosen to be wide on purpose, for example with a screening interval ranging from annually, to once every 60 years, this set contains many strategies which are not (clinically) relevant or feasible in practice. Hence, a practical constraint was introduced, which set the maximum age of the last screening moment to the age of 90 years. Removing all screening strategies containing a screening moment at age > 90 years resulted in a set comprising 40,864 strategies. The goal of the case study analysis was identifying the best performing screening strategy from this set of almost 41 thousand strategies. When additional screening parameters would be considered, for example, different FIT cut-off values or screening intervals depending on age, the number of unique strategies would run into the millions and the number of feasible strategies would increase very rapidly as well. Consequently, simply evaluating all feasible strategies with a complex (patient-level) simulation model is likely to pose substantial computational challenges.

APPENDIX 3: results of the optimization with NMB as outcome

Table S1. *Definitions and outcomes of screening strategies expected to be optimal in terms of NMB.*

Constraint	Screening strategy identified as optimal				Predicted strategy outcomes[1]			
Maximum colonoscopy capacity[2]	Start age screening (years)	Screening interval (years)	Number of screening rounds	FIT cut-off[3] for referral (ng/ml)	Number of colonosco-pies[2]	LYG	Incre-mental costs (€)	NMB[4]
300	32	1	18	150	298	0.058	-263	1430
450	40	1	20	150	441	0.081	-316	1943
550	33	2	21	150	546	0.092	-361	2202
650	33	2	22	100	649	0.097	-379	2313
800	35	2	19	50	797	0.102	-381	2413

[1]For the optimal screening strategy compared with no screening. No comparison between screening strategies was made, because their requirements in terms of colonoscopy capacity, which will determine feasibility in clinical practice, varies widely.
[2]Colonoscopy capacity in terms of life long colonoscopies, per 1,000 individuals.
[3]Only discrete cut-off values of 50, 75, 100, and 150 ng/ml were considered in the analysis.
[4]Net monetary benefit determined using a willingness-to-pay to €20,000 per lifeyear gained.

10

PART 4

Simulation modeling to evaluate and optimize targeted treatment strategies

CHAPTER 11

Simulating progression-free and overall survival for first-line doublet chemotherapy with or without bevacizumab in metastatic colorectal cancer patients based on real world registry data

Koen Degeling
Hui-Li Wong
Hendrik Koffijberg
Azim Jalali
Jeremy D. Shapiro
Suzanne Kosmider
Rachel Wong
Belinda Lee
Matthew E. Burge
Jeanne Tie
Desmond Yip
Louise M. Nott
Adnan Khattak
Stephanie H-S. Lim
Susan Caird
Peter Gibbs
Maarten J. IJzerman

ABSTRACT

Background: Simulations utilizing real world data have potential to optimize treatment sequencing strategies for specific patient subpopulations, including when conducting clinical trials is not feasible. This study simulated progression-free survival (PFS) and overall survival (OS) for first-line doublet chemotherapy with or without bevacizumab in metastatic colorectal cancer (mCRC) based on registry data.

Methods: Data from 867 patients were used to develop two survival models and a logistic regression model that populated a discrete event simulation (DES). Models' discrimination and calibration were used for internal validation, and predicted and observed medians and Kaplan-Meier plots were compared for DES. Bootstrapping and probabilistic sensitivity analysis were performed to correct for optimism and reflect uncertainty.

Results: Models showed reasonable discrimination and good calibration. A C-statistic of 0.66, regression slope of 0.82, and Brier-score of 0.09 were observed for the logistic regression model. For the survival models, C-statistics were 0.65, 0.62 and 0.62 (PFS), and 0.70, 0.68 and 0.67 (OS), at 0.5, 1, and 2 years respectively, and regression slopes were 0.996 and 0.998. Modified Hosmer-Lemeshow statistics showed good calibration, except for short-term predictions. Simulated medians and Kaplan-Meier plots matched those observed well. Exploratory analyses estimated that the median first-line PFS (95% confidence interval) of 219 (25%) patients could be improved from 175 days (156, 199) to 269 days (246, 294) by a personalized treatment targeting strategy based on expected PFS.

Conclusions: Clinical outcomes in mCRC may be improved based on simulations estimating effectiveness of different treatment combination and sequencing strategies using real world data.

INTRODUCTION

Clinical outcomes for metastatic colorectal cancer (mCRC) patients have improved substantially over the last two decades, partly due to increased efficacy of systemic treatment [1]. Initially, the availability of irinotecan and oxaliplatin made a major contribution to these improved outcomes [2-6]. Subsequently, the addition of bevacizumab, a vascular endothelial growth factor monoclonal antibody [7-11] was shown to further improve patient survival outcomes. This increase in treatment options, however, also created challenges regarding decisions about which (combination) of these drugs should be used for first-line and subsequent lines of treatment.

Numerous clinical trials have been performed to compare survival following first-line and second-line treatment in mCRC. For example, first-line irinotecan-based doublet chemotherapy (e.g., in combination with fluorouracil and leucovorin (FOLFIRI)) compared to oxaliplatin-based doublet chemotherapy (e.g., in combination with fluorouracil and leucovorin (FOLFOX)) has been studied [12], as well as first- and second-line sequencing of FOLFOX and FOLFIRI [13], and triplet chemotherapy (FOLFOXIRI) [14]. Furthermore, clinical trials have investigated adding bevacizumab to irinotecan- or oxaliplatin-based doublet chemotherapy [15, 16] or triplet chemotherapy [17, 18], as well as maintenance strategies [19]. However, although preferable, clinical trials cannot be used to evaluate all relevant treatment combinations and their optimal sequencing over multiple lines of treatment, because of the required amount of time and resources and the rate at which new therapies become available.

Further challenges remain regarding treatment sequencing in mCRC based on patient and disease characteristics and previous treatment outcomes [20, 21]. To address these, analyses of clinical trials have stratified efficacy estimates from broad patient populations into clinically relevant subgroups based on, for example, age [22], tumor biomarkers [23, 24], and primary tumor site [25]. Several of these factors, such as RAS mutation status and primary tumor site, now guide the use of epidermal growth factor receptor inhibitors, but there are as yet no established markers that inform treatment decisions for chemotherapy and bevacizumab. Furthermore, clinical trials are limited in the number of patient and disease characteristic combinations that can be investigated, and combinations of a limited number of characteristics only partly explain the variation in clinical outcomes. Additionally considering the success or failure of previous interventions in clinical trials for second and later lines of therapy adds to the required number of studies to be performed. Hence, it may not be possible to address challenges regarding treatment sequencing in mCRC by performing clinical trials only.

Observational studies based on real world data (RWD) have been recognized as a valuable source of evidence for comparative effectiveness research in oncology, when evidence from clinical trials is not available or when results from clinical trials only apply to a selected group of patients [26, 27]. In mCRC, for example, RWD have been used to describe observed treatment sequences, patient characteristics, and clinical outcomes [28] and to study the comparative effectiveness of FOLFOX and FOLFIRI sequencing, i.e. which chemotherapy doublet should be used in first-line and which in second-line treatment [29], and are being used to evaluate the prognostic role of inflammatory markers in patients treated with bevacizumab

[30]. Although observational studies are typically subject to selection bias, careful analysis of RWD has potential to provide valuable information on comparative effectiveness of different treatment sequences.

Regardless of whether clinical outcomes of different treatment sequences for a patient population are studied through clinical trials or observational studies, it is challenging to translate results from these studies into an optimal sequencing strategy that informs treatment decisions across multiple lines of treatment. Simulation models developed based on individual patient data of patient and disease characteristics and treatment history have the potential to estimate the effectiveness, and impact, of different treatment combinations and sequencing strategies for future specific subpopulations of patients. By applying such simulation models in combination with optimization techniques, optimal treatment sequencing strategies for specific patient subpopulations may be identified.

This study aimed to explore utilization of RWD using simulation techniques to optimize treatment sequencing in mCRC. A simulation model combining patient and disease characteristics, including tumor biomarkers, and conditional treatment outcomes to estimate effectiveness of treatment sequencing strategies was developed based on data from the Treatment of Recurrent and Advanced Colorectal Cancer registry (TRACC) [31]. Parametric survival models were used to populate a discrete event simulation (DES) that simulated progression-free survival (PFS) and overall survival (OS) for first-line mCRC treatment with doublet chemotherapy with or without bevacizumab.

METHODS

Figure 1 illustrates the structure of the DES (i.e., time-to-event simulation). Time-to-events in the DES were simulated using two parametric survival models and one logistic regression: a survival model representing PFS for first-line treatment (Model L1-Time), a logistic regression to distinguish between the competing events of death during first-line treatment versus physician-defined disease progression (Model L1-Event), and a survival model representing OS for subsequent treatment lines or follow up after first-line progression (Model LX-Time). More details on how these models and the DES were developed and validated is provided below.

Treatment of Recurrent and Advanced Colorectal Cancer Registry (TRACC)

TRACC was established to collect RWD on treatment and outcomes on consecutive patients diagnosed with mCRC at multiple public and private hospitals across and outside Australia [31]. To date, data on almost 3000 patients have been recorded in TRACC, including detailed patient and disease characteristics, treatment history of primary disease, and metastatic treatment motivation and outcomes across multiple treatment lines. TRACC data has been used previously to assess, for example, trends in biological agent use [32], primary tumor site impact on PFS and OS [33], and PFS, OS, and several safety endpoints for intact compared to resected primary tumors [34].

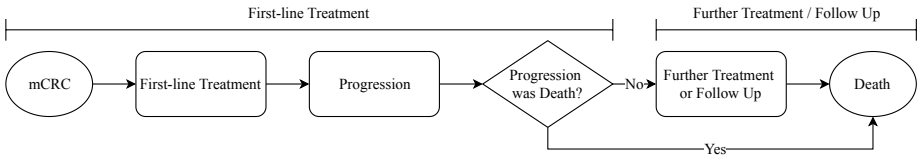


Figure 1. *Graphical representation of the discrete event simulation model structure.*

Data Cleaning

For every variable, unique values and distributions were assessed to check for data entry errors and to combine values with a similar meaning, which may occur following registry updates. As an example of a data entry error, some patients had progression dates prior to treatment start dates, resulting in a negative PFS, and were excluded. Additionally, two important data cleaning steps were performed: 1) categorization of treatments and 2) calculation of the time-to-events. Since patients typically receive different chemotherapy drugs in combination with a biological agent as part of first-line treatment, these different treatment combinations needed to be categorized to allow patient selection (see next section). Following a previously used categorization strategy, chemotherapy regimens were categorized according to their most intense therapy, i.e. singlet, doublet, or triplet chemotherapy, and multiple biological agents according to the first exposed biological [32]. Regarding the time-to-events, PFS was calculated from treatment start date to date of death or physician-defined progression. OS after first-line progression was calculated from the date of progression on first-line therapy to death. For patients without a date of progression or date of death, the treatment stop date or, if not available, date of last review was used as censoring date.

Patient Inclusion

Patients treated in an Australian hospital, indicated for palliative-intent treatment, with a treatment start date after 1 January 2009 and before 1 June 2017, and with recoded first-line treatment of doublet chemotherapy with or without bevacizumab, were included for analysis. Patients who were not exposed to bevacizumab but to other biological agents were excluded. Physician-defined treatment intent is captured in TRACC and was used to determine whether treatment was administered with curative (e.g., resectable liver metastases), potentially curative, or palliative goals. The timeframe was chosen because bevacizumab was listed on the Australian Pharmaceutical Benefits Scheme in 2009 and to allow for a reasonable follow up period to capture events.

Variables Considered for Inclusion in the Models

Candidate variables, transformations and interactions for both parametric survival models were identified considering their clinical relevance and number of observations. The following variables were identified: age, gender, hospital type (private or public), stage at diagnosis, primary tumor site, presence of liver metastases, presence of peritoneal metastases, number of metastatic sites, Eastern Cooperative Oncology Group (ECOG) performance status, Charlson comorbidity index (not corrected for age), RAS mutation status, and BRAF mutation status. An interaction between the number of metastatic sites (1 or 2 vs. more than 2) and

carcinoembryonic antigen (CEA) level (lower than 200 vs. 200 or more) was identified. A treatment indicator (chemotherapy and bevacizumab vs. chemotherapy only) was included for both survival models. First-line PFS was included as a candidate variable for modeling OS after first-line progression (Model LX-Time) to make this survival model conditional on first-line treatment outcomes. An overview of all considered transformations is provided in Appendix 1. For both survival models separately, each identified candidate variable and interaction was scored on its clinical relevance (i.e., low, moderate or high) to guide variable selection in the multivariable analysis (see Section Multivariable Modeling of Survival and Appendix 2).

Missing Data

Data were missing for the following variables (% missing): stage at initial diagnosis of colorectal cancer (1%), primary tumor site (2%), CEA level (9%), RAS mutation status (20%), and BRAF mutation status (57%). The substantial number of observations missing for RAS and BRAF mutation status were due to this molecular testing not being routinely assessed throughout the selected timeframe. Missing data for RAS and BRAF mutation status were partly addressed by assuming that RAS and BRAF mutations are mutually exclusive [35]. Remaining missing data for RAS and BRAF mutation status and for other variables were imputed 25 times using multiple imputation with chained equations in 20 iterations [36, 37]. CEA level was imputed using Bayesian linear regression, RAS and BRAF mutation status was imputed using logistic regression, and stage at diagnosis and primary tumor site were imputed using multinomial logit models. All variables were used as predictors. Post-processing was performed to correct a positive BRAF mutation status to negative in the presence of a RAS mutation.

Univariable Modeling

Univariable modeling was performed to determine which distribution types would be used to parameterize the survival models and which transformations would be used for each variable in the multivariable modeling (see Section Multivariable Modeling of Survival). First, a distribution type was selected for each survival model based on the likelihood and assessment of quantile-quantile plots and Kaplan-Meier plots. After considering exponential, Weibull, logistic and log-normal distributions, a Weibull distribution was selected to parameterize each survival model (see Appendix 3 for details). Next, best-performing transformations were selected for each survival model separately by fitting individual models for each candidate transformation and comparing performance based on Wald test p-values. All p-values were calculated for each of 25 imputed datasets, and for variables with multiple candidate transformations, transformations' p-value ranks were determined for each imputed dataset. Final selection of the transformation to be used was based on the average transformation rank over all 25 imputed datasets.

Multivariable Modeling of Survival

Two multivariable survival models Model L1-Time and Model LX-Time were developed to represent PFS for first-line treatment and OS following progression after first-line treatment accounting for subsequent treatment lines. To take both clinical relevance and statistical performance of variables into account, a comprehensive variable selection procedure based on

variables' p-values was implemented for both survival models. Variables considered highly relevant were included regardless of their statistical performance, whereas variables with a moderate or low clinical relevance were included based on a backward or forward selection procedure based on their p-value, respectively. Treatment interactions were subsequently considered for all included variables and included based on a backward selection procedure (see Appendix 4 for details). Clinical relevance was not considered for logistic regression Model L1-Event, as it is difficult to define variables' relevance in distinguishing between multiple competing events, and all candidate variables and their treatment interactions were included based on a forward selection procedure. To further value clinical relevance, relatively high p-value thresholds were chosen: 0.3 for all backward selections and 0.2 for all forward selections.

According to best practice for selection procedures using imputed data, an exhaustive approach was taken to calculate p-values for variables [38]. In each step of a selection procedure, a (survival or logistic regression) model was fitted to all imputed datasets and Rubin's rule [39] was applied to combine these 25 models into one model, including its total variance. Next, a Wald test p-value for each variable was calculated based on this pooled model, combining all coefficients that related to a specific variable. For example, variable Age was modeled using a second-order polynomial and, hence, inclusion or exclusion of Age as a variable was based on the two Age coefficients combined.

Discrete Event Simulation (DES)

A DES was populated using Model L1-Time, Model L1-Event, and Model LX-Time to simulate first-line PFS and OS on an individual-patient level, reflecting patient heterogeneity and stochastic uncertainty (i.e., first-order uncertainty) (Figure 1). To handle progression and death as competing events for first-line PFS, a unimodal joint-distribution and regression model approach was implemented using Model L1-Time and Model L1-Event [40]. Although general guidance on implementing competing risks in DESs suggests using multimodal distributions, this guidance is for uncensored data [40]. Because our data had 6.5% censoring (Appendix 5) and the performance of different approaches towards handling competing risks is expected to be different for censored data compared to uncensored data, and time-to-event distributions overlapped substantially for our data (73%, Appendix 5), it was decided to use a unimodal distribution. Simulations were performed for (subsets of) the TRACC patient population by randomly sampling 20,000 patients with replacement from the corresponding observed patient group.

Probabilistic Sensitivity Analysis

Probabilistic sensitivity analysis (PSA) was performed to reflect parameter uncertainty (i.e., second-order uncertainty) and obtain insights into the uncertainty surrounding simulation outcomes. According to best practice for DES [41], correlated sets of model parameter coefficients were obtained by applying Rubin's rule [39] over all imputed datasets in each of 10,000 bootstrap iterations performed to assess internal validity (see Section Validation and Figure 2). The resulting sets of Model L1-Time, Model L1-Event and Model LX-Time coefficients were used to perform 10,000 PSA runs.

Validation

Several statistical measures were used to assess models' internal validity in terms of model discrimination, i.e. ability to distinguish between high and low risk individuals, and calibration, i.e. agreement between observed and predicted outcomes [42]. Regarding survival model discrimination, Uno et al.'s [43] modification of Harrel's C-statistic [44] was used to account for censoring. Two measures were used to assess survival model calibration: 1) the regression slope of the predicted Weibull scale parameter in a Cox-regression model [45], and 2) Demler et al.'s Greenwood-Nam-D'Agostino statistic (GND) [46], which is a modification of Hosmer-Lemeshow's statistic by Nam and D'Agostino [47] and accounts for censoring using Greenwood's variance estimator [48]. Both C-statistic and GND are assessed at specific timepoints, which were selected to be at 0.5, 1, and 2 years. Additionally, GND requires a number of groups to be specified in which patients are stratified according to their predicted risk, for which 5 and 10 were chosen to assess impact of this arbitrary decision. Internal validity of logistic regression Model L1-Event was assessed by excluding patients for which no event was observed and using Harrel's C-statistic for model discrimination [44]. To assess calibration, the regression slope of the predicted response of Model L1-Event in a logistic regression model [49] and the Brier score [50] were used. To validate the DES, simulated median time-to-events and Kaplan-Meier plots (including confidence intervals based on 10,000 PSA runs, see Section Probabilistic Sensitivity Analysis) were compared to those observed for several subgroups to assure appropriate implementation of Model L1-Time, Model L1-Event and Model LX-Time.

Following best practices [51], nearly unbiased estimates for all statistics were obtained by performing 10,000 bootstrap iterations (Figure 2). In this process, the apparent performance of a model D_{app}, which is obtained by validating this model on all 25 imputed original datasets, is corrected for model optimism O. This optimism O is calculated as the difference between the internal validity of a bootstrap model based on the bootstrap dataset D_{boot} and the original dataset D_{origin} [44, 51].

Reproducibility and Availability of Materials

All analyses and simulations were performed in R version 3.5.2 [52]. An online tool was made publicly available in which dummy TRACC data can be visualized and simulations using the DES can be performed. Additionally, R code written to perform all analyses and simulations, including detailed intermediate outcomes, was made available to enhance transparency and reproducibility. All these materials are available from: http://www.personex.nl/research/mCRC-TRACC.

11

RESULTS

Of 2689 patients treated in an Australian hospital, 2041 received first-line treatment with chemotherapy and 1331 of those treatments had a palliative intent. Treatments were recoded as doublet chemotherapy for 988 of these patients, of which 958 were additionally treated with bevacizumab or without a biological agent. Of these patients, 871 patients were treated within the considered timeframe and 4 were excluded due to negative PFS, leaving

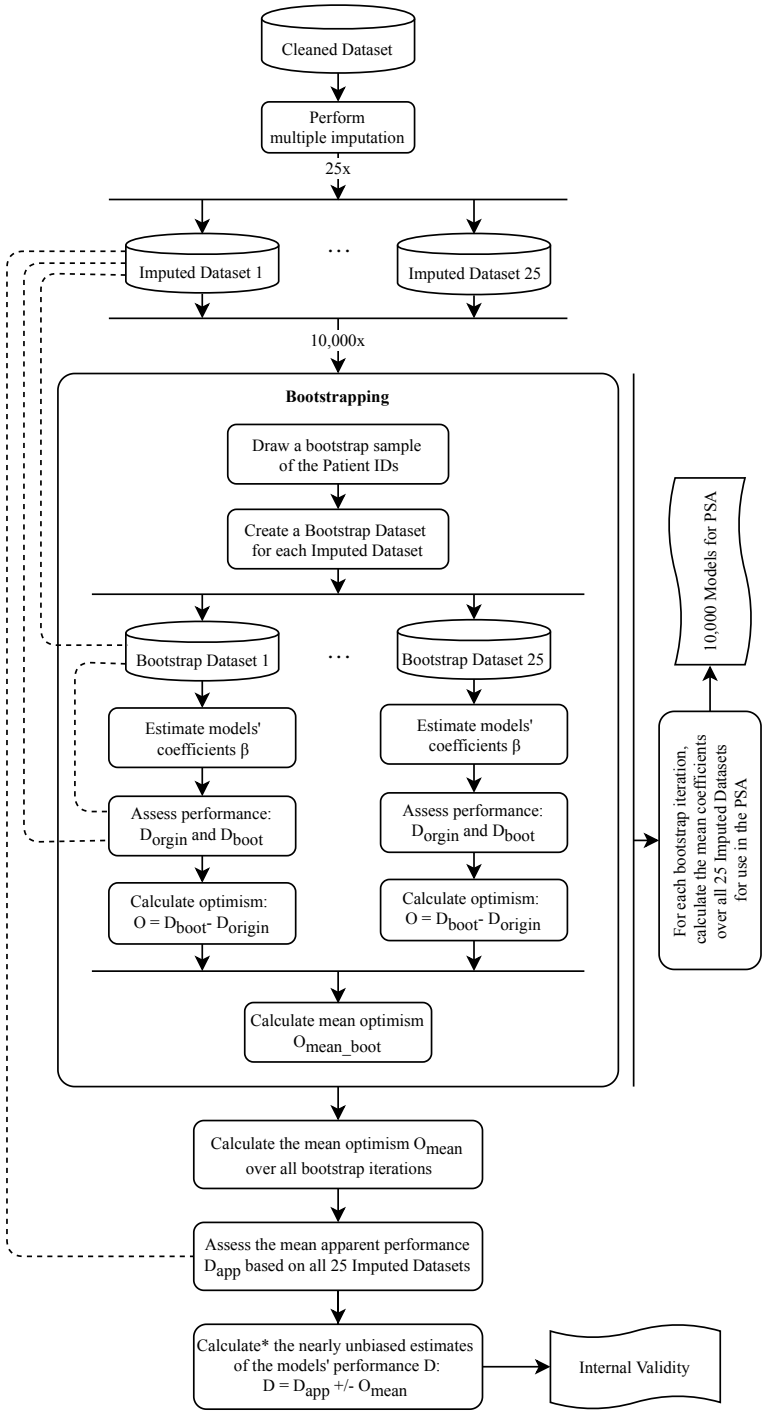


Figure 2. *Overview of the complete imputation and validation process. PSA, Probabilistic Sensitivity Analysis. * Whether* O_{mean} *is subtracted from or added to* D_{app} *depends on the performance measure.*

data of 867 patients for analysis. Table 1 provides an overview of patient characteristics for patients treated with or without bevacizumab in addition to doublet chemotherapy (abbreviated as ChemBev and ChemOnly, respectively). There were substantially more observations for the ChemBev group (n = 676) compared the ChemOnly group (n = 191). Significant differences in baseline characteristics were observed for the hospital type, ECOG performance status, Charlson index, RAS status, and BRAF status, though for these latter two differences are hard to interpret because of large amount of missing data. RAS status was missing for 175 (20%) patients and BRAF status was missing for 496 (57%) patients. These numbers were reduced to 174 (20%) and 354 (41%), respectively, by assuming RAS and BRAF mutations to be mutually exclusive (see Section Missing Data). Post-processing of BRAF mutation status following multiple imputation was only required in 0.35% of observations.

Table 1. *Patient characteristics for the doublet chemotherapy only (ChemOnly) and doublet chemotherapy with bevacizumab (ChemBev) treatment strategies (significance levels: * < 0.10; ** < 0.05; *** < 0.01). Values represent the count (%) for categorical variables and mean (SD) for continuous variables. NOS, not otherwise specified.*

		ChemOnly	ChemBev	p-value	
Number of Patients		191	676		
Age		63 (14)	63 (12)	0.571	
Gender	Male	118 (62)	407 (60)	0.739	
	Female	73 (38)	269 (40)		
Hospital Type	Public	115 (60)	337 (50)	0.014	**
	Private	76 (40)	339 (50)		
Stage at Diagnosis	Locally Advanced	8 (4)	35 (5)	0.294	
	Stage I	0 (0)	11 (2)		
	Stage II	10 (5)	35 (5)		
	Stage III	21 (11)	95 (14)		
	Stage IV	151 (79)	491 (73)		
	Missing	1 (1)	9 (1)		
Primary Tumor Site	Left colon	73 (38)	261 (39)	0.361	
	Right colon	51 (27)	213 (32)		
	Colon NOS	13 (7)	26 (4)		
	Rectum	51 (27)	165 (24)		
	Missing	3 (2)	11 (2)		
Number of Metastatic Sites	1	91 (48)	256 (38)	0.168	
	2	55 (29)	254 (38)		
	3	33 (17)	123 (18)		
	4	9 (5)	36 (5)		
	5	3 (2)	6 (1)		
	6	0 (0)	1 (0)		

Model coefficients and 95%-confidence intervals are presented in Appendix 2. For survival Model L1-Time, bias-corrected C-statistics were 0.65, 0.62, and 0.62 at 0.5, 1, and 2 years respectively, indicating a reasonable model discrimination. Regarding model calibration, a bias-corrected slope for the predicted Weibull Scale of 0.996 indicated good calibration. Bias-corrected GNDs (p-values) of 8.7 (0.07), 5.6 (0.23), and 4.0 (0.41) for 5 risk groups, and 18.6 (0.03), 17.6 (0.04), and 12.7 (0.17) for 10 risk groups, at 0.5, 1, and 2 years respectively, indicated a good model calibration, except for short-term predictions (see Discussion). A graphical representation of uncorrected GNDs at 1 year is provided in Figure 3, which shows predicted values compared to observed values for one randomly selected imputed dataset. For logistic regression Model L1-Event, a bias-corrected C-statistic of 0.66 showed reasonable model discrimination. A bias-corrected regression slope of 0.82 and Brier score of 0.09

Table 1 (continued). *Patient characteristics for the doublet chemotherapy only (ChemOnly) and doublet chemotherapy with bevacizumab (ChemBev) treatment strategies (significance levels: * < 0.10; ** < 0.05; *** < 0.01). Values represent the count (%) for categorical variables and mean (SD) for continuous variables. NOS, not otherwise specified.*

		ChemOnly	ChemBev	p-value	
Liver Metastases	Yes	127 (66)	465 (69)	0.606	
	No	64 (34)	211 (31)		
Peritoneum Metastases	Yes	53 (28)	185 (27)	0.926	
	No	138 (72)	491 (73)		
ECOG Performance Status	0	51 (27)	284 (42)	<0.001	***
	1	95 (50)	344 (51)		
	2	34 (18)	43 (6)		
	3	11 (6)	5 (1)		
Charlson Comorbidity Index (age-unadjusted)	0	127 (66)	487 (72)	0.001	***
	1	27 (14)	125 (18)		
	2	22 (12)	41 (6)		
	3	7 (4)	18 (3)		
	4 or more	8 (4)	5 (1)		
CEA Level	Recorded	726 (5254)	478 (2049)	0.546	
	Missing	19 (10)	61 (9)		
RAS Status	Mutated	56 (29)	247 (37)	<0.001	***
	Wild-type	72 (38)	317 (47)		
	Missing	63 (33)	112 (17)		
BRAF Status	Mutated	12 (6)	41 (6)	0.006	***
	Wild-type	52 (27)	266 (39)		
	Missing	127 (66)	369 (55)		

showed good model calibration. Bias-corrected C-statistics for survival Model LX-Time of 0.70, 0.68, and 0.67 at 0.5, 1, and 2 years respectively, indicated a reasonable model discrimination, and a bias-corrected slope for the predicted Weibull Scale of 0.998 indicated good calibration. Bias-corrected GNDs and p-values of 12.0 (0.02), 1.1 (0.90), and 5.6 (0.23) for 5 risk groups, and 26.9 (<0.01), 11.3 (0.26), and 13.3 (0.15) for 10 risk groups, at 0.5, 1, and 2 years respectively, indicated good calibration, except for short-term predictions (also see Figure 3).

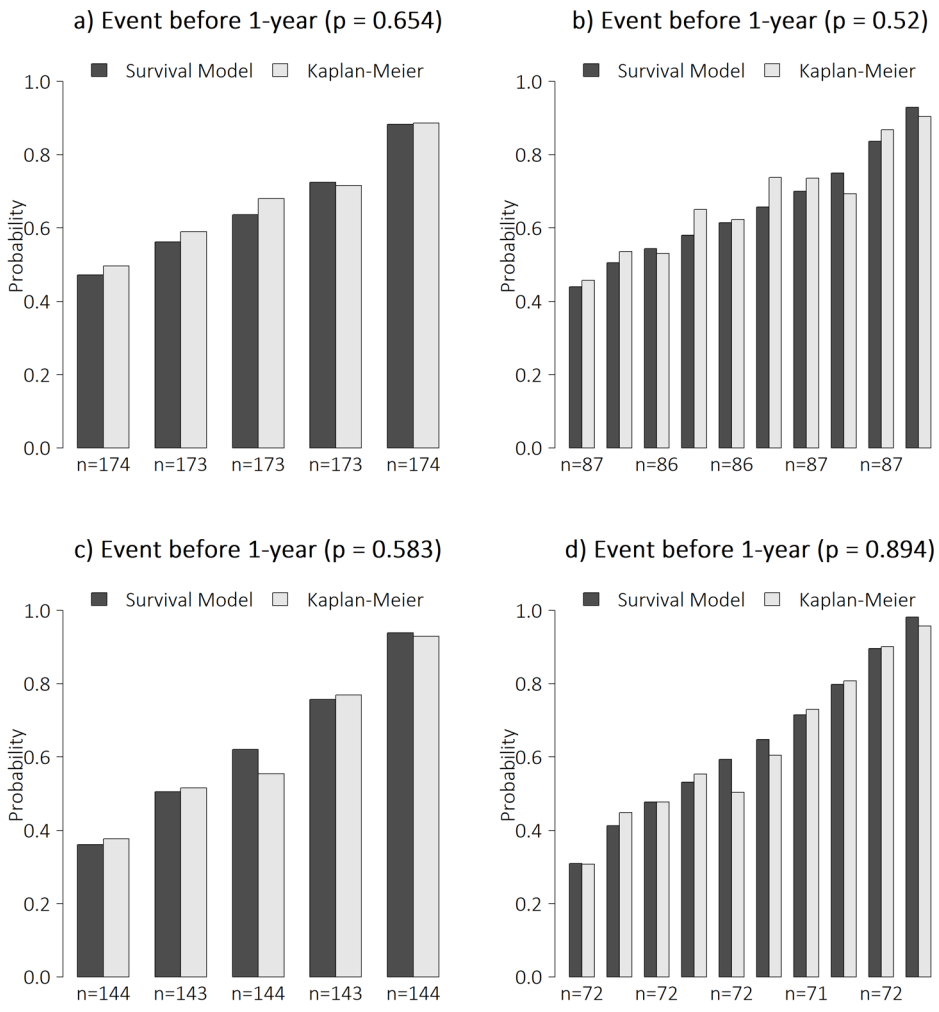


Figure 3. *Illustration of the Greenwood-Nam-D'Agostino statistic, showing predicted vs. observed risks for (a, b) first-line progression (Model L1-Time) and (c, d) overall survival after progression (Model LX-Time), and the impact of selecting (a, c) 5 risk groups and (b, d) 10 risk groups, at a 1-year timepoint and for one of the imputed datasets.*

Observed time-to-event Kaplan-Meier plots were similar to those simulated by the DES (Figure 4). For first-line treatment, observed median days until progression (95%-confidence intervals) were 171 (155, 210) and 304 (280, 329), and simulated medians were 199 (171, 228) and 308 (289, 328), for ChemOnly and ChemBev respectively. A median of 1055 days for time until death during first-line treatment was observed for ChemBev, and simulations yielded median 1171 days (1013, 1392). No observed confidence intervals for ChemBev nor any information for ChemOnly regarding this time-to-event could be estimated due to a low number of events. Regarding OS after first-line treatment, observed median days were 193 (153, 243) and 270 (237, 308), and simulated were 200 (161, 242) and 266 (243, 292), for ChemOnly and ChemBev respectively. Of all patients in the ChemOnly group 11.0% deaths were observed and for 6.8% of patients no event was observed, compared to 11.6% (7.1%, 16.6%) patients that died on first-line treatment in simulations. For ChemBev, 9.2% deaths were observed and 6.4% of patients were censored, whereas 10.3% (7.9%, 12.9%) deaths were simulated. Subgroup analyses showed similar results.

Exploratory scenario analyses using the DES estimated that median first-line PFS (95% confidence interval) of 219 (25%) patients could be improved from 175 days (156, 199) to 269 days (246, 294) by targeting a different treatment (Figure 5a). For 168 (77%) of these patients, ChemBev treatment was targeted rather than ChemOnly (observed), whereas ChemOnly rather than ChemBev (observed) was targeted for 51 (23%) patients. On a population level, this resulted in an estimated improved median first-line PFS of 288 days (270, 307) compared to 265 days (248, 280) as observed from the data (Figure 5b).

DISCUSSION

Incorporation of RWD into simulation techniques has the potential to identify optimal treatment sequencing strategies for specific patient subpopulations, which is unlikely to ever be achieved by performing clinical trials because of an infeasible number of required studies. Although simulations have been used in mCRC to estimate the cost-effectiveness of, for example, bevacizumab in multiple countries [53], several common treatment sequences [54, 55], and KRAS testing to target cetuximab treatment [56], these simulation studies did not account for patient heterogeneity and, therefore, cannot be used to identify optimal treatment strategies. To our knowledge, this is the first simulation that reflects patient heterogeneity based on RWD, combining patient and disease characteristics and conditional treatment outcomes, and can be used to estimate the effectiveness of different treatment combinations and sequencing strategies for specific patient subpopulations.

Although the internal validation showed reasonable to good performance, these statistics are hard to interpret and subject to several arbitrary decisions. For example, C-statistics for survival models do not take into account time differences between individuals, which implies that classifying a patient that progressed at day 100 as higher risk than a patient that progressed at day 95 is considered as wrong as classifying a patient that progressed at day 500 as higher risk than a patient that progressed at day 40. Hence, a discrimination statistic that considers the distance between individuals would be preferred. Another aspect that should

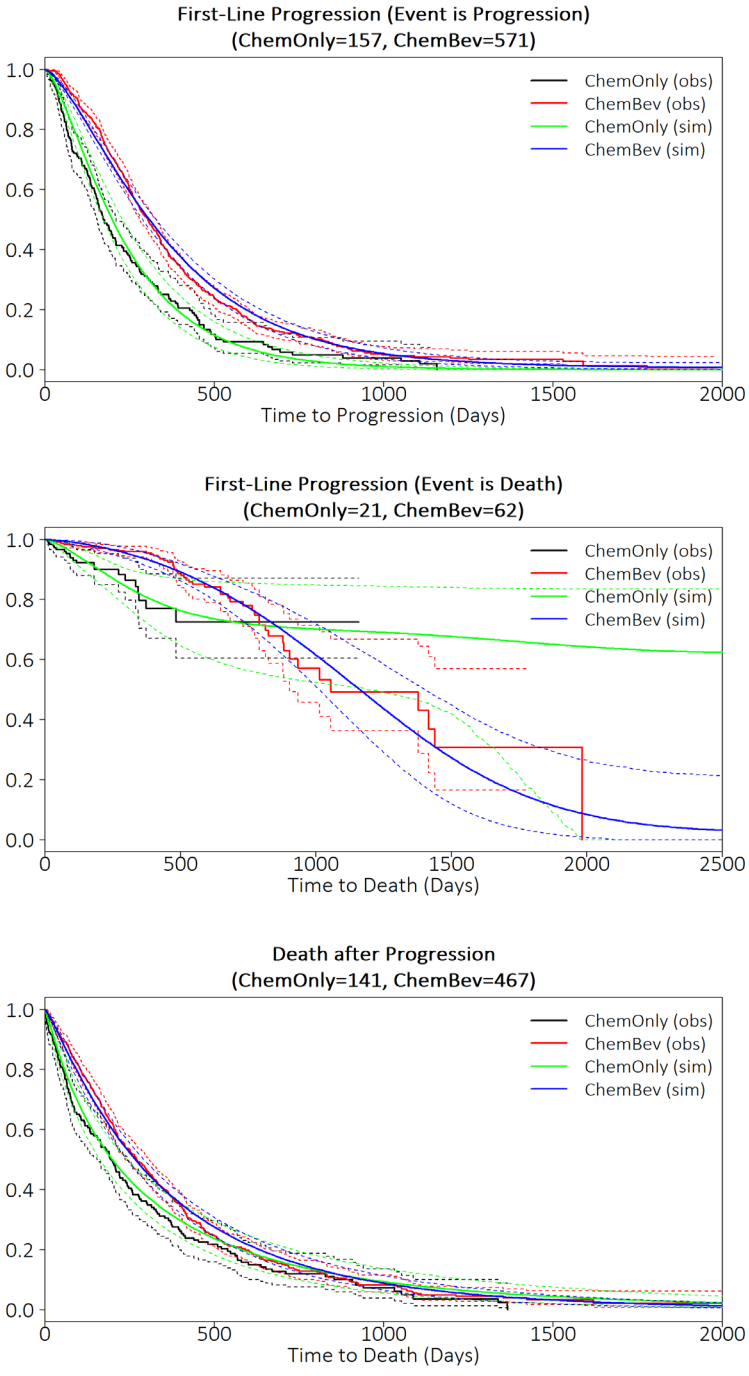


Figure 4. *Observed vs. simulated Kaplan-Meier plots for the discrete event simulation model, including the number of events in the corresponding patient groups and 95% confidence intervals (dashed lines) based on the Kaplan-Meier estimate (observed) and based on the probabilistic sensitivity analysis (simulated).*

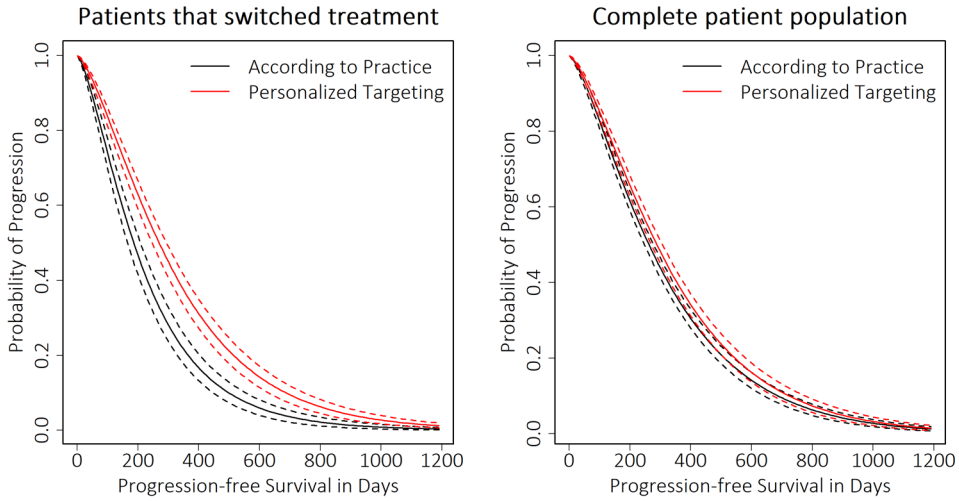


Figure 5. *Kaplan-Meier plots illustrating the progression-free survival (solid lines) and 95% confidence intervals (dashed lines) for the personalized treatment strategy compared to treatment as observed in the data, for the patients that switch treatment according to the personalized treatment strategy and the complete patient population.*

be considered when interpreting statistics such as the C-statistic and GND for survival models is the timepoint at which these are evaluated. For the GND specifically, also the selected number of risks groups may impact model performance outcomes. Furthermore, depending on the intended use of survival models, either model discrimination or calibration may be considered more important. Here, the aim is to predict time-to-events and model calibration may be considered most relevant. Finally, it is unknown how model performance is influenced by modeling the time-to-event of two competing events, i.e. progression and death, using one parametric survival model.

This study shows there is potential for translational research using RWD and simulation techniques, however, it is subject to limitations associated with using this source of evidence. First, only 191 patients were included for the ChemOnly treatment strategy compared to 676 patients for the ChemBev strategy, whereas clinical trials would have typically yielded groups of similar sizes. Additionally, significant differences were observed in important prognostic patient characteristics, such as ECOG performance status, highlighting the risk of selection bias as a potential driver of differences in survival outcomes. Although only patients with a palliative intent were included and multivariable models were developed, it is not guaranteed that this fully corrected for the impact of selection bias. Additional research efforts may be directed to investigating the use of propensity score matching to further address selection bias at an expected loss of observations given the low number of patients for the ChemOnly strategy, which may result in increased uncertainty surrounding simulation outcomes.

This study has certain additional limitations. A different treatment recoding strategy could have been applied, which might have resulted in inclusion of a different patient population.

Also, patients could have been restricted to those treated at a later moment in time to reduce the proportion of missing values for the RAS and BRAF mutational status. However, this would have decreased the number of observations substantially, being particularly troublesome for the ChemOnly treatment strategy. Finally, a rather inclusive variable selection procedure with high p-value thresholds was implemented to value clinical relevance. Although other thresholds, procedures, and criteria, such as using the Akaike information criterion, could have been used, it is unknown whether this would have resulted in different models.

Despite challenges regarding the selection and interpretation of validation measures, an important question is when models are to be considered "good enough", i.e. when are simulations accurate enough to be useful? Here, performance was considered sufficient to estimate population-level outcomes and, therefore, able to assess the impact of different treatment targeting decisions on a system level. An important next step in order to increase the informative value of DES is to add first-line treatment outcomes, such as best-achieved response, quality of life, and likelihood of receiving second-line treatment, as well as second-line treatment and outcomes conditional on first-line treatment outcomes. Additionally, treatment cost and cost of treating adverse events may be added to provide information on the cost-effectiveness of different treatment sequencing strategies. Using such a health economic simulation, optimization algorithms can be applied to obtain insights into key cost-effectiveness drivers and, ultimately, identify a personalized sequencing strategy that optimizes clinical outcomes given certain resource constraints. Although such personalized simulations have great potential for providing information on challenging treatment sequencing questions, they also pose questions regarding how complex these simulations can get before stakeholders no longer understand (and believe) them. Hence, additional research should also be done on how such simulations can potentially be used to support decision making on different levels, and how confidence in them can be established, as having a reasonable model could be better than having nothing at all.

ACKNOWLEDGEMENTS

This research was partly funded by the Netherlands Organisation for Health Research and Development (ZonMW) as part of the Translational Research Program (project number: 446001006). Roche Products Pty Limited (Australia) provided financial assistance for the development, installation and maintenance of the TRACC database. BioGrid Australia manages the TRACC database and provided data access and support.

REFERENCES

1. Van Cutsem E, et al. Metastatic colorectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2014;25(S3):1-9.

2. Douillard JY, et al. Irinotecan combined with fluorouracil compared with fluorouracil alone as first-line treatment for metastatic colorectal cancer: a multicentre randomised trial. Lancet. 2000;355(9209):1041-7.

3. Saltz LB, et al. Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer. N Engl J Med. 2000;343(13):905-14.

4. de Gramont A, et al. Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. J Clin Oncol. 2000;18(16):2938-47.

5. Giacchetti S, et al. Phase III multicenter randomized trial of oxaliplatin added to chronomodulated fluorouracil-leucovorin as first-line treatment of metastatic colorectal cancer. J Clin Oncol. 2000;18(1):136-47.

6. Kohne CH, et al. Phase III study of weekly high-dose infusional fluorouracil plus folinic acid with or without irinotecan in patients with metastatic colorectal cancer: European Organisation for Research and Treatment of Cancer Gastrointestinal Group Study 40986. J Clin Oncol. 2005;23(22):4856-65.

7. Hurwitz H, et al. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. N Engl J Med. 2004;350(23):2335-42.

8. Cunningham D, et al. Bevacizumab plus capecitabine versus capecitabine alone in elderly patients with previously untreated metastatic colorectal cancer (AVEX): an open-label, randomised phase 3 trial. Lancet Oncol. 2013;14(11):1077-85.

9. Guan ZZ, et al. Efficacy and safety of bevacizumab plus chemotherapy in Chinese patients with metastatic colorectal cancer: a randomized phase III ARTIST trial. Chin J Cancer. 2011;30(10):682-9.

10. Kabbinavar FF, et al. Addition of bevacizumab to bolus fluorouracil and leucovorin in first-Line metastatic colorectal cancer: results of a randomized phase II trial. J Clin Oncol. 2005;23(16):3697-705.

11. Tebbutt NC, et al. Capecitabine, bevacizumab, and mitomycin in first-line treatment of metastatic colorectal cancer: results of the Australasian Gastrointestinal Trials Group randomized phase III MAX Study. J Clin Oncol. 2010;28(19):3191-8.

12. Goldberg RM, et al. A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. J Clin Oncol. 2004;22(1):23-30.

13. Tournigand C, et al. FOLFIRI followed by FOLFOX6 or the reverse sequence in advanced colorectal cancer: a randomized GERCOR study. J Clin Oncol. 2004;22(2):229-37.

14. Falcone A, et al. Phase III trial of infusional fluorouracil, leucovorin, oxaliplatin, and irinotecan (FOLFOXIRI) compared with infusional fluorouracil, leucovorin, and irinotecan (FOLFIRI) as first-line treatment for metastatic colorectal cancer: the Gruppo Oncologico Nord Ovest. J Clin Oncol. 2007;25(13):1670-6.

15. Passardi A, et al. Effectiveness of bevacizumab added to standard chemotherapy in metastatic colorectal cancer: final results for first-line treatment from the ITACa randomized clinical trial. Ann Oncol. 2015;26(6):1201-7.

16. Saltz LB, et al. Bevacizumab in combination with oxaliplatin-based chemotherapy as first-line therapy in metastatic colorectal cancer: a randomized phase III study. J Clin Oncol. 2008;26(12):2013-9.

17. Cremolini C, et al. FOLFOXIRI plus bevacizumab versus FOLFIRI plus bevacizumab as first-line treatment of patients with metastatic colorectal cancer: updated overall survival and molecular subgroup analyses of the open-label, phase 3 TRIBE study. Lancet Oncol. 2015;16(13):1306-15.

18. Loupakis F, et al. Initial therapy with FOLFOXIRI and bevacizumab for metastatic colorectal cancer. J Engl J Med. 2014;371(17):1609-18.

19. Simkens LHJ, et al. Maintenance treatment with capecitabine and bevacizumab in metastatic colorectal cancer (CAIRO3): a phase 3 randomised controlled trial of the Dutch Colorectal Cancer Group. Lancet. 2015;385(9980):1843-52.

20. Loree JM, Kopetz S. Recent developments in the treatment of metastatic colorectal cancer. Ther Adv Med Oncol. 2017;9(8):551-64.

21. Temraz S, et al. Sequencing of treatment in metastatic colorectal cancer: where to fit the target. World J Gastroenterol. 2014;20(8):1993-2004.

22. Aparicio T, et al. Oxaliplatin- or irinotecan-based chemotherapy for metastatic colorectal cancer in the elderly. Br J Cancer. 2003;89(8):1439-44.

23. Stintzing S, et al. FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab for metastatic colorectal cancer (FIRE-3): a post-hoc analysis of tumour dynamics in the final RAS wild-type subgroup of this randomised open-label phase 3 trial. Lancet Oncol. 2016;17(10):1426-34.

24. Loupakis F, et al. FOLFOXIRI plus bevacizumab as first-line treatment in BRAF mutant metastatic colorectal cancer. Eur J Cancer. 2014;50(1):57-63.

25. Arnold D, et al. Prognostic and predictive value of primary tumour side in patients with RAS wild-type metastatic colorectal cancer treated with chemotherapy and EGFR directed antibodies in six randomized trials. Ann Oncol. 2017;28(8):1713-29.

26. Giordano SH. Comparative effectiveness research in cancer with observational data. Am Soc Clin Oncol Educ Book. 2015:e330-5.

27. Hershman DL, Wright JD. Comparative effectiveness research in oncology methodology: observational data. J Clin Oncol. 2012;30(34):4215-22.

28. McLean J, et al. Clinical practice patterns in chemotherapeutic treatment regimens for metastatic colorectal cancer. Clin Colorectal Cancer. 2016;15(2):135-40.

29. Teng C-LJ, et al. Optimal sequence of irinotecan and oxaliplatin-based regimens in metastatic colorectal cancer: a population-based observational study. PloS One. 2015;10(8):e0135673.

30. Clarke S, et al. An Australian translational study to evaluate the prognostic role of inflammatory markers in patients with metastatic ColorEctal caNcer Treated with bevacizumab (Avastin) [ASCENT]. BMC Cancer. 2013;13:120.

31. Field K, et al. Developing a national database for metastatic colorectal cancer management: perspectives and challenges. Intern Med J. 2013;43(11):1224-31.

32. Semira C, et al. Chemotherapy and biologic use in the routine management of metastatic colorectal cancer in Australia: is clinical practice following the evidence? Intern Med J. 2018.

33. Wong H-l, et al. Impact of primary tumor site on bevacizumab efficacy in metastatic colorectal cancer. Clin Colorectal Cancer. 2016;15(2):e9-e15.

34. Lee B, et al. The impact of bevacizumab in metastatic colorectal cancer with an intact primary tumor: Results from a large prospective cohort study. Asia Pac J Clin Oncol. 2017;13(4):314-21.

35. Rajagopalan H, et al. Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. Nature. 2002;418(6901):934.

36. Van Buuren S, et al. Fully conditional specification in multivariate imputation. J Stat Comput Simul. 2006;76(12):1049-64.

37. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. 2011. 2011;45(3):67.

38. Wood AM, et al. How should variable selection be performed with multiply imputed data? Stat Med. 2008;27(17):3227-46.

39. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J A Stat Ass. 1986;81(394):366-74.

40. Degeling K, et al. Comparing strategies for modeling competing risks in discrete event simulations: a simulation study and illustration in colorectal cancer. Med Decis Making. 2019;39(1):57-73.

41. Degeling K, et al. Accounting for parameter uncertainty in the definition of parametric distributions used to describe individual patient variation in health economic models. BMC Med Res Methodol. 2017;17(1):170.

42. Rahman MS, et al. Review and evaluation of performance measures for survival prediction models in external validation settings. BMC Med Red Methodol. 2017;17(1):60.

43. Uno H, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30(10):1105-17.

44. Harrell FE, et al. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-87.

45. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Stat Med. 2000;19(24):3401-15.

46. Demler OV, et al. Tests of calibration and goodness-of-fit in the survival setting. Stat Med. 2015;34(10):1659-80.

47. D'Agostino RB, Nam B-H. Evaluation of the performance of survival analysis models: discrimination and calibration measures. Handbook of Statistics. 2003; 23:1-25

48. Crowson CS, et al. Assessing calibration of prognostic risk scores. Stat Methods Med Res. 2016;25(4):1692-706.

49. Miller ME, et al. Validation of probabilistic predictions. Med Decis Making. 1993;13(1):49-57.

50. Arkes HR, et al. The covariance decomposition of the probability score and its use in evaluating prognostic estimates. Med Decis Making. 1995;15(2):120-31.

51. Steyerberg EW, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774-81.

52. R Core Team. R: a language and environment for statistical computing. Version 3.5.2 [software]. Available from: https://www.r-project.org/

53. Goldstein DA, et al. Bevacizumab for metastatic colorectal cancer: a global cost-effectiveness analysis. Oncologist. 2017;22(6):694-9.

54. Parikh RC, et al. Cost-effectiveness of treatment sequences of chemotherapies and targeted biologics for elderly metastatic colorectal cancer patients. J Manag Care Spec Pharm. 2017;23(1):64-73.

55. Toumazis I, et al. Comparative effectiveness of up to three lines of chemotherapy treatment plans for metastatic colorectal cancer. MDM Policy Pract. 2017;2(2):2381468317729650.

56. Westwood M, al. KRAS mutation testing of tumours in adults with metastatic colorectal cancer: a systematic review and cost-effectiveness analysis. Health Technol Assess. 2014;18(62).

APPENDIX 1: considered variable transformations

Table S1. *Candidate variables considered in the univariable modeling for the first-line progression-free survival (PFS) (Model L1-Time) and overall survival (OS) after progression (Model LX-Time) and their average p-values and ranks over all imputed datasets.*

		First-line PFS		OS after Progression	
Variables	**Transformations**	**p-value**	**Rank**	**p-value**	**Rank**
Age	Logarithm	0.874	4	0.006	5
	Square root	0.914	5	0.002	4
	1st Order Polynomial	0.722	3	0.001	3
	2nd Order Polynomial	0.005	1	<0.001	1
	3rd Order Polynomial	0.015	2	<0.001	2
Gender	Binary: Female (yes) vs. Male (no)	0.041		0.029	
Hospital Type	Binary: Private (yes) vs. Public (no)	0.132		0.009	
Stage at Diagnosis	Binary: Stage IV (yes) vs. Lower Stage (no)	0.018		0.465	
Primary Tumor Site	Categorical: Group(NOS and Multiple Primaries)	0.004	2	0.053	2
	Categorical: Group(Left Colon and Rectum) & Group(NOS and Multiple Primaries)	0.004	1	0.024	1
Metastatic Sites	Categorical: Group(3 or more)	0.402		0.679	
Liver Metastases	Binary: Yes vs. No	0.055		0.719	
Peritoneum Metastases	Binary: Yes vs. No	0.109		0.012	
ECOG Performance	Categorical: Group(2 and 3)	<0.001	1	<0.001	1
	Binary: 2 or more (yes) vs. 0 or 1 (no)	<0.001	2	<0.001	2
Charlson Index	Categorical: Group(2 or more)	0.173	2	0.994	3
	Categorical: Group(0 and 1) & Group(2 or more)	0.850	3	0.942	2
	Binary: 1 or more (yes) vs. 0 (no)	0.165	1	0.913	1
CEA Level	Binary: 200 or more (yes) vs. less than 200 (no)	0.006		0.267	
RAS Status	Binary: Mutated (yes) vs. Wild-type (no)	0.208		0.011	

Table S1 (continued). *Candidate variables considered in the univariable modeling for the first-line progression-free survival (PFS) (Model L1-Time) and overall survival (OS) after progression (Model LX-Time) and their average p-values and ranks over all imputed datasets.*

		First-line PFS		OS after Progression	
Variables	**Transformations**	**p-value**	**Rank**	**p-value**	**Rank**
BRAF Status	Binary: Mutated (yes) vs. Wild-type (no)	0.152		0.033	
First-line PFS*	Logarithm***	0.360		<0.001	2.5
	Square root	0.005		<0.001	2.5
	1st Order Polynomial	<0.001		<0.001	5
	2nd Order Polynomial**	<0.001		<0.001	2.5
	3rd Order Polynomial	<0.001		<0.001	2.5

** For First-line PFS the equivalent variable First-line PFS refers to the logistic regression model used to distinguish between progression or death as competing events (Model L1-Event) and not to survival Model L1-Time. ** For logistic regression Model L1-Event a second-order polynomial was selected, whereas a third-order polynomial had a slightly lower p-value, because problems occurred when fitting the third-order polynomial together with its treatment interaction. Also, for this variable no average rank could be calculated, as this variable is not imputed. *** For Model LX-Time, following equal ranks for several transformations, the logarithm was selected.*

APPENDIX 2: variable importance, and corresponding model coefficients and confidence intervals

See subsequent pages for Table S2.

Table S2. *Variables, their clinical relevance, and their corresponding model coefficients, for all three models: first-line progression-free survival (PFS) model (Model L1-Time), logistic regression Model L1-Event to distinguish between competing events progression and death, and overall survival (OS) after progression model (Model LX-Time) (NB: grey cells indicate that these variables were not applicable, hyphens indicate these coefficients were not included).*

	First-line Treatment			Subsequent Treatment / Follow Up	
	Model L1-Time		Model L1-Event	Model LX-Time	
Variable	**Relevance**	**Coefficient**	**Coefficient**	**Relevance**	**Coefficient**
Weibull Shape		1.3944 (1.3229, 1.4697)			1.1648 (1.0938, 1.2403)
Intercept		4.5352 (3.6608, 5.4097)	-2.0699 (-3.0113,-1.1284)		3.2892 (2.2621, 4.3163)
Progression-free Survival (PFS)					
PFS			-0.0002 (-0.0082, 0.0078)		
PFS2			0.0000 (0.0000, 0.0000)		
Progression-free Survival (PFS)				Moderate	
Log(PFS)					0.5752 (0.4054, 0.7449)
Treatment	High			High	
Chemotherapy Only		Reference	Reference		Reference
Chemotherapy and Bevacizumab		0.4904 (0.2176, 0.7631)	0.1738 (-1.2292, 1.5768)		0.8818 (-0.2813, 2.0449)
Age	High			High	
Age		0.0467 (0.0194, 0.0741)	-		-0.0251 (-0.037,-0.0132)
Age2		-0.0004 (-0.0006,-0.0002)	-		-0.0006 (-0.0012, 0.0000)
Gender	Low			Low	
Male		Reference	Reference		Reference
Female		-	-		-

	First-line Treatment			Subsequent Treatment / Follow Up	
	Model L1-Time		Model L1-Event	Model LX-Time	
Variable	**Relevance**	**Coefficient**	**Coefficient**	**Relevance**	**Coefficient**
Hospital Type	Low			Low	
Public		Reference	Reference		Reference
Private		0.3952 (0.1641, 0.6263)	-		0.1053 (-0.0459, 0.2565)
Stage at Diagnosis	Moderate			Moderate	
Lower than Stage IV		Reference	Reference		Reference
Stage IV		-	-		-0.5243 (-0.9757,-0.0729)
Primary Tumor Site	Moderate			High	
Colon NOS or Multiple		Reference	Reference		Reference
Left or Rectum		-0.0320 (-0.2807, 0.2167)	-		-0.1579 (-0.5383, 0.2224)
Right		-0.2048 (-0.4595, 0.0499)	-		-0.2365 (-0.6212, 0.1481)
Number of Metastatic Sites	Moderate			Moderate	
1		Reference	Reference		Reference
2		-	-		-
3 or more		-	-		-
Liver Metastases	Moderate			Moderate	
No		Reference	Reference		Reference
Yes		-0.1007 (-0.2197, 0.0183)	-		-
Peritoneum Metastases	Moderate			High	
No		Reference	Reference		Reference
Yes		-0.0967 (-0.2155, 0.0220)	-		0.1514 (-0.1774, 0.4801)

	First-line Treatment			Subsequent Treatment / Follow Up	
	Model L1-Time		Model L1-Event	Model LX-Time	
Variable	**Relevance**	**Coefficient**	**Coefficient**	**Relevance**	**Coefficient**
ECOG at Treatment Start	High			High	
0		Reference	Reference		Reference
1		-0.4420 (-0.7086,-0.1754)	-		-0.2762 (-0.666, 0.1136)
2 or 3		-0.4466 (-0.7602,-0.1331)	-		-1.2023 (-1.6308,-0.7738)
Charlson Index	Moderate			Moderate	
0		Reference	Reference		Reference
1 or more		-	0.6461 (0.1594, 1.1328)		0.1045 (-0.0566, 0.2656)
CEA Level x 1-2 Metastatic Sites	Low			Moderate	
CEA < 200 & 1-2 Metastatic Sites		Reference	Reference		Reference
CEA >= 200 & 1-2 Metastatic Sites		-0.0350 (-0.3387, 0.2687)	-		-
RAS Status	Low			Low	
Wild-type		Reference	Reference		Reference
Mutant		-0.0959 (-0.2360, 0.0443)	-		0.1092 (-0.3022, 0.5206)
BRAF Status	High			High	
Wild-type		Reference	Reference		Reference
Mutant		-0.1587 (-0.3947, 0.0773)	-		0.2746 (-0.3313, 0.8804)
PFS x Treatment					
PFS & ChemOnly			0.0023 (-0.006, 0.0106)		
PFS2 & ChemOnly			0.0000 (0.0000, 0.0000)		

	First-line Treatment			Subsequent Treatment / Follow Up	
	Model L1-Time		Model L1-Event	Model LX-Time	
Variable	**Relevance**	**Coefficient**	**Coefficient**	**Relevance**	**Coefficient**
PFS x Treatment					
Log(PFS) & ChemOnly					-0.1487 (-0.3475, 0.0501)
Age x Treatment					
Age & ChemOnly		-	-		0.0174 (0.0035, 0.0314)
Age2 & ChemOnly		-	-		0.0004 (-0.0003, 0.0011)
Gender x Treatment					
Reference & ChemOnly | ChemBev			Reference		
Female & ChemOnly			-		
Hospital Type x Treatment					
Reference & ChemOnly | ChemBev		Reference	Reference		Reference
Private & ChemOnly		-0.4110 (-0.6654,-0.1566)	-		-
Stage at Diagnosis x Treatment					
Reference & ChemOnly | ChemBev			Reference		Reference
Stage IV & Chemo Only			-		0.4904 (-0.0222, 1.0030)

	First-line Treatment			Subsequent Treatment / Follow Up	
	Model L1-Time		Model L1-Event	Model LX-Time	
Variable	Relevance	Coefficient	Coefficient	Relevance	Coefficient
Primary Tumor Site x Treatment					
Reference & ChemOnly | ChemBev		Reference	Reference		Reference
Left or Rectum & Chemo Only		-	-1.4623 (-2.4591,-0.4655)		-
Right & Chemo Only		-	-1.4690 (-2.5294,-0.4087)		-
Num. Metastatic Sites x Treatment					
Reference & ChemOnly | ChemBev			Reference		
2 & ChemOnly			-		
3 or more & ChemOnly			-		
Liver Metastases x Treatment					
Reference & ChemOnly | ChemBev		Reference	Reference		
Yes & ChemOnly		-	-		
Peritoneum Metastases x Treatment					
Reference & ChemOnly | ChemBev		Reference	Reference		Reference
Yes & ChemOnly		-	-		-0.2387 (-0.6130, 0.1356)

	First-line Treatment			Subsequent Treatment / Follow Up	
	Model L1-Time		Model L1-Event	Model LX-Time	
Variable	**Relevance**	**Coefficient**	**Coefficient**	**Relevance**	**Coefficient**
ECOG x Treatment					
Reference & ChemOnly | ChemBev		Reference	Reference		Reference
1 & ChemOnly		0.2337 (-0.0577, 0.5251)	-		-0.0931 (-0.5166, 0.3305)
2 or 3 & ChemOnly		-0.0886 (-0.4774, 0.3001)	-		0.3116 (-0.2211, 0.8442)
Charlson Index x Treatment					
Reference & ChemOnly | ChemBev			Reference		Reference
1 or more & ChemOnly			-		-
CEA x Metastatic Sites x Treatment					
Reference & ChemOnly | ChemBev		Reference	Reference		
CEA >= 200 & 1-2 Sites & ChemOnly		-0.2022 (-0.5296, 0.1253)	-		
RAS Status x Treatment					
Reference & ChemOnly | ChemBev		Reference	Reference		Reference
Mutant & ChemOnly		-	-		-0.4072 (-0.8556, 0.0413)
BRAF Status x Treatment					
Reference & ChemOnly | ChemBev		Reference	Reference		Reference
Mutant & ChemOnly		-	-		-0.6795 (-1.3769, 0.0178)

APPENDIX 3: selection of distribution type

For both parametric survival models that were used to simulate progression-free survival (PFS) for first-line treatment (Model L1-Time) and overall survival (OS) after progression for subsequent treatment lines or follow up (Model LX-Time) exponential, Weibull, logistic, and lognormal distributions were considered. The Weibull distribution was selected for both survival models, based on the log-likelihood (see Table S3 below, a higher log-likelihood is better), observed compared to simulated Kaplan-Meier plots (see Figure S1), and quantile-quantile (Q-Q) plots (see Figure S2 and Figure S3).

Table S3. *Log-likelihood for all considered distributions for both first-line PFS and OS after progression.*

Distribution	First-line PFS	OS after Progression
Exponential	-5522.443	-4207.651
Weibull	-5483.372	-4207.648
Logistic	-5646.513	-4430.506
Lognormal	-5512.667	-4245.090

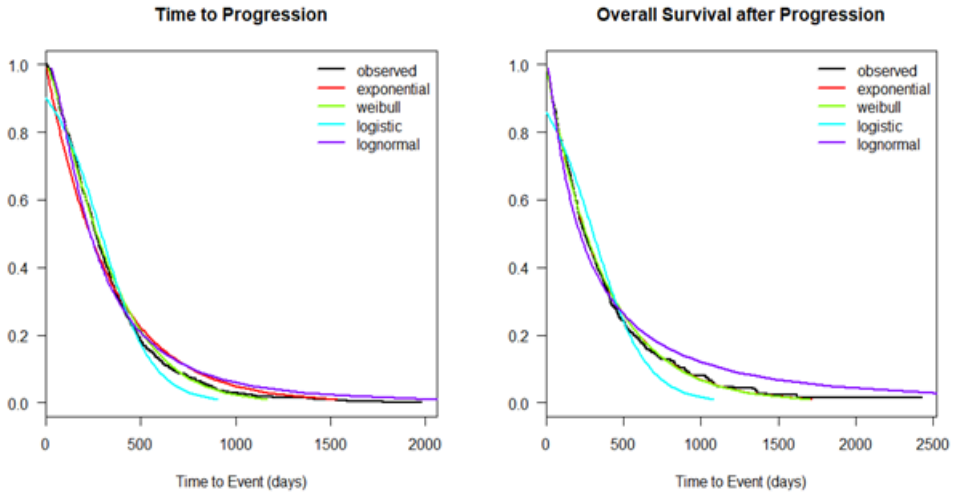


Figure S1. *Kaplan-Meier plots for the observed data compared to the considered distributions for both first-line PFS and OS after progression (NB: the red line for the exponential distribution almost completely overlaps with the green line for the Weibull distribution).*

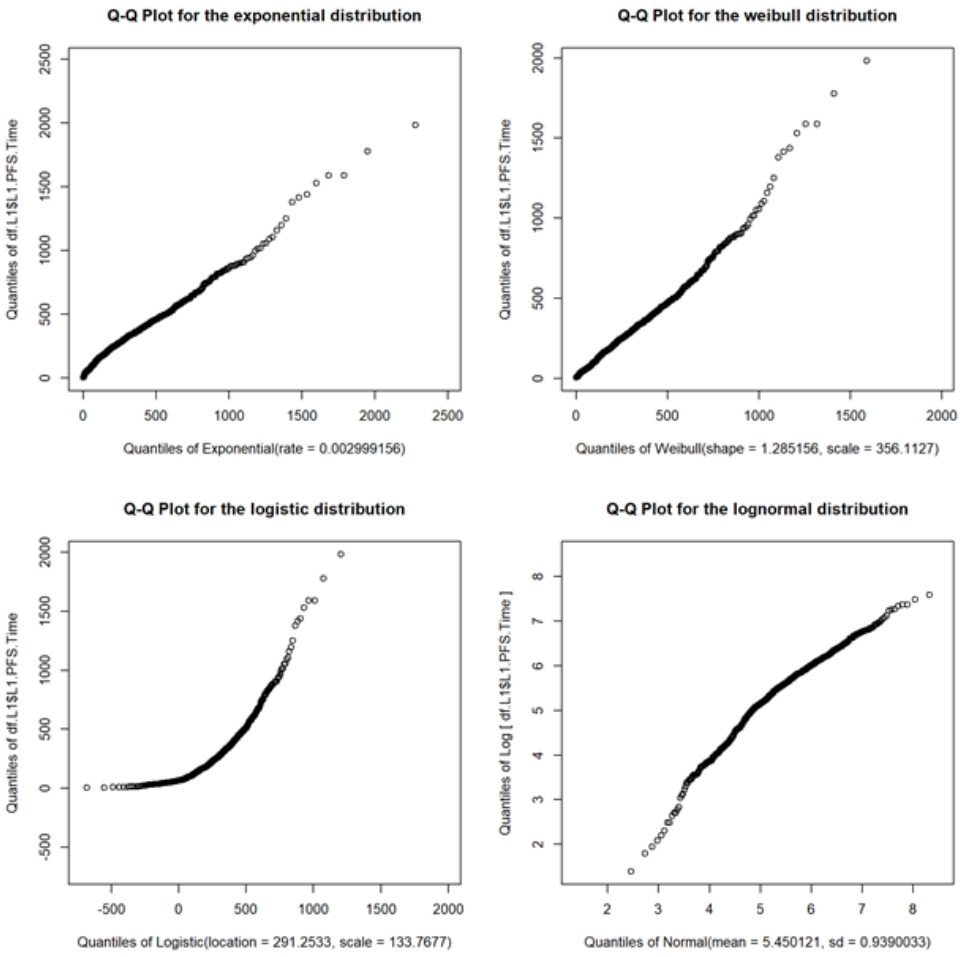


Figure S2. *Q-Q plots for the considered distributions for first-line PFS.*

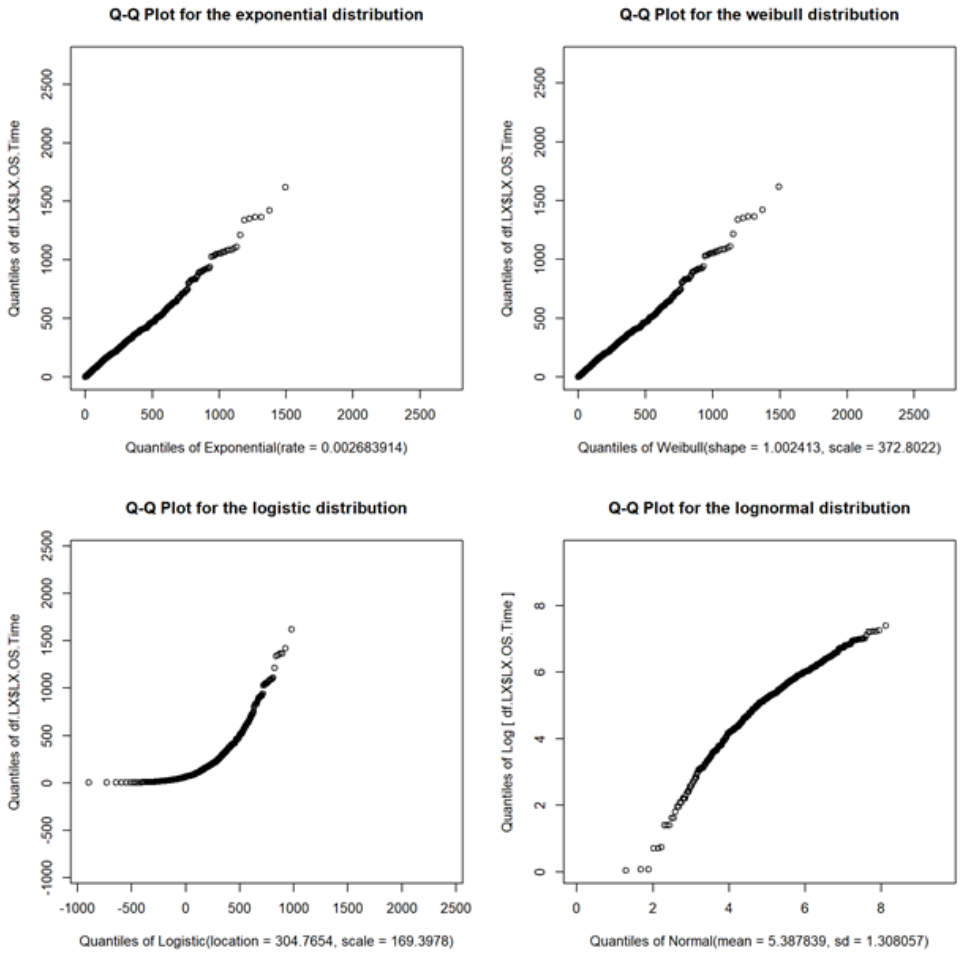


Figure S3. *Q-Q plots for the considered distributions for OS after progression.*

APPENDIX 4: variable selection process for survival models

For both survival models separately, variables with a high clinical relevance were always included. Moderately relevant variables were subject to a backward selection procedure in which variables with p-values higher than 0.3 were iteratively excluded. If multiple variables had a p-value higher than 0.3, the variable corresponding to the highest p-value was excluded first, and new p-values for all remaining candidate variables were obtained by fitting a new survival model without the excluded variable. All variables excluded in this backward selection procedure were added to a list of candidate variables for a forward selection procedure, together with all variables scored at a low clinical relevance. In this forward selection procedure, candidate variables with a p-value lower than 0.2 were iteratively included one at a time. After the backward and forward selection, treatment interaction effects were considered for all included variables in another iterative backward selection procedure with a p-value threshold of 0.3, to allow for differences between treatments based on patient and disease characteristics and biomarkers. Also see Figure S4 for a graphical representation of this variable selection process. In each step of a selection procedure, the survival model was fitted to all imputed datasets and Rubin's rule was applied to obtain one pooled model. A Wald test p-value for each variable was calculated based on this pooled model, combining all coefficients that related to a specific variable.

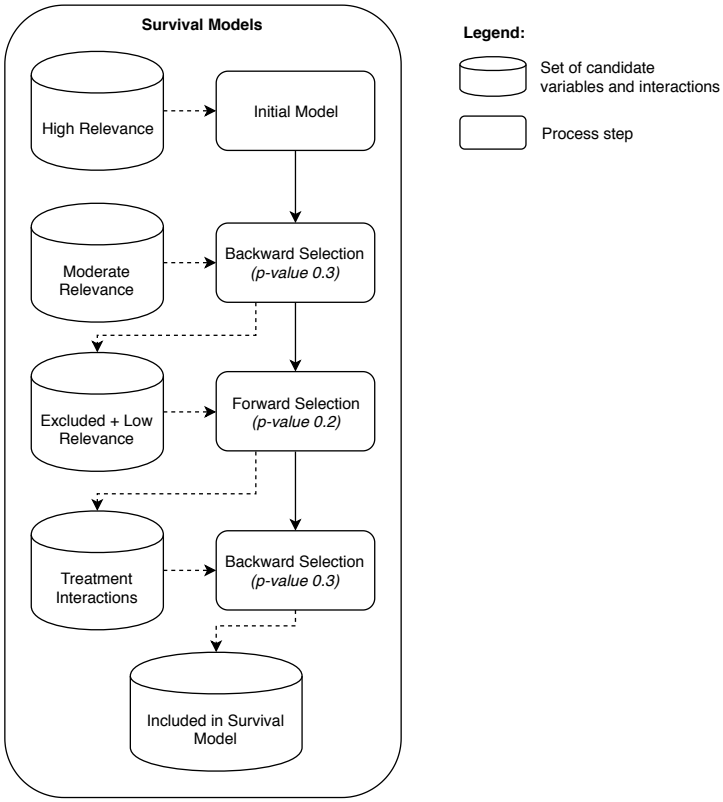


Figure S4. *Overview of the variable selection process.*

APPENDIX 5: overlap of time-to-event distributions

Overlap was assessed graphically and numerically. According to the approach explained below to quantify the overlap, overlap was 72.5%. A graphical representation of the overlap is provided in Figure S5. From both the percentage of overlap and the density plots, it can be concluded that there is a very substantial amount of overlap.

No gold standard approach for calculating the overlap between multiple (competing) distributions is available. The approach used in the current study defines the overlap between distributions as the mean overlap percentage over all distribution-specific overlaps, for which the distribution-specific overlap is defined as the percentage of the distribution that is overlapping with any other distribution. More precisely, the calculation can be performed according to the following steps:

1. Define a pre-specified number of equally sized bins that cover the observations of all distributions
2. Calculate the density of all distributions for each bin (1)
3. For each distribution, calculate the "overlap density" in each of the bins (1) by determining how much the density (2) overlaps with the corresponding highest density of the other distributions
4. Calculate the distribution-specific overlap by dividing the area under the "overlap density" curve (3) by the area under overall density curve (2)
5. Calculate the overall overlap between all distributions as the average of the distribution-specific overlaps (4)

R code used to calculate the overlap between competing time-to-event distributions is provided on the next page.

Figure S5. *Density plots of the time-to-event distributions illustrating the amount of overlap.*

```
# This function calculates the overlap between competing
# time-to-event distributions based on a 'data' object
# containing two columns, i.e. 'time' and 'event'.

getOverlap <- function(data, n.bins=512, digits=3) {

 # Function inputs:
 # - data:     data.frame with two columns:
 #           + time:     column containing the times
 #           + event:    column containing the events corresponding to the times
 # - n.bins:   the number of bins used for the 'density' function
 # - digits:   the number of digits to round the results to

 # Load package for calculating the Area Under the Curve (AUC)
 library(flux);

 # Get the events in the data
 events <- sort(unique(data$event));
 n.events <- length(events);

 # Get event-specific times
 times <- lapply(events, function(e) {
  subset(data, event==e)$time;
 });

 # Density settings
 min <- min(data$time);
 max <- max(data$time);
 n <- n.bins;
 x <- seq(from=min, to=max, length.out=n);

 # Get event-specific densities
 y <- sapply(1:n.events, function(i1) {
  density(times[[i1]], from=min, to=max, n=n)$y;
 });

 # Get event-specific density overlap
 yo <- sapply(1:n.events, function(i2) {

  sapply(1:n, function(j){

    own <- y[j,i2];
    other <- max(y[j,which(c(1:n.events)!=i2)]);
```

```
    if(own<=0) {
      return(0)
    } else
      return(min(own,other))

  })
});

# Calculate event-specific overlap
o <- sapply(1:n.events, function(i3) {
  (auc(x=x, y=yo[,i3]) / auc(x=x, y=y[,i3])) * 100;
});

return(round(mean(o), digits=digits));
}
```

CHAPTER 12

General discussion

GENERAL DISCUSSION

The discovery and introduction of new therapies and biomarkers in cancer treatment have accelerated the development of personalized oncology, aiming to provide patients with the best possible treatment based on patient and disease characteristics. These developments have contributed to improved health outcomes, though also present considerable challenges. Excessive cancer drug prices, for example, challenge the sustainability of healthcare systems, and the numerous therapies and biomarkers available pose challenges regarding their selection and combination across multiple lines of treatment. This complexity is likely to result in suboptimal use of therapies and biomarkers, increasing clinical practice variation and the delivery of ineffective and inefficient treatment. Suboptimal use of targeted and immunotherapies in clinical practice may also cause their real-world health economic impact to be different from that estimated based on clinical trial evidence, as these trials are typically not subject to such inefficiencies.

To address these challenges by evaluating new therapies and different treatment sequencing strategies, adaptive clinical trial designs and the use of real-world data have been proposed as alternatives to traditional randomized controlled trials. In addition, in order to optimize personalized oncology, this thesis advocates and illustrates that patient-level simulation models are needed to combine evidence from different (types of) sources, to extrapolate beyond time horizons of clinical studies, and to perform what-if scenario analyses. In doing so, papers presented in this thesis addressed questions regarding the added value and transparency of patient-level simulation methods compared to their cohort-level equivalents, and regarding their selection, appropriate use and potential.

Simulation modeling of health and economic outcomes for personalized clinical pathways

Simulation models need to adequately represent clinical practice and reflect the true nature of the evidence used to define them. In a personalized oncology context, this entails reflecting patient and disease characteristics and their relations with outcomes. Neglecting to do so may result in biased outcomes and, potentially, suboptimal resource allocation and treatment decisions [1]. This has also been recognized by policy makers, such as the National Institute for Health and Care Excellence (NICE) [2]. It is therefore remarkable that Chapter 2 found both the extent to which aspects of personalized clinical pathways are represented in simulation models, and the uptake of patient-level simulation modeling methods, to be limited within health economics. This observation was confirmed for personalized oncology specifically by a review of 13 recent NICE technology appraisals on cancer treatment sequencing, showing 12 out of 13 appraisals to be based on cohort-level simulation methods [3].

The limited uptake of patient-level simulation modeling methods is surprising because of several reasons. First, it has been widely questioned whether the most commonly used cohort-level state-transition modeling method is able to adequately represent aspects of personalized clinical pathways, whereas this is a strength of patient-level methods [4]. Second, when attempting to represent aspects of personalized clinical pathways in cohort-level simulation models, the number of model states required may increase exponentially with every additional characteristic considered [5]. This is likely to result in poor model transparency

compared to a patient-level simulation models developed for the same purpose [1, 6]. As was illustrated in Chapter 4, discrete event simulation and timed automata, which are both patient-level methods, can be used to transparently model biomarker-based treatment decisions in metastatic castration-resistant prostate cancer. Furthermore, Chapter 3 illustrated that the evidence structure of a discrete event simulation developed based on individual patient data, was considered more straightforward to interpret compared to that of a cohort-level state-transition model. This suggests that patient-level simulation models may also be more transparent when representing pathways that are not personalized. Third, it is a misconception that patient-level simulation modeling methods require more evidence compared to cohort-level methods, which was illustrated in Chapter 3, as additional evidence is only required when (more) aspects of personalized pathways are considered.

Despite patient-level simulation modeling methods can better represent personalized oncology pathways, there apparently are insufficient incentives for modelers to brave remaining barriers regarding the use of these methods. A lack of experience in applying patient-level methods may be an important barrier for those that do have experience with cohort-level methods. Additionally, several more technical barriers may have detained modelers from adopting patient-level methods, including a lack of guidance on applying these methods and the computational cost of evaluating the resulting models [1, 7]. This thesis has therefore made a significant contribution to the guidance for developing patient-level simulation models (Chapters 5 to 7). Furthermore, this thesis also provides solutions to the computational challenges associated with the analyses of patient-level simulation models (Chapters 8 to 10).

Although this thesis made a contribution to widespread use of patient-level simulation methods, it still remains a question where patient-level models add most value. For instance, when modeling treatment pathways without reflecting patient and disease characteristics, use of patient-level methods may not yield better outcome estimates. However, use of patient-level methods might still result in models that are more straightforward to interpret. Furthermore, developing a patient-level model provides modelers the flexibility to update the model at a later stage by adding aspects of personalization. When modeling personalized oncology care pathways and considering patient and disease characteristics, patient-level methods are required to obtain transparent models and (potentially) unbiased outcome estimates.

When applying patient-level simulation modeling methods, extensive literature is available to guide selection between alternatives [4, 6, 8, 9]. Patient-level state-transition modeling and discrete event simulation are two frequently applied patient-level methods that are potentially useful for modeling personalized oncology pathways. Alternative methods that have been introduced and applied, such as agent-based or dynamic transmission modeling [10] and timed automata (Chapter 4), are especially suitable for modeling interactions and behavior. When selecting between state-transition modeling and discrete event simulation, outcomes are not expected to be meaningfully different if the evidence used and level of detail is the same and no resource constraints are considered in the discrete event simulation [11-14]. However, discrete event simulation may represent clinical practice more realistically due to its flexibility towards defining states and handling of time (Chapter 3). This flexibility also

makes it possible to resemble a state-transition model throughout, or just in a specific part of a discrete event simulation model, by handling time in discrete cycles (Chapter 4). Dedicated software environments and technical skills required to develop discrete event simulation models have been considered an important disadvantage of this method, as state-transition models can be implemented using spreadsheets, for example. However, nowadays discrete event simulation can also be implemented in widely-used software, such as R [15, 16], and concerns regarding the transparency of using spreadsheets further mitigate this disadvantage. This thesis therefore argues discrete event simulation to be preferred over patient-level state-transition modeling for evaluating and optimizing personalized oncology pathways.

Methodological guidance for developing patient-level simulation models

Patient-level simulation modeling methods provide great flexibility in terms of model structure and evidence format and, therefore, modelers need to make several design choices. Methodological and practical guidance is essential to inform about which choices are to be made, alternatives to these choices, and to provide suggestions for the selection and use of alternative approaches. Although a general body of guidance for applying patient-level simulation modeling methods was already available, comprehensive methodological and practical guidance was lacking with regard to two important aspects: reflecting uncertainty in patient-level models and implementing competing risks in discrete event simulations.

It is crucial for simulation models to reflect uncertainty surrounding the outcomes in order to be meaningful to decision makers, as has been recognized by policy makers and in good simulation modeling practices guidelines [17, 18]. Chapter 5 has bridged an important gap with regard to guidance on accounting for parameter uncertainty (i.e., second order uncertainty) in the definition of parametric distributions used to describe individual patient variation (i.e., stochastic uncertainty or first-order uncertainty). Another topic of continued debate is structural uncertainty, as it has been shown that assumptions regarding model structure impact modeling outcomes [19]. Although guidance is available on accounting for structural uncertainty [20-22], this form of uncertainty is still rarely reflected in modeling studies and a need for additional practical guidance has been identified [23]. Further research may therefore explore a combined approach for representing structural and parameter uncertainty simultaneously, and provide practical guidance if this approach shows to be effective. Importantly, structural uncertainty is not specific to patient-level simulation models and the need for further guidance should, therefore, not be considered a disadvantage of patient-level methods compared to cohort-level methods.

In addition to uncertainty, accounting for competing risks is another important aspect to developing health economic simulation models [24]. Chapter 6 and Chapter 7 provided methodological and practical guidance on how competing events can be implemented in discrete event simulation models. These studies considered both uncensored and censored individual patient time-to-event data. Although the study regarding uncensored data (Chapter 6) confirmed general simulation modeling good practices recommendations [25], the study concerning censored data illustrated that the presence of censoring resulted in too much complexity to make a single general recommendation (Chapter 7). Another need for guidance

was identified while modeling the effectiveness of different treatment targeting strategies in metastatic colorectal cancer (Chapter 11), that is, how reflecting patient heterogeneity using multivariable survival models impacts the performance of different modeling approaches for implementing competing events. However, since representing heterogeneity adds to the complexity resulting from censoring alone, it is unlikely that general recommendations in this context can be made. Modelers are therefore left to assess the (internal) validity of their models to determine which approach is to be used when implementing competing events in discrete event simulations that comprise (multiple) multivariable survival models.

The level of complexity resulting from considering censoring amongst other data and disease pathway characteristics highlights an essential question: How complex can simulation become so that it is possible to maintain confidence and actually use their outcomes for decision making? Although this question will be discussed from a multidisciplinary stakeholder perspective later, there is potential for methodological guidance on validation of simulation models to contribute to this discussion. Guidance on the validation of health economic simulation models is available [26, 27], though specific measures that can be used to assess simulation model accuracy are not discussed. Health economic simulation models represent survival outcomes (i.e., time-to-event outcomes) and developing accuracy measures for such outcomes is challenging due to presence of competing risks and censoring. Several statistical measures have been modified to assess discrimination and calibration of single survival models [28-31], but these are difficult to interpret, lack clinical relevance, are subject to arbitrary decisions (Chapter 11). Furthermore, health economic simulation models typically include multiple survival models, and measures to assess the validity of such composite simulation models are lacking. Hence, research is desired into informative (i.e., clinically relevant) measures that can be used to validate simulation models that comprise multiple survival models.

As for structural uncertainty, questions regarding the complexity of simulation models are not only applicable to patient-level simulation models and the need for further guidance should, therefore, not be considered a disadvantage of patient-level methods only. With several chapters providing insight into the appropriate use of patient-level simulation modeling methods, modelers are now provided with additional guidance and tools to support them in adequately implementing these methods.

Metamodeling to reduce computational burden of analyzing simulation models

The amount of computational resources required to simulate (sufficiently) large cohorts of individual patients has been a barrier to the adoption of patient-level simulation methods [7]. Currently, simulating tens of thousands of hypothetical patients to obtain stable outcome estimates in a single model evaluation, has become less of an issue due to improved computational power. However, repeatedly doing so to perform advanced analyses or apply techniques that require thousands to millions of model evaluations, e.g. value of information analysis or optimization algorithms, still may not be possible within feasible timeframes. Metamodels have been used for decades in many fields of research to address issues regarding the computational cost of simulation models and analyses [32], and have also been proposed in health economic simulation modeling good practices guidelines [25]. Nevertheless, Chap-

ter 8 found that metamodeling methods have not yet been frequently used to the reduce computational cost of analyses with health economic simulation models. This study also found that computational challenges are applicable to both patient-level and cohort-level simulation models, confirming that not only the model type itself, but also the analysis performed determines whether runtime issues may occur.

An explanation for the limited uptake of metamodeling methods within health economics might be that computational burden is just not (yet) an issue. However, given the uptake of patient-level simulation modeling methods and advanced analyses, runtime issues seem inevitable. Other reasons for the limited uptake of metamodeling methods may include a limited awareness of the potential of these methods and lack of guidance on applying them in a health economic context. Chapter 9 addressed the lack of guidance by providing a structured overview of metamodeling methods appropriate for modeling health economic outcomes, including directions for their selection and use. Chapter 10 illustrated how a metamodel enabled optimization of a colorectal cancer screening program in the Netherlands. Combined with the potential applications as identified by the review, such as to perform probabilistic sensitivity analysis or enable simulations to be performed in decisions-support tools that require immediate responses, there is clear potential for metamodeling to reduce computational burden of (advanced) analyses of health economic models. Assuming adequate simulation models of personalized oncology pathways to be available, metamodeling methods may enable evaluation and optimization of such pathways when this is not possible within feasible timeframes using the original simulation models.

Simulation modeling to evaluate and optimize targeted treatment strategies

The remaining question to be answered in this thesis concerns the potential of simulation models to evaluate and optimize personalized oncology pathways and, thereby, contribute to addressing the challenges regarding the sustainability of healthcare systems and complexity of treatment sequencing. This potential was illustrated in Chapter 4 by modeling and evaluating the use of circulating tumor cells to inform treatment switching decisions in metastatic castration-resistant prostate cancer. Additionally, Chapter 11 illustrated how a simulation model developed based on real-world registry data was used to improve health outcomes of metastatic colorectal cancer patients, by selecting the treatment most likely to be effective in terms of progression-free survival on an individual level. To further illustrate their potential, Chapter 10 combined simulation modeling methods with metamodeling to optimize colorectal cancer screening in the Netherlands, given a constraint on the colonoscopy capacity. Although results from these studies cannot be directly implemented into clinical practice at this point, obtaining similar insights by performing (additional) randomized studies would have been very time and resource consuming, if feasible at all. Furthermore, although observational real-world data, for example from clinical registries, can provide valuable insights into practice variation and the real world impact of therapies, optimal care pathways cannot be derived from such data without further analysis. Altogether, there is a clear potential for simulation modeling methods to go beyond clinical studies and utilize evidence from different (types of) sources to optimize personalized oncology care strategies.

Despite their potential, it is important to interpret results from simulation models with care, as limitations may arise from model structures and the evidence used. One of the main reasons for developing simulation models is to extrapolate beyond the time horizon of available data or scenarios represented in those data. As illustrated by the case study on colorectal cancer screening in the Netherlands (Chapter 10), the best performing screening strategies may be extrapolations beyond those covered by the data used to populate the simulation model, and outcomes of these strategies should therefore be interpreted with care. Furthermore, the case study on personalized metastatic colorectal cancer treatment used real-world data from a clinical registry to estimate the effectiveness of different treatment targeting strategies (Chapter 11). Consequently, conclusions based on outcomes from this simulation model should be drawn with care, as these data used to populate the model are subject to selection bias. If stakeholders are not aware of issues like these, results from simulation modeling studies may be misinterpreted. Considerations regarding the interpretation of simulation outcomes resulting from assumptions and the data underlying a simulation model should, therefore, be clearly communicated about.

Remaining challenges and future perspectives

It is apparent that patient-level simulation modeling methods are valuable tools to address challenges regarding the sustainability and complexity of personalized oncology. However, to translate these methods from a rather isolated modeling environment into (clinical) decision making practice, several questions are to be answered: Which questions are to be answered using simulation models, and under which circumstances? How will be determined whether simulation models are good enough to inform decisions, and what is meant with good? Finally, when optimizing personalized oncology care pathways, what is going to be optimized and given which constraints?

Multidisciplinary collaborations should identify whether simulation models will be used mainly for direct comparisons of pre-defined care strategies (i.e., evaluation), or can also be used to identify what might be the best care strategy from a set of candidate strategies (i.e., optimization). Additionally, they can be used to inform clinical trial designs [33], or to structure and visualize (big) datasets and care pathways to better understand the complexity of personalized oncology. It is also important to discuss whether data from randomized clinical studies will be required to populate simulation models, or that real world data suffice if bias is accounted for as much as possible and uncertainty in the outcomes is adequately reflected. As long as high-quality simulation models are developed and their outcomes are interpreted carefully, having estimates based on models populated using observational data may be better than not having any estimates at all.

As models are becoming increasingly complex when more patient and disease characteristics are being considered, more effort is required to establish confidence in them. Although external code review and making source codes publicly available enhance transparency and reproducibility [34], this does not guarantee all relevant aspects are reflected and outcomes to be accurate enough to inform decisions. Therefore, multidisciplinary research efforts should establish a set of key aspects of personalized oncology pathways that should at least be rep-

resented in simulation models. Furthermore, as discussed before, relevant performance measures should be identified, including a minimum performance level for each of them.

Finally, personalized oncology pathways can be optimized from different perspectives. For example, optimization can be performed from a population perspective (e.g., clinical guidelines) or from a patient perspective (e.g., treatment targeting decisions). Whereas clinicians and patients may be interested in optimizing health outcomes while limiting the out-of-pocket expenses to patients, policy makers may aim to optimize population outcomes given a limited amount of resources that are available. Multidisciplinary research should aim to develop aggregated optimization objectives that represent all relevant perspectives, for example, by applying multi-criteria decision analysis methods.

Conclusions

There is great potential for patient-level simulation modeling methods to address challenges regarding the economic burden and complexity of personalized oncology. By estimating health and economic outcomes of personalized care pathways based on patient and disease characteristics, simulation models can support decisions about the selection and combination of therapies and biomarkers across multiple lines of treatments. Methodological and practical guidance on the application of patient-level simulation methods is now available and provides modelers with the tools required to utilize these methods and appropriately estimate health and economic outcomes of personalized oncology care pathways. It has been illustrated that simulation methods can be used to improve health outcomes for patients, while making efficient use of scarce health care resources. Further multidisciplinary research efforts should aim to translate this conceptual and methodological work into practice. These efforts may identify for what type of research questions and under which circumstances patient-level simulation modeling methods are most valuable, how confidence in simulation models can be established, and which outcomes are key in optimizing personalized oncology care pathways.

REFERENCES

1. Caro JJ, et al. Discrete event simulation: the preferred technique for health economic evaluations? Value Health. 2010;13(8):1056-60.

2. Davis S, et al. NICE Decision Support Unit Technical Support Documents. NICE DSU Technical Support Document 15: Cost-effectiveness modelling using patient-level simulation. London: National Institute for Health and Care Excellence (NICE); 2014.

3. Zheng Y, et al. Modeling treatment sequences in pharmacoeconomic models. PharmacoEconomics. 2017;35(1):15-24.

4. Stahl JE. Modelling methods for pharmacoeconomics and health technology assessment. Pharmacoeconomics. 2008;26(2):131-48.

5. Siebert U, et al. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. Value Health. 2012;15(6):812-20.

6. Brennan A, et al. A taxonomy of model structures for economic evaluation of health technologies. Health Econ. 2006;15(12):1295-310.

7. Karnon J, Haji Ali Afzali H. When to use discrete event simulation (DES) for the economic evaluation of health technologies? A review and critique of the costs and benefits of DES. Pharmacoeconomics. 2014;32(6):547-58.

8. Caro JJ, et al. Modeling good research practices—overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Value Health. 2012;15(6):796-803.

9. Marshall DA, et al. Selecting a dynamic simulation modeling method for health care delivery research—Part 2: report of the ISPOR Dynamic Simulation Modeling Emerging Good Practices Task Force. Value Health. 2015;18(2):147-60.

10. Pitman R, et al. Dynamic transmission modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-5. Value Health. 2012;15(6):828-34.

11. Jahn B, et al. Lessons learned from a cross-model validation between a discrete event simulation model and a cohort state-transition model for personalized breast cancer treatment. Med Decis Making. 2016;36(3):375-90.

12. Karnon J. Alternative decision modelling techniques for the evaluation of health care technologies: markov processes versus discrete event simulation. Health Econ. 2003;12(10):837-48.

13. Simpson KN, et al. Comparison of markov model and discrete-event simulation techniques for HIV. Pharmacoeconomics. 2009;27(2):159-65.

14. Standfield LB, et al. An empirical comparison of markov cohort modeling and discrete event simulation in a capacity-constrained health care setting. Eur J Health Econ. 2017;18(1):33-47.

15. R Core Team. R: a language and environment for statistical computing [software]. Available from: https://www.r-project.org/

16. Ucar I, et al. simmer: discrete-event simulation for R. arXiv preprint arXiv:170509746. 2017.

17. Briggs AH, et al. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. Value Health. 2012;15(6):835-42.

18. Claxton K, al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. Health Econ. 2005;14(4):339-47.

19. Frederix GWJ, et al. The impact of structural uncertainty on cost-effectiveness models for adjuvant endocrine breast cancer treatments: the need for disease-specific model standardization and improved guidance. Pharmacoeconomics. 2014;32(1):47-61.

20. Haji Ali Afzali H, Karnon J. Exploring structural uncertainty in model-based economic evaluations. Pharmacoeconomics. 2015;33(5):435-43.

21. Jackson CH, et al. A framework for addressing structural uncertainty in decision models. Med Decis Making. 2011;31(4):662-74.

22. Strong M, Oakley JE. When is a model good enough? Deriving the expected value of model improvement via specifying internal model discrepancies. SIAM/ASA J Uncertain. 2014;2(1):106-25.

23. Ghabri S, et al. Towards a new framework for addressing structural uncertainty in health technology assessment guidelines. Pharmacoeconomics. 2018;36(2):127-30.

24. Caro JJ, Möller J. Decision-analytic models: current methodological challenges. Pharmacoeconomics. 2014;32(10):943-50.

25. Karnon J, et al. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. Value Health. 2012;15(6):821-7.

26. Eddy DM, et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Value Health. 2012;15(6):843-50.

27. Vemer P, et al. AdViSHE: a validation-assessment tool of health-economic models for decision makers and model users. Pharmacoeconomics. 2016;34(4):349-61.

28. Demler OV, et al. Tests of calibration and goodness-of-fit in the survival setting. Stats Med. 2015;34(10):1659-80.

29. Rahman MS, et al. Review and evaluation of performance measures for survival prediction models in external validation settings. BMC Med Res Method. 2017;17(1):60.

30. Uno H, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stats Med. 2011;30(10):1105-17.

31. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Stats Med. 2000;19(24):3401-15.

32. Sacks J, et al. Design and analysis of computer experiments. Statist Sci. 1989;4(4):409-23.

33. McKenna C, Claxton K. Addressing adoption and research design decisions simultaneously: the role of value of sample information analysis. Med Decis Making. 2011;31(6):853-65.

34. Incerti D, et al. A flexible open-source decision model for value assessment of biologic treatment for theumatoid arthritis. Pharmacoeconomics. 2019.

SUMMARY

SUMMARY

Many novel cancer therapies and biomarkers have been developed and adopted over the last two decades. Treatment landscapes in metastatic cancers now typically comprise different targeted and immunotherapies across multiple treatment lines, with several predictive and prognostic biomarkers available to stratify patients into subpopulations to maximize clinical efficacy and patient health outcomes. These advances have paved the way for personalized oncology, aiming to identify the best treatment for each patient individually based on specific patient and disease characteristics. Although these developments have contributed to improved health outcomes, they are also associated with considerable cost. These cost are not only monetary in terms of excessive cancer drug prices, for example, but also in terms of increased complexity regarding the selection and combination of therapies and biomarkers across multiple lines of treatment. Patient-level simulation modeling methods have potential to address these challenges, as simulation models are able to estimate the health and economic impact of different treatment strategies based on specific patient and disease characteristics. Despite their clear potential, several questions regarding the use of patient-level simulation modeling methods haven been raised. For example, in what scenarios should patient-level simulation methods be applied? How should these methods be implemented? And how can challenges regarding their use be addressed?

Aiming to make a substantial contribution to answering these questions, the objectives of this thesis were to review the current status of health economic modeling in personalized medicine (Part 1), provide guidance regarding the implementation of patient-level simulation modeling methods (Part 2), address challenges regarding the computational burden of simulation models (Part 3), and illustrate the potential of using simulation models to optimize personalized oncology pathways (Part 4). Although these aspects are applicable to all patient-level simulation modeling methods in general, focus in this thesis was on discrete event simulation.

As a starting point, Part 1 (Chapters 2, 3, and 4) of this thesis provided an overview of the current status of simulation modeling in personalized medicine, including comparisons of different simulation modeling methods.

The review presented in Chapter 2 showed that published health economic simulation models that were classified as personalized, only represent the complex dynamics of current clinical pathways to a limited extent. The use of patient-level simulation modeling methods to model these pathways also showed to be limited, as only 6 out of 31 identified models were developed using patient-level methods. A lack of incentives and guidance for developing personalized simulation models, as well as the complexity of implementing and understanding patient-level simulation models, were amongst reasons discussed to explain these observations.

Chapter 3 compared discrete event simulation to cohort-level discrete-time state-transition modeling in a case study on metastatic colorectal cancer treatment. This case study estimated the cost-effectiveness of capecitabine and bevacizumab maintenance treatment for

patients with stable disease or better after six cycles of capecitabine, oxaliplatin, and bevacizumab induction therapy, compared to an observation strategy. Although the discrete event simulation was found to provide slightly more accurate time-to-event predictions, incremental cost-effectiveness ratios were similar: €172,443 and €168,383 per quality adjusted life year gained for the state-transition model and discrete event simulation, respectively. This study additionally found that the discrete event simulation represented a patient flow more naturally, as the parametric distributions used to populate this model were less prone to irregularities and more straightforward to interpret compared to the time-dependent transition probabilities used to populate the state-transition model.

Chapter 4 illustrated how discrete event simulation and timed automata, a modeling paradigm new to the field of health economics, can go beyond traditional health economic simulation modeling methods and transparently model the health and economic impact of preference-sensitive treatment decisions. This case study investigated the potential impact of using circulating tumor cell analysis for informing treatment switching decisions in metastatic castration-resistant prostate cancer, compared to current practice in which prostate-specific antigen and bone scans are used as response markers. It was estimated that first- and second-line overtreatment could be reduced by 6.99 and 7.02 weeks by applying circulating tumor cells as response marker, for the timed automata and discrete event simulation model respectively. At a willingness to pay of €50,000 per quality adjusted life year gained, this translated to estimated net monetary benefits of €1033 and €1104, respectively. Implementing interactions was more straightforward in the timed automata model, which benefitted from its compositional structure. However, there were several software specific issues to this modeling technique and, hence, discrete event simulation was considered to be the preferred method.

Aiming to address challenges regarding the implementation of patient-level simulation modeling methods, Part 2 (Chapters 5, 6, and 7) provided methodological and practical guidance for the application of these methods.

The study presented in Chapter 5 illustrated the importance of appropriately reflecting parameter uncertainty (i.e., second-order uncertainty) when parametric distributions are used to reflect stochastic uncertainty (i.e., first-order uncertainty) in patient-level simulation models. This study compared two approaches to reflect second-order uncertainty in distribution parameters based on a simulation study and a case study. An approach based on non-parametric bootstrapping was found to be preferred over an approach using multivariate Normal distributions. An important benefit of the non-parametric bootstrapping approach was that it maintains correlations not only between distribution parameters, but also between different distributions estimated based on the same dataset. Results from this study also highlighted that neglecting to adequately account for parameter uncertainty in patient-level simulation models, will result in underestimation of the total amount of uncertainty surrounding simulation outcomes and may result in cost-effectiveness point-estimates to be biased.

Chapters 6 and 7 provided methodological and practical guidance for implementing com-

peting risks in discrete event simulations, by illustrating and comparing different approaches to do so in a simulation study and a case study. Both studies found substantial differences between modeling approaches, both with regard to the approaches' performance as well as in health economic outcomes and the uncertainty surrounding these outcomes. More specifically, Chapter 6 investigated the implementation of competing risks in discrete event simulation models based on uncensored individual patient time-to-event data. This study showed that the preferred modeling approach is to sample the time-to-event first from a single multimodal distribution representing all competing events, and to subsequently sample the corresponding event using a (multinomial) logistic regression model. However, a more straightforward (to implement) modeling approach that uses event-specific probabilities and distributions to sample an event first and the time-to-event second, showed similar performance when sufficient observations were available. Chapter 7 demonstrated and compared modeling approaches based on censored time-to-event data. This study showed that the presence of censoring resulted in too much complexity to make a general recommendation, as different approaches showed to be preferred depending on the level of censoring and overlap between competing time-to-event distributions.

Part 3 of this thesis (Chapters 8, 9, and 10) focused on the use of metamodeling methods to address challenges regarding the computational burden of simulating individual patients, which has been an important barrier to the application of patient-level modeling methods.

Chapter 8 presented a review on the use of metamodeling methods in health economics, which only found 13 previous applications. The study showed that computational issues are applicable to both patient-level and cohort-level simulation models, as 7 out of 13 studies used cohort-level methods to develop the original health economic simulation model. This confirms that not only the model type itself, but also the analysis performed determines whether runtime issues may occur. Although the number of applications in health economics was found to be limited, the applications that were identified highlighted the potential of metamodeling methods to perform computationally demanding analyses, such as probabilistic sensitivity analyses and value of information analyses.

Chapter 9 presented a structured overview of metamodeling methods for use in health economics, including considerations regarding their selection and implementation, to provide health economic modelers with essential tools for applying these methods. All identified methods were structured in a comprehensive six-step process: 1) identification of suitable techniques, 2) data simulation, 3) metamodel fitting, 4) performance assessment, 5) analysis execution, and 6) results verification. This process does not only guides modelers in the selection and application of metamodeling methods, it also provides guidance on what should be communicated about when reporting metamodeling studies.

Chapter 10 presented a case study on colorectal cancer screening in the Netherlands, which further illustrated the potential of using metamodeling methods in health economics. In this study, a metamodel was used to address computational challenges with the validated Adenoma and Serrated pathway to Colorectal CAncer (ASCCA) model, to allow optimization of

the Dutch colorectal cancer screening program, while accounting for colonoscopy capacity constraints. Results from this study suggested that for a maximum of 550 colonoscopies lifelong per 1,000 individuals, which reflects current colonoscopy capacity in the Netherlands, the optimal screening strategy is to perform 21 biennial screening rounds, starting at age 33 years and using a fecal immunochemical test cut-off of 150 ng/ml. This strategy is expected to yield 0.086 life years and savings of €361 per individual compared to a scenario without screening. Although these results suggested screening should start at an earlier age compared to current practice, which is to start screening at age 55, strategies involving screening at relatively young ages were extrapolations beyond the data used to populate the ASCCA model and they should, therefore, be interpreted with care.

Part 4 (Chapter 11) answered the remaining question regarding the potential of simulation models to evaluate and optimize personalized oncology pathways.

Chapter 4 already illustrated that patient-level simulation methods can be used to model and evaluate biomarker-based, preference-sensitive treatment decisions. Additionally, Chapter 10 illustrated how simulation modeling methods combined with metamodeling methods can be used to optimize screening strategies while accounting for resource constraints. To further illustrate their potential, Chapter 11 applied simulation modeling methods based on real-world registry data to inform treatment targeting decisions in metastatic colorectal cancer. This study showed that health outcomes of patients could be substantially improved by selecting the treatment most likely to be effective in terms of progression-free survival, based on specific patient and disease characteristics. By targeting a different treatment for 219 patients, which represented 25% of the complete cohort, median progression-free survival for these patients was estimated to improve significantly from 175 days to 269 days. For the complete patient cohort, this resulted in an estimated improved median first-line progression-free survival of 288 days compared to 265 days as observed from the data. Although multivariable (survival) models were used to account for selection bias, these promising results obtained based on observational data should be interpreted with care.

This thesis has illustrated the potential of patient-level simulation modeling methods to address challenges regarding the economic burden and complexity of personalized oncology. It contributed methodological and practical guidance on the application of these methods, providing modelers with essential tools to adequately evaluate and optimize personalized oncology pathways. The case studies performed, provided insights that would have been very time and resource consuming to obtain by performing (additional) randomized studies, if feasible at all. Although it was illustrated that simulation methods can be used to identify strategies that improve health outcomes while making efficient use of scarce health care resources, further multidisciplinary research is needed to translate this conceptual and methodological work into practice. These efforts may identify for what type of research questions and under which circumstances patient-level simulation modeling methods are most valuable, how confidence in simulation models can be established, and which outcomes are key in optimizing personalized oncology pathways.

SAMENVATTING

Dutch summary

SAMENVATTING

De afgelopen twee decennia zijn er veel nieuwe kankerbehandelingen en biologische markers (biomarkers) ontwikkeld en in gebruik genomen. Voor de meeste gemetastaseerde kankersoorten zijn er nu meerdere behandelopties in verschillende lijnen van behandeling. Ook zijn er vaak meerdere voorspellende biomarkers beschikbaar om patiënten op te delen in subpopulaties waarvoor de effectiviteit en uitkomsten van een specifieke behandeling het beste zijn. Deze ontwikkelingen hebben geleid tot meer gepersonaliseerde processen voor de diagnose en behandeling van kanker (personalized oncology) met als doel om patiënten te voorzien van de best mogelijke zorg op basis van specifieke patiënt- en ziektekarakteristieken. Ondanks dat dit alles heeft bijgedragen aan verbeterde behandeluitkomsten, zijn deze ontwikkelingen ook gepaard gegaan met belangrijke vraagstukken en uitdagingen. Niet alleen de kosten van bijvoorbeeld dure kankermedicijnen zijn een probleem, ook de behandelbeslissingen zijn erg complex geworden door het grote aantal behandelopties en biomarkers die in verschillende lijnen van behandeling kunnen worden gebruikt. Simulatiemodellen op patiëntniveau (patient-level) kunnen mogelijk worden gebruikt om inzichten te geven in de kosten en complexiteit van kankerbehandelingen, doordat deze modellen de economische- en gezondheidsgevolgen van verschillende behandelstrategieën in kunnen schatten. Hoewel het gebruik van patient-level simulatiemodellen veelbelovend is, zijn er verschillende vragen die beantwoord moeten worden rondom het gebruik van deze modellen. Wanneer moeten deze patient-level simulatiemodellen bijvoorbeeld worden gebruikt? Hoe moeten deze modellen worden ontwikkeld? En hoe kunnen problemen, die gepaard gaan met het ontwikkelen van deze modellen, worden opgelost?

Met als ambitie een substantiële bijdrage te leveren aan het beantwoorden van deze vragen, waren de doelen van dit proefschrift om: de huidige stand van zaken rondom het modelleren van gepersonaliseerde strategieën in kaart te brengen (Deel 1), aanbevelingen te doen voor het ontwikkelen van patient-level simulatiemodellen (Deel 2), problemen met betrekking tot de duur van het analyseren van simulatiemodellen op te lossen (Deel 3) en de mogelijkheden te illustreren van simulatiemodellen om oncologische behandelstrategieën te optimaliseren (Deel 4). Alhoewel deze aspecten in principe van toepassing zijn op alle patient-level simulatie methodes, lag de focus in dit proefschrift op “discrete event simulation” als methode.

Het eerste deel van dit proefschrift (Hoofdstukken 2, 3 en 4) gaf een overzicht van de huidige stand van zaken rondom het modelleren van gepersonaliseerde behandelstrategieën, inclusief vergelijkingen van verschillende simulatie methodes.

Het literatuuronderzoek gepresenteerd in Hoofdstuk 2 liet zien dat gepubliceerde gezondheids-economische simulatiemodellen slechts in beperkte mate de relevante aspecten van gepersonaliseerde behandelstrategieën weergeven. Ook het gebruik van patient-level simulatie methodes bleek beperkt, aangezien slechts 6 van de 31 gevonden studies patient-level methodes had gebruikt. Een gebrek aan stimulans en richtlijnen voor het ontwikkelen van patient-level simulatiemodellen zijn onder andere aangedragen als mogelijke oorzaken hiervan.

Hoofdstuk 3 vergeleek discrete event simulation met ‘’cohort-level discrete-time state-tran-

sition modeling'' in een casus over de behandeling van gemetastaseerde darmkanker. Deze studie keek naar de kosteneffectiviteit van voortgezette behandeling met capecitabine en bevacizumab voor patiënten met op zijn minst een stabiele ziekte na zes behandelcyclussen met capecitabine, oxaliplatin en bevacizumab, in vergelijking met een observatie strategie. Ondanks dat het discrete event simulation model de timing van gebeurtenissen iets beter wist te voorspellen, waren er geen grote verschillen in kosteneffectiviteit. Namelijk €172,443 en €168,383 per voor kwaliteit gecorrigeerd gewonnen levensjaar (quality adjusted life year, QALY) voor respectievelijk het state-transition en discrete event simulation model. Deze studie liet verder zien dat discrete event simulation het behandelproces op een meer natuurlijke wijze weergaf, doordat de statistische verdelingen die in dit model werden gebruikt robuuster en beter te interpreteren waren dan de tijdsafhankelijk transitiekansen die in het state-transition model werden gebruikt.

Hoofdstuk 4 liet zien hoe discrete event simulation en "timed automata", een nieuwe methode in deze context, de gezondheids-economische uitkomsten van voorkeursgevoelige behandelbeslissingen kunnen inschatten. De gebruikte casus onderzocht het gebruik van ''circulating tumor cells analysis'' om strategieën voor de behandeling van gemetastaseerde prostaat kanker te informeren. Het analyseren van circulating tumor cells werd daarin vergeleken met het gebruik van het prostaat specifieke antigeen en botscans. De resultaten lieten zien dat overbehandeling kan worden verminderd met naar schatting 6.99 of 7.02 weken als circulating tumor cells analysis zou worden gebruikt, voor respectievelijk het timed automata en discrete event simulation model. Indien men bereid is om €50.000 per QALY te betalen, vertaalt dit zich naar een ''net monetary benefit'' van respectievelijk €1033 of €1104. Vanwege de modulaire structuur van het timed automata model, konden interactief hierin eenvoudiger worden gemodelleerd. Desalniettemin waren er meerdere specifieke software problemen met deze methode, waardoor discrete event simulation de voorkeur kreeg.

Het tweede deel van dit proefschrift (Hoofdstukken 5, 6 en 7) voorzag in aanbevelingen voor het het ontwikkelen van patient-level simulatiemodellen.

Hoofstuk 5 illustreerde het belang van het kwantificeren van de onzekerheid rondom modeluitkomsten ontstaande uit de informatie die wordt gebruikt om modellen te ontwikkelen (parameter uncertainty), wanneer statistische verdelingen worden gebruikt om op patiëntniveau variatie te simuleren (stochastic uncertainty). Deze studie vergeleek twee methodes om deze onzekerheid correct weer te geven op basis van een gesimuleerde studie en op basis van een casus. Een methode gebaseerd op "non-parametric bootstrapping" bleek de voorkeur te hebben boven een methode die gebruik maakt van een multivariate Normaalverdeling. Een belangrijk voordeel van de non-parametric bootstrapping methode was dat deze niet alleen de correlatie tussen verdelingsparameters in stand hield, maar ook de correlatie tussen verdelingen. De resultaten lieten verder zien hoe belangrijk het is om deze onzekerheid correct te kwantificeren. Als dat niet wordt gedaan, wordt de totale hoeveelheid onzekerheid onderschat en kunnen puntschattingen van uitkomsten afwijken.

Hoofdstukken 6 en 7 deden aanbevelingen voor het modelleren van gebeurtenissen in dis-

crete event simulations, door het illustreren en vergelijken van verschillende methodes. Beide studies lieten zien dat de gebruikte methode niet alleen de nauwkeurigheid van het model bepalen, maar daarmee ook invloed hebben op de uitkomsten en de onzekerheid in die uitkomsten. Hoofdstuk 6 keek naar het modelleren van gebeurtenissen op basis van data waarin voor iedere patiënt bekend is welke gebeurtenis heeft plaatsgevonden en wanneer dat was (uncensored data). Deze studie liet zien dat de meest nauwkeurige methode is om eerst het tijdstip te bepalen aan de hand van één statistische verdeling, om vervolgens te bepalen welke gebeurtenis daarbij hoort met behulp van een (multinomiaal) logistisch regressie model. Een eenvoudiger toe te passen methode, waarbij er voor elke gebeurtenis een aparte kans en verdeling wordt gebruikt, bleek echter vergelijkbaar te presteren indien er voldoende observaties waren. Hoofdstuk 7 keek naar het modelleren van gebeurtenissen wanneer niet voor iedere patiënt bekend is of een gebeurtenis heeft plaatsgevonden en wanneer dat was (censored data). Deze studie liet zien dat de complexiteit die ontstaat door het missen van observaties ervoor zorgt dat een algemene aanbeveling, zoals in Hoofdstuk 6 kon worden gedaan, niet kon worden gedaan voor censored data. Afhankelijk van de hoeveelheid missende data en de mate waarin de verschillende gebeurtenissen rondom dezelfde tijdsperiode plaatsvinden, bleken verschillende methodes de voorkeur te hebben.

Het derde deel (Hoofdstukken 8, 9 en 10) heeft zich gericht op het gebruik van metamodellen om problemen met de duur van het analyseren van simulatiemodellen op te lossen.

Hoofdstuk 8 presenteerde een literatuuronderzoek naar het gebruik van metamodeling methodes binnen de gezondheidseconomie en vond slechts 13 toepassingen. Deze studie liet zien dat problemen met betrekking tot de duur van het analyseren van simulatiemodellen van toepassing waren op zowel patient-level als cohort-level simulatiemodellen. Dit suggereerde dat niet alleen het modeltype, maar ook de soort analyse bepaalt of er problemen met betrekking tot de duur van analyses kunnen optreden. Ondanks dat het aantal toepassingen van metamodeling methodes beperkt was, lieten de studies die deze methodes wel hebben gebruikt, zien dat dit soort methodes succesvol kunnen worden toegepast om geavanceerde analyses uit te voeren welke anders niet binnen een realistisch tijdsbestek zouden kunnen zijn uitgevoerd. Een voorbeeld van dit soort analyses zijn “value of information” analyses.

Hoofdstuk 9 presenteerde een gestructureerd overzicht van metamodeling methodes die geschikt zijn voor gebruik binnen de gezondheidseconomie, inclusief richtlijnen voor het selecteren en implementeren van deze methodes. Alle methodes zijn gestructureerd in een proces van zes stappen: 1) geschikte technieken identificeren, 2) data simuleren om te trainen en testen, 3) metamodel trainen, 4) metamodel testen, 5) analyses uitvoeren en 6) resultaten verifiëren. Dit proces voor het toepassen van metamodeling methodes biedt niet alleen handvatten voor de selectie en het gebruik van methodes. Het suggereert ook waarover gecommuniceerd dient te worden wanneer studies, die gebruik maken van metamodeling methodes, worden gepubliceerd.

Hoofdstuk 10 presenteerde een casus over darmkankerscreening in Nederland, welke verder illustreerde hoe metamodeling methodes kunnen worden gebruikt binnen de gezondheid-

seconomie. Een metamodel werd gebruikt om het langzame gevalideerde Adenoma and Serrated pathway to Colorectal CAncer (ASCCA) model snel te benaderen. Hierdoor kon het Nederlandse darmkankerscreeningprogramma worden geoptimaliseerd, rekening houdend met de beschikbare colonoscopiecapaciteit. Resultaten van deze studie suggereerden dat, voor de huidige colonoscopiecapaciteit van 550 over het gehele leven van 1000 individuen, de beste strategie is om vanaf 33 jarige leeftijd 21 tweejaarlijkse screening rondes uit te voeren met een drempelwaarde van 150 ng/ml voor de gebruikte fecal immunochemical test. Deze strategie levert naar verwachting een winst van 0.086 levensjaren en een kostenbesparing van €361 op per individu, in vergelijking tot een scenario zonder screening. Deze resultaten suggereerden dat screening eerder zou moeten beginnen dan de huidige startleeftijd van 55 jaar. Dit soort strategieën zijn echter extrapolaties van de data waarop het ASCCA model is ontwikkeld en men moet daarom voorzichtig zijn met het trekken van conclusies.

Het vierde deel (Hoofdstuk 11) gaf antwoord op het overgebleven vraagstuk rondom het gebruik van patient-level simulatiemodellen om behandelstrategieën te optimaliseren.

Hoofstuk 4 liet al zien dat patient-level simulatiemodellen kunnen worden gebruikt om gepersonaliseerde behandelbeslissingen te modelleren. Hoofdstuk 10 liet zien hoe simulatie en metamodeling methodes kunnen worden gebruikt om strategieën te optimaliseren, rekening houdende met bepaalde restricties. Om de mogelijkheden van patient-level simulatiemodellen verder te illustreren, werd in Hoofdstuk 11 een simulatiemodel ontwikkeld op basis van real-world data om behandelbeslissingen voor gemetastaseerde darmkanker te informeren. Deze studie liet zien dat uitkomsten konden worden verbeterd door de behandeling te selecteren die op basis van patiënt- en ziektekarakteristieken als meest succesvol werd ingeschat. Door voor 219 patiënten (25%) een andere behandeling te kiezen, kon de mediane progressie-vrije overleving naar scharring significant worden verbeterd van 175 dagen naar 269 dagen. Voor het complete cohort van patiënten resulteerde dit in een verbetering in de mediane progressie-vrije overleving van 265 dagen naar 288 dagen. Ondanks dat multivariable (survival) modellen werden gebruikt, konden er geen definitieve conclusies worden getrokken vanwege de mogelijke selectiebias in de gebruikte observationele data.

Dit proefschrift heeft de mogelijkheden geïllustreerd van patient-level simulatiemodellen om uitdagingen rondom de kosten en complexiteit van personalized oncology aan te gaan. Door het bijdragen van aanbevelingen voor het ontwikkelen van dit soort simulatiemodellen, hebben ontwikkelaars belangrijke bouwstenen hebben gekregen om behandelstrategieën te evalueren en optimaliseren. De behandelde casussen hebben geresulteerd in inzichten die waarschijnlijk veel meer geld en tijd hadden gevergd wanneer deze via klinische studies hadden moeten worden verkregen, als dit laatste al mogelijk was geweest. Ondanks dat er duidelijk mogelijkheden zijn voor simulatiemodellen om behandelstrategieën te informeren, is verder multidisciplinair onderzoek nodig om deze modellen vanuit hun methodologische omgeving te vertalen naar de klinische praktijk. Er zal gekeken moeten worden naar voor welk type vraagstukken en onder welke omstandigheden patient-level simulatiemodellen het meest waardevol zijn, hoe vertrouwen in deze modellen kan worden gevestigd en welke uitkomsten belangrijk zijn in het optimaliseren van gepersonaliseerde behandelstrategieën.

ACKNOWLEDGEMENTS

Dankwoord

ACKNOWLEDGEMENTS

Toen ik in 2015 tijdens het schrijven van mijn afstudeerscriptie op de Health Technology and Services Research vakgroep een promotieplek aangeboden kreeg, was ik in eerste instantie niet louter enthousiast. Het beeld dat ik van promoveren had bestond vooral uit veel literatuur lezen en achter een bureau zitten. Na bijna vier jaar realiseer ik me dat ik inderdaad veel achter een bureau heb gezeten en veel literatuur heb gelezen, maar ook dat ik mij geen andere “eerste baan” kan voorstellen waarbij ik zoveel van de wereld had kunnen zien en mijzelf tegelijkertijd persoonlijk en professioneel zo had kunnen ontwikkelen. Kijkend naar mijn proefschrift en terugdenkend aan de ervaringen die ik heb opgedaan, ben ik erg blij dat ik deze uitdaging ben aangegaan en ik heb geprobeerd er zoveel mogelijk uit te halen.

Maarten, jij was diegene die mij de promotieplek aanbood en herhaaldelijk aangaf welke kansen daarmee gemoeid konden gaan. Hiervoor ben ik je tot op de dag van vandaag erg dankbaar. Ook ben ik dankbaar voor jouw vertrouwen in mij en alle kansen die jij mij hebt gegeven. Met jouw kritische blik en kennis over wat er op nationaal en internationaal niveau speelt, heb jij ervoor gezorgd dat naast de methodologische kant van het onderzoek ook de andere perspectieven voldoende werden belicht. Bedankt voor de gesprekken over verschillende carrière perspectieven en de inspiratie om mezelf, ook buiten het doen van onderzoek, te ontwikkelen. Mijn research visit aan de University of Melbourne is voor mij een absoluut hoogtepunt geweest en ik kijk er dan ook naar uit om de komende jaren deel uit te maken van de Cancer Health Services Research Unit in Melbourne.

Erik, jouw expertise en kritische blik hebben een cruciale rol gespeeld in de totstandkoming van dit proefschrift. Ondanks dat onze discussies achteraf gezien soms best korter hadden kunnen zijn, hebben deze er wel voor gezorgd dat ik me op methodologisch vlak en schrijf technisch verder heb kunnen ontwikkelen dan ik aanvankelijk had gedacht. Bedankt voor de kansen die je me hebt gegeven om samen onderwijs te ontwikkelen en te geven. Een mooi resultaat van onze samenwerking was de workshop over discrete event simulation die we samen hebben opgezet en gepresenteerd in Glasgow. Ook jouw gevoel voor humor en onze gesprekken over het promoveren op zich heb ik erg gewaardeerd. Ik kijk er naar uit onze samenwerking voort te zetten en elkaar te treffen op congressen onder het genot van een speciaal biertje.

Maarten en Erik, voor mij zijn jullie het perfecte supervisieteam geweest. Jullie hebben er altijd voor gezorgd dat ik me voornamelijk kon focussen op mijn werk en me geen zorgen hoefde te maken over randzaken. Bedankt daarvoor.

Dear members of the doctoral commission, thank you for reading and approving this thesis, and for being part of the defense ceremony.

Dear coauthors and collaborators, thank you for all interesting projects we have completed, your timely and constructive feedback, and your willingness and patience to explain the different clinical contexts. I very much enjoyed working with all of you and look forward to further joint research projects.

Sofie en Annelies, ik ben jullie als paranimfen erg dankbaar voor alle gezelligheid en support. Sofie, samen hebben we over de hele wereld over gereisd, veel borrels gedronken, initiatieven ondernomen, mijlpalen bereikt en overwinningen gevierd. Bedankt dat je mij bij dat laatste altijd direct weer (ongevraagd) met beide benen op de grond zette. Nog meer wil ik je bedanken voor alle gesprekken en jouw humor, welke er niet alleen voor hebben gezorgd dat ik meermaals met tranen van het lachen achter mijn bureau of aan de lunchtafel heb gezeten, maar ook dat ik met erg veel plezier terugkijk op mijn tijd in Raffa 5155.

Annelies, toen ik begin augustus 2015 mijn eerste werkdag op de vakgroep had, zorgde jij ervoor dat ik me vanaf dat moment direct welkom en onderdeel van het team voelde. Als steun en toeverlaat op het secretariaat heb je ervoor gezorgd dat ik snel mijn weg kon vinden binnen de universiteit. Ondanks dat we niet altijd directe collega's zijn gebleven, ben ik je erg dankbaar voor je onverminderde betrokkenheid. Bedankt voor alle gezelligheid, lachbuien en je support als paranimf.

Ook mijn overige collega's binnen de vakgroep wil ik graag bedanken. Niet alleen voor de gezelligheid tijdens de uitjes en congressen, maar vooral ook voor de koffie- en lunchpauzes, waar ik altijd even stoom af kon blazen. Jorien, Sarah en Annemieke, bedankt voor het wegwijs maken toen ik begon als promovendus. Ingrid en Maaike, bedankt voor alle ondersteuning en dat jullie me door de bureaucratie van de universiteit hebben geloodst. Michelle en Marieke, bedankt voor alle gezelligheid, goede gesprekken en uiteraard ook voor de baksels en andere lekkernijen. Deon, thanks for all the good times, your kindness, and for putting up with all my nonsense.

I would also like to express my gratitude towards the University of Melbourne colleagues that I met during my research visit. Amanda, thanks for all the great serious and fun conversations. Thanks Jay, Camille, Josh, and Jesse, you made me feel very welcome, and my stay unforgettable.

Also, I would like to thank my kind and involved colleagues at the PhD Network of the University of Twente (P-NUT), who made my time as a PhD candidate at the University of Twente unforgettable. Koen, bedankt voor de mooie ervaringen die we samen hebben gedeeld, de goede gesprekken en dat je me hebt laten genieten van je kookkunsten en brouw skills.

Special thanks to all my colleagues and fellow students at the Professional Society of Health Economics and Outcomes Research (ISPOR), who made my time as Chapter President, Committee Co-chair and Network Chair an incredibly valuable and enjoyable experience. Jason, thank you for our close collaboration and your support. Blythe, thank you for being such a great inspiration. Zeba, thank you for being a great mentor. ISPOR Students, Committee Members and Committee Co-chairs, thank you for making my time as the ISPOR Student Network Chair an unforgettable experience.

Mijn vrienden hebben ervoor gezorgd dat ik tijdens mijn promotie meer dan genoeg afleiding heb gekregen. Ik wil iedereen dan ook graag bedanken voor alle mooie momenten die we

samen hebben meegemaakt, maar ook voor jullie interesse in mijn werk, ondanks dat we niet altijd over de inhoud konden praten. Alhoewel er velen zijn die mij hebben gesteund en met wie ik geweldige herinneringen heb mogen maken, zijn er drie personen die voor mij in deze periode in het bijzonder belangrijk zijn geweest. Deck, bedankt voor je support en dat je altijd voor de nodige afleiding hebt gezorgd. Steen, bedankt voor alle goede gesprekken en potjes FIFA onder het genot van een “glaasie”. Noordam, bedankt voor je interesse, de goede gesprekken en alle gezelligheid.

Ook mijn familie heeft voor mij een belangrijke rol gespeeld. Pap en mam, bedankt dat jullie ervoor hebben gezorgd dat ik me altijd heb kunnen ontwikkelen zonder dat de nadruk ooit heeft gelegen op presteren, maar altijd op doen waar ik gelukkig van word. Jullie hebben me altijd gestimuleerd om kansen te pakken, ook al hield dit in dat ik voor een korte of langere periode naar het buitenland moest, en ook jullie trots heeft mij erg gemotiveerd. Annewil en Femke, bedankt voor jullie support en de gezellige etentjes en familiale aangelegenheden. Pap, mam, Annewil en Femke, bedankt dat jullie voortdurend interesse zijn blijven tonen.

Vivian, bedankt voor jouw oneindige steun, waardoor geen uitdaging mij te groot was, en voor jouw begrip. Ondanks dat jij bijvoorbeeld terecht vroeg of het gezien mijn ambities op het werk wel verstandig was om er een vierde bestuursfunctie bij te nemen, heb jij me altijd gemotiveerd om kansen te benutten. En natuurlijk bedankt voor het checken van de Nederlandse teksten in dit proefschrift. Ik ben blij dat we mede door mijn promotie de afgelopen jaren veel mooie herinneringen hebben kunnen maken en kijk er naar uit om er daar samen nog velen aan toe te voegen.

CURRICULUM VITAE

CURRICULUM VITAE

Koen Degeling was born on the 20th of March 1991 in Hengelo, the Netherlands. Directly after completing his secondary education at the SG Twickel (Hengelo) high school in 2009, he started as a bachelor student in the Mechanical Engineering program at the University of Twente in Enschede, the Netherlands. Soon he realized he was missing the human touch in this program and switched to the Industrial Engineering and Management program, which he completed in 2013. Subsequently, he pursued a master degree in Industrial Engineering and Management with a specialization in Health Care Technology and Management, which he was awarded Cum Laude in 2015. His master thesis "The Health Economic Impact of Circulating Tumour Cells Analysis in Metastatic Castration Resistant Prostate Cancer Treatment" was accepted as a podium presentation at the Professional Society for Health Economics and Outcomes Research (ISPOR) conference in Milan (2015), and awarded with the "Best Student Podium Presentation Award".

After obtaining his master degree, Koen started as a PhD Candidate within the Health Technology and Services Department at the University of Twente. Throughout his PhD project, his contributions to international conferences included a short course, workshop, and multiple podium and poster presentations. He has also been substantially involved as a board member in several local and international organizations, including the PhD Network of the University of Twente (P-NUT) and ISPOR. After founding the ISPOR University of Twente Student Chapter, which was awarded "Best New Student Chapter 2017-2018", and Co-chairing the ISPOR Student Network Chapter Development Committee for the 2017-2018 term, he received a "Distinguished Service Award" for his contributions to the society in 2018 and in 2019. His absolute highlight within ISPOR has been chairing the global ISPOR Student Network for the 2018-2019 term. With regard to research, his absolute highlight was being awarded a personal travel grant by the Netherlands Organisation for Health Research and Development (ZonMw) for a four-month research visit to the University of Melbourne, Australia. This research visit enabled him to translate his research on the use of simulation modeling methods to evaluate and optimize personalized oncology into practice, by performing a case study on treatment targeting in metastatic colorectal cancer based on real-world data. He received his second "Best Student Podium Presentation Award" for presenting this research at the ISPOR conference in New Orleans (2019).

Koen will continue his academic career at the University of Melbourne in Australia. Although he will continue to develop methodological and practical guidance for applying patient-level simulation modeling methods, his primary focus will be on addressing challenges regarding the sustainability and complexity of personalized oncology using simulation models.

PUBLICATIONS

PUBLICATIONS

Published papers

Degeling K, IJzerman MJ, Koffijberg H. A scoping review of metamodeling applications and opportunities for advanced health economic analyses. Expert Review of Pharmacoeconomics & Outcomes Research. 2019;19(2):181-7

Degeling K, Koffijberg H, Franken MD, Koopman M, IJzerman MJ. Comparing strategies for modeling competing risks in discrete-event simulations: a simulation study and illustration in colorectal cancer. Medical Decision Making. 2019;39(1):57-73

Degeling K, Franken MD, May AM, van Oijen MGH, Koopman M, Punt CJA, IJzerman MJ, Koffijberg H. Matching the model with the evidence: comparing discrete event simulation and state-transition modeling for time-to-event predictions in a cost-effectiveness analysis of treatment in metastatic colorectal cancer patients. Cancer Epidemiology. 2018;57:60-7

Degeling K, IJzerman MJ, Koopman M, Koffijberg H. Accounting for parameter uncertainty in the definition of parametric distributions used to describe individual patient variation in health economic models. BMC medical research methodology. 2017;17(1):170

Degeling K, Schivo S, Mehra N, Koffijberg H, Langerak R, de Bono JS, IJzerman MJ. Comparison of timed automata with discrete event simulation for modeling of biomarker-based treatment decisions: an illustration for metastatic castration-resistant prostate cancer. Value in Health. 2017;20(10):1411-9

Degeling K, Koffijberg H, IJzerman MJ. A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models. Expert Review of Pharmacoeconomics & Outcomes Research. 2017;17(1):17-25

Submitted papers

Degeling K, IJzerman MJ, Lavieri MS, Strong M, Koffijberg H. How to use metamodeling for reducing computational burden of advanced analyses with health economic models: a structured overview of metamodeling methods in a six-step application process. *Submitted.*

Degeling K, IJzerman MJ, Groothuis-Oudshoorn CGM, Franken MD, Koopman M, Koffijberg H. Comparing modeling approach for discrete event simulations of competing events based on censored individual patient data: a simulation study and illustration in colorectal cancer.

Koffijberg H, Degeling K, IJzerman MJ, Coupe VMH, Greuter MJE. From evaluation to optimization of diagnostic strategies: maximizing the benefits of colorectal screening using a metamodel.

Degeling K, Wong H-L, Koffijberg H, Jalali A, Shapiro JD, Kosmider S, Wong R, Lee B, Burge ME, Tie J, Yip D, Nott LM, Khattak A, Lim SH-S, Caird S, Gibbs P, IJzerman MJ. Simulating progression-free and overall survival for first-line doublet chemotherapy with or without bevacizumab in metastatic colorectal cancer patients based on real world registry data.

CONFERENCE CONTRIBUTIONS

Short courses

Marshall DA, Tosh J, Degeling K, Fenwick E. Fitting the structure to the task: choosing the right dynamic simulation model to inform decisions about health care delivery. International Society for Health Economics and Outcomes Research, ISPOR 2018 Meeting, Baltimore, 2018

Workshops

Fenwick E, Koffijberg H, Jahn B, Degeling K. Why is your outcome different from mine? Discussing the potential impact of design choices in the development of discrete event simulation models. International Society for Pharmacoeconomics and Outcomes Research, 20th Annual European Congress Glasgow, 2017

Podium presentations

Wong H-L, Degeling K, Jalali A, Shapiro JD, Kosmider S, Wong R, Lee B, Burge ME, Tie J, Yip D, Nott LM, Khattak A, Lim SH-S, Caird S, IJzerman MJ, Gibbs P. Understanding variation in treatment sequences and outcomes in metastatic colorectal cancer: using real world data to answer real world questions. Professional Society for Health Economics and Outcomes Research, New Orleans, 2019

Degeling K, Koffijberg H, Schivo S, Langerak R, IJzerman MJ. Comparison of timed automata with discrete event simulation for modeling personalized treatment decisions: the case of metastatic castration resistant prostate cancer. International Society for Pharmacoeconomics and Outcomes Research, 18th Annual European Congress Milan, 2015

Poster presentations

Degeling K, Wong H-L, Johns J, Koffijberg H, Gibbs P, IJzerman MJ. Personalized discrete event simulation of progression-free survival and overall survival for first-line doublet chemotherapy with or without bevacizumab. Professional Society for Health Economics and Outcomes Research, New Orleans, 2019

Degeling K, IJzerman MJ, Lavieri MS, Strong M, Koffijberg H. An overview of metamodeling methods to address runtime Issues associated with advanced analyses of health economic models. Professional Society for Health Economics and Outcomes Research, 21th Annual European Congress Barcelona, 2018

Degeling K, IJzerman MJ, Koffijberg H. Scoping review of metamodeling applications and opportunities for advanced analyses with health economic models. Professional Society for Health Economics and Outcomes Research, 21th Annual European Congress Barcelona, 2018

Degeling K, Mehra N, Koffijberg H, de Bono JS, IJzerman MJ. Physicians' behavior influences the health and economic impact of applying circulating tumor cells as response marker in metastatic castration-resistant prostate cancer. International Society for Pharmacoeconomics and Outcomes Research, 19th Annual European Congress Vienna, 2016

Degeling K, Koffijberg H, IJzerman MJ. A review on the current status of health economic modeling in personalized medicine. International Society for Pharmacoeconomics and Outcomes Research, 19th Annual European Congress Vienna, 2016

Koffijberg H, Degeling K, Janssen MP, IJzerman MJ. Reflecting uncertainty in early HTA when evidence on parameter distributions is lacking: a comparison of methods. International Society for Pharmacoeconomics and Outcomes Research, 19th Annual European Congress Vienna, 2016

Schivo S, Degeling K, Koffijberg H, IJzerman MJ, Langerak R. Timed automata modeling of the personalized treatment decisions in metastatic castration-resistant prostate cancer. Society of Medical Decision Making, 37th Annual North American Meeting St. Louis, 2015

Degeling K, Koffijberg H, Schivo S, Langerak R, IJzerman MJ. Modeling personalized treatment decisions: comparison of timed automata with discrete event simulation. Society of Medical Decision Making, 37th Annual North American Meeting St. Louis, 2015

Schivo S, Degeling K, Koffijberg H, IJzerman MJ, Langerak R. Timed automata modeling of the personalized treatment decisions in metastatic castration-resistant prostate cancer. International Society for Pharmacoeconomics and Outcomes Research, 18th Annual European Congress Milan, 2015