technische universität
dortmund

Faculty of Statistics

Statistical topics in clinical biosimilar development

Johanna Isabel Mielke

Dissertation

submitted to the Department of Statistics at TU Dortmund University in partial fulfilment of the requirements for the degree *Doktor der Naturwissenschaften*

Advisors and referees: Prof. Dr. Joachim Kunert, Prof. Dr. Byron Jones
Referee: Prof. Dr. Joerg Rahnenfuehrer
Chair of the thesis committee: Prof. Dr. Guido Knapp
Date of defense: 29.06.2018

Acknowledgment

I am grateful to many people who have, through their support and encouragement, supported me along the journey of my PhD. First, I would like to thank my advisor Byron Jones for his continuous support and guidance during the last three years. Your enthusiasm and reliability and the optimal mixture of leaving me enough freedom to follow my own research ideas while providing support at critical decision points were the key to finish this dissertation in less than three years! Thanks also to the Statistical Methodology & Consulting group at Novartis in Basel for making my time as a PhD student so enjoyable.

I am also grateful to Joachim Kunert for being my internal advisor at TU Dortmund, for giving advice on optimal design theory and hosting me during a very productive research visit in Bozen. Furthermore thanks to Joerg Rahnenfuehrer for refereeing the thesis and to Guido Knapp for chairing the committee.

Many collaborators were involved in this dissertation: I would like to thank Franz Koenig and the Section of Medical Statistics at the Medical University of Vienna for hosting me during two secondments and for the successful collaboration. Thanks to Bernd Jilma for giving insights into medical topics and helping with the preparation of the manuscripts for a clinical audience. I am also grateful to Heike Woehling and her team at Sandoz in Oberhaching for supporting me during my secondment and for advice on practical considerations of biosimilar development. Thanks also to Heinz Schmidli for helping me with my first steps into Bayesian statistics.

My PhD research was part of the IDEAS European training network and this network made my experiences as a PhD student really extraordinary. I would like to thank all the Early Stage Researchers for travelling with me to conferences and summer schools during the last three years. Marius, thanks for sharing the office with me and for providing

constant supply with fresh tea! Special thanks also to Julia and Nico for spending time with me in Vienna and to Haiyan, Enya, Saswati, Arsenio, Nico and Jose for visiting me in Basel. In addition, thanks to everyone else involved in the training network for giving feedback on my research and for the perfect organisation of the summer schools.

Last, many thanks to my family and friends for unconditional support and to Stefan for being willing to spend so much time in trains so that I could follow my dreams.

This dissertation was supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 999754557. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Swiss Government. The project is part of the IDEAS European training network from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567.

List of contributed papers

Kunert, J. and Mielke, J. (2018): Efficient designs for the estimation of mixed and self carryover effects. *SFB 823, Discussion paper,* 18 (8). DOI: 10.17877/DE290R-18820.

Mielke, J., Jilma, B., Jones, B. and Koenig, F. (2018a): An update on the clinical evidence that supports biosimilar approvals in Europe. *British Journal of Clinical Pharmacology*, 84 (7), 1415–1431.

Mielke, J., Jilma, B., Koenig, F. and Jones, B. (2016): Clinical trials for authorized biosimilars in the European Union: a systematic review. *British Journal of Clinical Pharmacology*, 82 (6), 1444–1457.

Mielke, J., Jones, B., Jilma, B. and König, F. (2018b): Sample size for multiple hypothesis testing in biosimilar development. *Statistics in Biopharmaceutical Research*, 10 (1), 39–49.

Mielke, J. and Kunert, J. (2018): Universally optimal crossover designs for the estimation of mixed-carryover effects with an application to biosimilar development. *SFB 823, Discussion paper*, 18 (3). DOI: 10.17877/DE290R-18786.

Mielke, J., Schmidli, H. and Jones, B. (2018c): Incorporating historical information in biosimilar trials: challenges and a hybrid Bayesian-frequentist approach. *Biometrical Journal*, 60 (3), 564–582.

Mielke, J., Woehling, H. and Jones, B. (2018d): Longitudinal assessment of the impact of multiple switches between a biosimilar and its reference product on efficacy parameters. *Pharmaceutical Statistics*, 17 (3), 231–247.

Contents

Chapter 1

Introduction

Innovative medicines are patent protected for a limited period of time. After the medicine goes off patent, any pharmaceutical company may produce their own version of the drug and apply to a health authority for market authorisation. This concept is very well established in the context of small molecule drugs (generics), however, it was introduced much later for biologics. A biologic is formally defined by the European Medicines Agency (EMA) as a "medicine that contains one or more active substances made by or derived from a biological source" (EMA, 2012a). Biologics have revolutionised the treatment in important disease areas like cancer or diabetes. However, since biologics are very expensive, the access of patients to these innovative treatment options is often limited, especially in low-income countries (Putrik et al., 2014). The hope that more competition on the market will lower these high prices, combined with the fact that the patents of several biological blockbusters (e.g., etanercept, infliximab, adalimumab) have expired in the last few years, makes the development of copies of biologics, so-called biosimilars, a topic of high interest both for the pharmaceutical industry and the general public.

So far, there exists no unified definition for biosimilars worldwide, but the way of thinking is comparable in the highly-regulated markets (e.g., European Union (EU), United States (US)). The EMA states that (EMA, 2012a) a "biosimilar medicine is a biological medicine that is developed to be similar to an existing biological medicine (the 'reference medicine'). [...] When approved, its variability and any differences between it and its reference medicine will have been shown not to affect safety or effectiveness." It is important to note that, while the overarching guideline on biosimilarity of the EMA is applicable to all types of biological products, all other guidelines published by the EMA refer to biotechnologically-derived proteins and, so far, the Food and Drug Administration (FDA), the regulatory agency in the US, considers biosimilarity for proteins only. Therefore, we only refer to biotechnologically-derived proteins when we mention biologics. However, the general concepts might also be applicable to other types of biologics.

It is the responsibility of the developer of the proposed biosimilar (also known as the sponsor) to convince the regulatory authority that the proposed product (also known as the test product) is *biosimilar*, i.e., it is required to show that a patient taking the biosimilar can expect the same treatment effect and safety profile as with the reference product. Since the main idea of biosimilars and generics is comparable, one might wonder if it is possible to use the well-established regulatory pathway for generics (EMA, 2012b) for showing biosimilarity. However, even though the main idea of biosimilars has some similarities to the concept of generics, there exist fundamental differences between small molecule drugs and biologics which are not only related to the product itself, but also to the manufacturing process. A brief overview of these differences, based on Crommelin et al. (2005), can be found in the following paragraph.

On the product side, small molecule drugs tend to have a well-defined and stable chemical structure which can be easily identified. Biologics, on the other hand, are more complex proteins with heterogeneous structures. The primary structure is a sequence of amino acids which is comparable to the structure of small molecule drugs. This structure is folded into structural elements which are stabilised by a secondary structure which is again folded into a three-dimensional tertiary structure. Many proteins are glycosylated and the pattern of the glycosylation, which depends, among others, on the condition under which the protein is produced, might impact the clinical outcome. Also, interactions with other molecules (e.g., cell-surface receptors, binding proteins and nucleic acids) influence the biological activity. While small molecule drugs can be fully characterised by their chemical structure, biological proteins can be so complex that current analytical methods cannot fully characterise them which makes the establishment of similarity based on chemical attributes very challenging. Also, small molecule drugs are chemically synthesised while biologics are produced in living cells or organisms. For the manufacturing of a biologic, a host cell (bacteria or eukaryotic) is created by inserting into it the chosen DNA sequences of the target protein. Afterwards, cell screening and selection processes are used for selecting a unique master cell bank. This cell bank differs between the production of batches, introducing a high amount of variability, even within different batches of the reference product. Then, the cells are grown on a large scale. Small changes at this stage (e.g., the physical conditions like temperature) might alter the protein. Since the cells also produce other substances apart from the target protein, it is next necessary to separate out the protein of interest. Finally, the protein is assessed with analytical methods for purity and potency. Even for the manufacturer of the reference product that possesses most knowledge about the product, the high complexity and the sensitive manufacturing process make it impossible to produce an exact copy of the biologic. That is why, in contrast to generics which are chemically identical to the original small molecule drug, biosimilars are only required to be *similar* to the reference product.

It is important to note that the differences present between small molecule drugs and biologics do not only complicate the development and production but make the characterisation of the molecules difficult. Also, the natural variability between batches of the reference product, and the fact that biosimilars are only similar, but not identical to the reference product, lead to a higher uncertainty as to whether the proposed biosimilar has the same efficacy and safety as the reference product. Therefore, in contrast to the approval process for generics which requires the showing of identical chemical properties and the confirmation that the concentration of the drug in the blood is equivalent after injection of the proposed generic or the original drug into healthy volunteers (Chow, 2013), this is not sufficient for getting approval as a biosimilar. Indeed, an extensive comparability exercise which requires analytical studies (comparison of chemical attributes of the molecules), non-clinical studies (in vitro studies and in vivo studies in animals) and clinical studies in healthy volunteers or patients is required. This implies that the resources and time required for bringing a biosimilar to the market are closer to the effort needed for an innovative drug than to that for a generic medicine (Blackstone and Fuhr Jr, 2012).

However, it is important to keep in mind that the objective of a biosimilar development programme is different to the objective of a development programme for an innovative drug: when an innovative product is developed, the existing knowledge about the molecule and its effect on humans is limited. Therefore, as part of the clinical development programme, it is, for example, necessary to determine safe and effective doses and treatment regimens. Sponsors might also aim to find additional indications in which the treatment is favourable or to identify subgroups which respond better or worse to the treatment (Friedman et al., 2015). For biosimilar development, all this information is already available from the research that was undertaken during the development of the reference product and the goal of the biosimilar development programme is to confirm that a patient who takes the biosimilar can expect the same efficacy and safety as if taking the reference product. This is also emphasised by regulators; for example, the EMA clearly states that (CHMP, 2014a) "efficacy trials of biosimilar medicinal products do not aim at demonstrating efficacy per se since this has already been established with the reference product. The purpose of the efficacy trials is to confirm comparable clinical performance of the biosimilar and the reference product." In summary, the regulatory pathway which is established for innovative products is not suitable for biosimilar development.

Since neither the approval pathway for generics nor the approval pathway for innovative drugs is applicable in biosimilar development, regulatory authorities have introduced a new approval pathway for biosimilars. We focus on the regulatory pathway and important concepts which are used in the EU, however, it should be emphasised that the approval pathway in the US shares the same main concepts. For getting approval as a biosimilar in the EU, "similarity to the reference medicinal product in terms of quality characteristics, biological activity, safety and efficacy based on a comprehensive comparability exercise needs to be established" (CHMP, 2014c). The EMA recommends a step-wise approach that consists of quality considerations, non-clinical studies and clinical studies. In the following, we introduce the general concepts that are important for biosimilar development. Since the methodological contributions presented in this thesis relate exclusively to the clinical studies, we focus our attention on the regulatory requirements for the clinical part of development. More information on quality and non-clinical studies can be found in the respective guidelines (CHMP, 2014a,b).

The guiding principle for biosimilar approval in Europe is the idea of the *totality of the data* (EMA, 2016). This means that there is not one pivotal step or study in the development programme, but all information is considered important and the final decision as to whether a product is approved or not will be based on all provided data (i.e., on quality, non-clinical and clinical data). In particular, it is possible to gain approval even in cases in which a single study or a single analysis failed (CHMP, 2014a) as long as justification is provided.

The clinical development programme consists in most cases of at least one Phase I study in which the pharmacokinetics (PK, i.e., what the body does to the drug) and the pharmacodynamics (PD, i.e., what the drug does to the body) of the biosimilar are compared to the reference product. The assessment of PK is comparable to the showing of bioequivalence for generics (Patterson and Jones, 2017) and follows mostly a well-established and standardised approach with only a small degree of flexibility for sponsors: the drug is administered to the subject and the concentration of the drug in the blood over time is measured. Measures like the area under the drug concentration vs. time curve (AUC) and the maximum concentration over time (Cmax) are reported for each subject. If it can be shown that the ratio of the reference product and the proposed biosimilar for each of these measures as a percentage lies within 80% and 125% with a pre-specified confidence level $1 - \alpha$, commonly $\alpha = 0.1$, bioequivalence is established. In contrast, the assessment of PD markers (i.e., surrogate markers for efficacy in patients) is less standardised since the importance of PD markers highly depends on the availability of biomarkers that are well-accepted surrogates of and are strongly related to the clinical outcome. This availability clearly depends on the active substance (the active ingredient) and is therefore product specific. If no PD marker exists that can be considered relevant to predict the efficacy of the biosimilar, no PD markers may be included in the Phase I studies. On the other hand, if there is a well-established PD marker, additional confirmatory clinical trials in patients may be waived. (CHMP, 2014a)

If Phase III studies have to be conducted, these are performed in patients in at least one indication of the reference product. The chosen indication should be sensitive to detect potential differences in efficacy and safety. That is why an indication with a large treatment effect is typically chosen. For that matter, it is not required that the indication is the one that was studied in the regulatory application of the reference product. Equivalent efficacy is demonstrated at a pre-specified point in time for a chosen endpoint, which can, but does not have to be, the one studied for the reference product. From a statistical point of view, using an equivalence test (Wellek, 2010) with a pre-specified equivalence margin (the maximum difference in the endpoint between the biosimilar and the reference product which is considered not to be clinically relevant) is recommended, but also a non-inferiority approach (Ng, 2014) might be acceptable if justified. In addition to showing equivalence in efficacy, establishing the equivalence of the safety and immunogenicity profile is the main goal of the Phase III studies. Immunogenicity refers to the ability of a drug to induce an immune response (e.g., anaphylaxis). While a formal statistical testing procedure is mostly required for showing equivalence on PK, PD and efficacy endpoints, safety and immunogenicity are only analysed descriptively. (CHMP, 2014a)

Since the drug is typically only studied in patients in one or, at most, two or three indications while the reference product is usually approved for multiple indications, the concept of *extrapolation* plays an important role in biosimilar development: the information that was gathered during analytical and non-clinical development, the knowledge about the mechanism of action of the drug and the limited clinical data allow the regulatory approval of the biosimilar in indications which were not explicitly studied by appealing to scientific judgement (CHMP, 2014a). Sponsors tend to conduct post-marketing studies in extrapolated indications and, so far, no concerns have been reported concerning the extrapolated indications (Weise et al., 2014).

In this cumulative thesis, several statistical topics in clinical biosimilar development are presented. Even though the chapters cover a wide spread of statistical methodology, it is important to emphasise that there are some common features in all topics of this thesis: first, all results are applicable in the area of clinical biosimilar development, i.e., there is a joint field of application. Second, the aim of all presented contributions is to establish equivalence. Thus, the null hypothesis is that the absolute difference between two quantities is larger than a pre-specified value and the alternative is that the absolute difference is smaller than this value. More formally, let γ_T be a characteristic of interest of the biosimilar (the test product, T) and γ_R be the same characteristic of interest of the reference product (R). Then, we test the hypotheses (Wellek, 2010)

$$H_0: \ |\gamma_T - \gamma_R| \geq \Delta \text{ vs. } H_1: \ |\gamma_T - \gamma_R| < \Delta,$$

where $\Delta \in \mathbb{R}_+$ has to be pre-specified and, in the context of biosimilar trials, justified from a clinical and statistical perspective (CHMP, 2014a). Even though many concepts which were developed for superiority testing are also applicable to equivalence testing, it is important to keep in mind that some approaches for superiority testing are not applicable in this context (e.g., non-parametric permutation tests).

The rest of this thesis is structured as follows: Chapter 2 deals with regulatory requirements for clinical biosimilar development in the EU. In the recent past, the EMA has established itself as the leading regulatory agency for biosimilar approval. It approved the first biosimilar (Omnitrope, Sandoz) in 2006 and has since then published several guidelines that support sponsors in conducting appropriate clinical development programmes. Currently, there exist one overarching guideline, two guidelines focussing on the development of biotechnology-derived proteins (clinical and non-clinical development and quality issues), eight product-specific guidelines which give advice for a specific active substance and four other related guidelines. However, at the start of this research, it was unclear how these guidelines were put into practice, i.e., how closely sponsors were following these guidelines and how the regulatory agency was handling deviations from the guidelines. In Chapter 2, we present the results of two systematic reviews in which key characteristics of the clinical development programmes (e.g., sample sizes, endpoints, study populations) are compared with the aim of providing a comprehensive overview of biosimilar development in practice in the EU.

One of the results of our systematic reviews (Chapter 2) is that sponsors often conduct multiple trials in different indications, study several dosing regimens or analyse multiple endpoints. Therefore, it is necessary to discuss the implication of multiple testing on the interpretation of the test results. In principle, sponsors aim to demonstrate equivalence in all indications, dosing regimens and endpoints. In that case, the control of the Type I error rate (the rate of false positive test decision, i.e., the control of the patient's risk) is not jeopardised, but the Type II error rate (the rate of false negative test decisions, i.e., the sponsor's risk) is high compared to the Type II error rate for a single test. Consequently, it might be necessary to increase the number of subjects in the study to have a reasonable chance of claiming equivalence in all indications, dosing regimens and endpoints. Let m denote the number of hypotheses of interest. In Chapter 3, we first illustrate the impact of multiple hypothesis testing on the required sample size when all m hypotheses have to be rejected and show that there are scenarios in which the required sample size is unrealistically high for biosimilar development. Since we noted in Chapter 2 that biosimilars gained approval in the past even in situations where equivalence was not shown for all indications, dosing regimens and endpoints by using the concept of *totality of the data*, we develop in Chapter 3 a strategy to test if at least k out of these m tests are successful where $k < m$. This reduces the required sample size to a feasible level and has, in contrast to the currently used practice of deciding post-hoc if the provided evidence is sufficient, the advantage of being a well-defined formal testing procedure with known operating characteristics. However, a multiplicity adjustment might be required for limiting the Type I error rate and we discuss the impact of different types of adjustment both in simulations and in case studies.

Keeping the sample size as low as possible is a key requirement in biosimilar development because biosimilars are supposed to be sold at a cheaper price than the reference product. Consequently, the development costs have to be noticeably lower for the biosimilar compared to the reference product. The approach proposed in Chapter 3 is one way to limit the burden on the sponsor while providing a high chance of regulatory approval. A different strategy for reducing the number of subjects involved in the clinical trials, and therefore to reduce costs, is to make the best use of all available information. Since the reference product has already been studied in several trials before the start of the development of the biosimilar, it seems attractive to include all available knowledge about the reference product into the analysis of the biosimilar clinical trial that has been run for showing equivalent efficacy. However, it is well-known that the incorporation of historical information in an analysis can lead to an inflation of the Type I error rate if the data from the historical studies do not match the results from the new study, i.e., in the case of a prior-data conflict. Since all biosimilar trials are confirmatory trials, control of the Type I error rate in all situations which are realistic in practice is expected to be a regulatory requirement and so making the use of historical data challenging in this context. In Chapter 4, we first show that the aims of complete Type I error rate control and a gain in power compared to the standard frequentist approach, which considers only the data from the new biosimilar trial, are incompatible. To illustrate this, we use the robustified meta-analytic-predictive (MAP) approach (Schmidli et al., 2014) as an example methodology which incorporates historical data. Afterwards, we introduce a novel hybrid Bayesian-frequentist approach for binary endpoints which guarantees partial Type I error rate control, i.e., Type I error rate control in all situations that are realistic in practice while providing an advantage in terms of power. We study the performance of the proposed approach in an extensive simulation study and present a case study for illustrating the application of the proposed methodology in practice.

The case study discussed in Chapter 4 focusses on the assessment of the treatment effect under continuous treatment with the biosimilar or the reference product. The associated analysis provides most information on the direct comparability of the biosimilar and the reference product for treatment-naive patients who start with one of the treatments and continue to be treated with this product of choice for the complete duration of their treatment. In practice, however, biologics are often used for treating chronic diseases and during the long period of treatment, patients might wish to switch between the biosimilar and the reference product once or even multiple times for several different reasons (e.g., change of reimbursement policy of health care provider, introduction of another biosimilar to the market, change of doctor). But even after more than ten years of experience with biosimilars in practice in Europe, there is still uncertainty if patients can switch between a biosimilar and its reference product without impacting the expected efficacy and safety of the treatment. Regulatory agencies approach this topic with different strategies (see, for example, Tóthfalusi et al., 2014): while the EMA states that "the Agency's evaluations do not include recommendations on whether a biosimilar should be used interchangeably with its reference medicine" and recommends that "for questions related to switching from one biological medicine to another, patients should speak to their doctor or pharmacist" (EMA, 2012a), the FDA has, as stated in the *Biologics Price Competition and Innovation Act* (BPCI Act) that was introduced in 2009, the legal requirement to offer the opportunity to approve a biosimilar as an *interchangeable biosimilar* (FDA, 2009). However, so far no interchangeable biosimilars have been approved in the US. In this thesis, we use the general term *switchability* to refer to the property of the biosimilar that switching once or multiple times with the reference product has no relevant impact on the patient's response to the treatment.

Very often, no clinical data on patients switching between the biosimilar and its reference product are available at the time the biosimilar gets to the market since the large Phase III studies, which are required prior to market authorisation, are mostly conducted using parallel groups designs (see Chapter 2). Studies using parallel groups designs cannot answer the question if patients can switch safely between the biosimilar and its reference product because no transitions between the biosimilar and the reference product are included in the study design thus making it impossible to study the effect of switching. Crossover designs may be more appropriate and we discuss in Chapter 5 efficient study designs for estimating the so-called mixed and self-carryover effects, which are closely related to the effect of switching. The model that includes the self and mixed-carryover effects was first proposed by Afsarinejad and Hedayat (2002). They introduced a model in which the usual, first-order carryover effect that only depends on the treatment given in the immediately previous period is replaced by two different carryover effects per treatment. These carryover effects do not only depend on the treatment given in the immediately previous period, but also on the treatment given in the current period: if the treatments in both periods (current and previous) are the same, a potential self-carryover effect is introduced, if the treatments differ in the two periods, a mixed-carryover effect may be present. Previously, self and mixed-carryover effects were considered as nuisance parameters (Kunert and Stufken, 2002, 2008) and the focus was on the estimation of the direct effects of the treatments, adjusted for self and mixed-carryover effects. In the assessment of switchability, we can assume that equality of the direct treatment effects has been established previously and instead the mixed-carryover effects, which describe the impact of switching, and the self-carryover effects, which relate to continuous treatment with the biosimilar or the reference product, are of most interest. Therefore, in

Chapter 5, we derive efficient designs for estimating the mixed-carryover effects separately and for estimating self and mixed-carryover effects simultaneously.

In Chapter 6, we introduce three statistical tests for formally establishing switchability based on the data obtained from several periods of treatment, i.e., longitudinal data. One of these tests uses an idea that is related to the estimation of self and mixed-carryover effects (which we refer to as the estimation method), even though the details are slightly different. For all our proposals for testing for switchability, we assume a study design which is motivated by a study design that has already been used in practice (the EGALITY study, see Griffiths et al. (2017)) and not the efficient designs derived in Chapter 5. Therefore, while Chapter 5 gives theoretical results to a fundamental problem which might also be relevant in other fields of application, in Chapter 6 we tailor the proposed methodologies to the goal of establishing switchability. This might make the approaches more attractive for an application in practice. The part of the design of the EGALITY study which is related to the assessment of switchability consists of four periods of treatment and four sequences (orders of treatment), which can be split into switching (switch from the biosimilar to the reference product and back after each period) and non-switching sequences (continuous treatment with the biosimilar or the reference product). Generally speaking, the two of the three developed methods which differ substantially from the one that directly compares the self and mixed-carryover effects, use the idea of comparing switching with non-switching sequences to assess switchability. In Chapter 6, we first discuss which patterns in the data have to be visible such that we would consider a proposed biosimilar to be switchable or not switchable. These considerations are used for formally defining the null and the alternative hypothesis that are to be assessed. Afterwards, we describe the three different approaches which use the longitudinal data from the patients for testing these hypotheses and discuss the strengths and weaknesses of these tests in a simulation study. We also use the data of the EGALITY study (Griffiths et al., 2017) to illustrate the performance of the proposed methods in practice. Lastly, we discuss the efficiency of some study designs for the estimation method building up upon the results in Chapter 5. We show that the efficient study design for the estimation of self and mixed-carryover effects which was derived in Chapter 5 is more efficient for the estimation method than the design used in the EGALITY study.

Chapter 2

Clinical biosimilar development in practice in the European Union

2.1 Contributed material

Mielke, J., Jilma, B., Koenig, F. and Jones, B. (2016): Clinical trials for authorized biosimilars in the European Union: a systematic review. *British Journal of Clinical Pharmacology*, 82 (6), 1444–1457.

Mielke, J., Jilma, B., Jones, B. and Koenig, F. (2018a): An update on the clinical evidence that supports biosimilar approvals in Europe. *British Journal of Clinical Pharmacology*, 84 (7), 1415–1431.

Co-authors' contribution: Bernd Jilma proposed the project and gave advice on questions related to the clinical interpretation of the results. This project was partially completed during a research stay at the Medical University of Vienna under the supervision of Franz Koenig. Byron Jones helped with the interpretation of the results and the presentation of the material.

2.2 Key results

Biosimilars are still a fairly new concept with the first biosimilar approved in the EU in 2006 (Omnitrope, Sandoz). That is why there is still some uncertainty among sponsors on the regulatory expectation of the amount and type of evidence that has to be provided for gaining approval as a biosimilar. The regulatory agency in the EU, the EMA, has published several guidelines in order to advise sponsors on their biosimilar development strategies. These guidelines provide non-binding recommendations on scientific questions of biosimilar development. However, prior to the start of this research, it was not clear how sponsors and regulators have put these guidelines into practice. In the two contributed papers, we systematically compare the clinical studies of successful biosimilar development programmes which were submitted to the EMA with the aim of providing a comprehensive overview of clinical biosimilar development in practice. We focus on two different aspects. First, we compare the clinical development programmes of biosimilars and specifically focus on the situation where two or more approved biosimilars contain the same active substance, i.e., share the same reference product. This analysis aims to provide insights into the question of whether biosimilar development is, even though experience with biosimilars is limited, already standardised or if companies are still discordant in their development strategies. Second, we compare the regulatory guidelines with the biosimilar development programmes in practice to clarify if there were deviations from the guidelines. In the case of identified deviations from the guidelines, we analyse the way regulators dealt with these deviations.

Currently, there are 43 approved biosimilars in Europe (EMA, 2018). The biosimilars which gained approval prior to August 2016 are discussed in Mielke et al. (2016). In Mielke et al. (2018a), we discuss special features of biosimilars which were approved between September 2016 and November 2017 and compare additional key characteristics for all approved biosimilars. In both contributed papers, we confirm that there is a high variability in the development strategies. Interestingly, this is also true for biosimilars with the same active substance (the same reference product): for example, we report in Mielke et al. (2016) that some companies reduced the size of their studies for showing equivalent efficacy and safety in patients and provided, as compensation, more evidence in the pharmacokinetics (PK)/pharmacodynamics (PD) part of the development programme, while other companies conducted extensive clinical trials in patients. In Mielke et al. (2018a), we compare the study populations used for the efficacy and safety assessment and report that these are not necessarily identical even if the biosimilars are approved for the same reference product. In total, we conclude that there seems to be a fair amount of flexibility for sponsors to set-up the development programme according to their preferences.

While comparing the recommendations presented in the guidelines to the studies conducted in practice, we first note that often the product-specific guideline (the guideline that gives detailed recommendations for a specific class of products) was not available at the time the development of the first biosimilar within a class started (Mielke et al., 2018a). This might indicate that these guidelines are, in most cases, not prepared prior to the first sponsor approaching the regulators with questions regarding a specific product. In some cases, the companies followed all recommendations, but we report cases in which companies deviated in a few aspects (e.g., the study design was not the one recommended as for the biosimilars Epoetin Alfa Hexal/Abseamed/Binocrit, Mielke et al., 2016). In addition, we also identify cases with major deviations from the guideline that was in operation during the time of development. For example, in one case a sponsor provided substantially less evidence than requested in the guideline (biosimilars Inhixa/Thorinane, Mielke et al., 2018a). In this case, the product-specific guideline was changed after the approval in order to reflect the development programme of the sponsor. In contrast, we notice examples in which sponsors provided more evidence than explicitly requested (insulin biosimilars, Mielke et al., 2018a). We conclude that European regulators are willing to accept deviations from their guidelines as long as sound scientific justification is provided.

The main resources for the two systematic reviews are the so-called European public assessment reports (EPARs) which are publicly available and offer insights into the studies and analyses presented to the regulators when a sponsor applies for market authorisation. For example, for the clinical part, details on the planning of the study (e.g., sample sizes, endpoints, equivalence margins) are stated and the study results are reported. The EPARs of the first approved biosimilars are mostly short and do not provide detailed information (Mielke et al., 2016), however, we notice that the quality of the EPARs has improved recently (Mielke et al., 2018a). Nonetheless, often there is still some information missing (e.g., justification of the equivalence margins) and we propose in Mielke et al. (2018a) a checklist with the minimal information that is recommended to be included in an EPAR.

During the review of clinical development programmes, we noted that often multiple endpoints, treatment regimens and doses or study populations were analysed. These are all situations in which one needs to decide if multiple testing has to be considered in the interpretation of the test results. However, we only identified a few EPARs in which this was explicitly stated (Mielke et al., 2018a). We discuss the implication of multiple testing and propose solutions for handling multiplicity in a manageable way in biosimilar development in Chapter 3.

Chapter 3

Sample size for multiple hypothesis testing in biosimilar development

3.1 Contributed material

Mielke, J., Jones, B., Jilma, B. and König, F. (2018b): Sample size for multiple hypothesis testing in biosimilar development. *Statistics in Biopharmaceutical Research*, 10 (1), 39–49.

Co-authors' contribution: Franz Koenig proposed the project and gave advice during a research stay at the Medical University of Vienna. Bernd Jilma contributed to the case studies. Byron Jones helped with the interpretation of the results and the presentation of the material.

3.2 Key results

One of the results of the systematic reviews presented in Chapter 2 was that sponsors often set-up the clinical development programme of biosimilars in such a way that the impact of multiple testing needs to be taken into account during the interpretation of the results. In practice, however, this need is often ignored. For example, we identified situations in which sponsors considered multiple treatment regimens within one study (e.g., the pharmacokinetics (PK)/pharmacodynamics (PD)-trial undertaken for the application of Tevagrastim (Lubenau et al., 2009) in which several doses and routes of administration were compared) or situations in which the drug was tested in several patient populations (e.g., for the application of Abasaglar, the sponsor conducted one Phase III study in patients with diabetes mellitus type 1 (Blevins et al., 2015) and one Phase III study in patients with diabetes mellitus type 2 (Rosenstock et al., 2015)). Also, in the regulatory guideline on non-clinical and clinical issues in biosimilar development published by the EMA (CHMP, 2014a), it is recommended to show equivalence in PK studies both for AUC (area under the drug concentration vs. time curve) and Cmax (maximum concentration of the drug over time). Thus, multiple co-primary endpoints must be considered.

In all these examples, it is desired that equivalence is shown for all treatment regimens, endpoints and study populations. Since we need to reject *all* hypotheses, the Type I error rate is controlled and no multiplicity adjustment is needed. This is also explicitly stated in the European public assessment report (EPAR) for Lusduna (CHMP, 2016): "It should be noted that within the pharmacodynamics hypothesis and within the pharmacokinetic hypothesis no multiplicity adjustment is applied, since the mean treatment ratio for each endpoint needs to lie within (0.8, 1.25) to support the particular primary hypothesis." However, the Type II error rate (the sponsor's risk that a study fails) may be increased. Obviously, it is possible to reduce the Type II error rate to the nominal level by increasing the sample size, but this consequence of using multiple tests seems to be rarely considered in biosimilar development in practice. In what follows, we use the term *test* in order to refer to the multiple treatment arms, studies or endpoints.

In the contributed paper, we first show the impact of multiple testing on the sample size and illustrate that if equivalence on multiple tests has to be confirmed, the sample size drastically increases. For example, in a situation in which 76 subjects would be required for 80% power for an individual test, 134 subjects would be necessary for 80% power for five uncorrelated tests. For details we refer to the contributed paper. Although a higher correlation between the tests decreases the required sample size, it is important to emphasise that the correlation structure of the tests is often not known in practice and this makes it difficult to make a reasonable assumption a priori. If uncorrelated tests have to be assumed, the required sample size might be unrealistically high in the framework of biosimilar development.

That is why we propose in the contributed paper a novel strategy for controlling the effect of multiple testing in a manageable way in clinical biosimilar development. The proposed strategy is motivated by an observation that we made in the first systematic review which is presented in Chapter 2 (Mielke et al., 2016): we noticed that biosimilars gained approval, using the concept of *totality of the data*, even though not all primary endpoints were successful. This happened, for example, in the application for Zarzio by Sandoz (CHMP, 2008) which is a biosimilar with the active substance filgrastim. The sponsor submitted four PK/PD-studies in which four different doses (1, 2.5, 5, and 10 μg/kg) were assessed and one of the doses was studied in two routes of administration. For the lower doses and after multiple subcutaneous doses, both the endpoints Cmax and AUC failed to show equivalence of the biosimilar and the reference product. Nevertheless, the EMA approved the product.

Therefore, in situations in which a very high sample size would be required to achieve a reasonable probability to claim equivalence for all m multiple tests ($m \in \mathbb{N}$), we propose to claim success on the overarching hypothesis of biosimilarity if at least k ($k \in \{1, \dots, m\}$) out of the m tests are successful and the choice of k has to be made during discussions with the regulatory agency. The idea of rejecting an overarching hypothesis if at least k out of m tests are successful was first introduced by Rüger (1978), but results on the operating characteristics of such a test were, to the best of our knowledge, not published. Formally, we assume that a finite number of statistical hypotheses $H^{(1)}, \dots, H^{(m)}$ are tested. Then, the overarching null hypothesis for the k-out-of-m-test is defined by

$$H_0: \text{ At least } m - k + 1 \text{ null hypotheses } H^{(i)} \text{ are true, } i = 1, \dots, m,$$

and the overarching alternative hypothesis is given by

$$H_1: \text{ Less than } m - k + 1 \text{ null hypotheses } H^{(i)} \text{ are true, } i = 1, \dots, m.$$

In other words, the aim is to reject the overarching null hypothesis if at least k individual null hypotheses $H^{(i)}$ are false. The test decision of the k-out-of-m-test for this overarching null hypothesis is to reject if and only if

$$\sum_{i=1}^{m} r^{(i)} \geq k,$$

where $r^{(i)}$ gives the test decision for the individual hypothesis $H^{(i)}$ and is therefore the realisation of a random variable $R^{(i)}$ that represents the test decision for hypothesis $H^{(i)}$ with

$$R^{(i)} = \begin{cases} 0, \text{ if the null hypothesis of test } i \text{ is not rejected,} \\ 1, \text{ if the null hypothesis of test } i \text{ is rejected.} \end{cases}$$

If multiple tests are conducted and it is not required that all tests are successful, then the control of the familywise error rate (FWER, see, for example, Bretz et al. (2016)) is not guaranteed if the significance level is not adjusted to account for multiple testing. A simple multiplicity adjustment that leads to a controlled FWER is the Bonferroni adjustment (see, for example, Bretz et al., 2016), where instead of the desired significance level α, all m tests are conducted with a level α^*. The level α^* is given by

$$\alpha^* = \frac{1}{m} \cdot \alpha.$$

The control of the FWER is very stringent. Therefore, the less stringent, so-called k-FWER error rate control might be more appropriate for an application in practice. The k-FWER was introduced by Lehmann and Romano (2005) and is given by

$$k\text{-FWER} = P\{\text{reject at least } k \text{ hypothesis } H^{(i)} \text{ with } i \in I\},$$

where $I \subseteq \{1,\ldots,m\}$ is the set of true null hypotheses. Hommel and Hoffmann (1988) proposed an adjustment of the significance level that guarantees the control of the k-FWER. Let α be the desired significance level. Then, the significance level α^* for the m individual tests is given by

$$\alpha^* = \frac{k}{m} \cdot \alpha.$$

We call this adjustment the k-adjustment in the following. In the contributed paper, we investigate the performance of the k-out-of-m-test in a simulation study. We show that this approach noticeably reduces the required sample size. For example, for the scenario mentioned above for which 76 subjects are required for an individual test and 134 subjects are required for five uncorrelated tests, the sample size is reduced to 90 subjects for 4 out of 5 uncorrelated tests if the k-adjustment is used. A higher correlation does not necessarily reduce the required sample size if the k-out-of-m-test is used, which is contradictory to the situation in which equivalence has to be claimed on all hypotheses.

We also compare the different multiplicity adjustments (Bonferroni adjustment, k-adjustment, no adjustment) in the simulation study and find that the number of additional subjects, which are required if the k-adjustment is used instead of no adjustment is limited if k is chosen to be close to m, which is the most relevant situation in practice. It is well-known that the Bonferroni adjustment is conservative even in the case of $k = 1$ and even more conservative for higher values of k. That is why the increase in sample size compared to the k-adjustment or no adjustment can be extreme if the Bonferroni adjustment is used.

The contributed paper concludes with case studies to illustrate the increase in sample size in practice if multiple testing is taken into account. In addition, the advantage of using the k-out-of-m-test is demonstrated.

3.3 Type I error rate control with the k-out-of-m-test

While the k-adjustment provides, especially in the context of *totality of the data* (see Chapter 1), an attractive compromise between the control of false positive decisions on the individual hypotheses $H^{(i)}$ $(i = 1, \ldots, m)$ and a realistic sample size, it is important to emphasise that the Type I error rate of the k-out-of-m-test is not controlled with the k-adjustment: the k-adjustment controls the k-FWER only, i.e., the risk to observe more than k false positive decisions on the individual hypotheses $H^{(i)}$. It is, however, not necessary to make k wrong decisions on the individual hypotheses $H^{(i)}$ for making a wrong decision on the overarching null hypothesis H_0. Let us assume that $k = 3$ and $m = 5$ are considered. Then, there could be, for example, three individual hypotheses under the null hypothesis and two individual hypotheses for which the alternative hypothesis would be the correct decision. In this situation, the correct test decision on the overarching hypotheses would be for the null hypothesis. However, with just *one* false positive decision for an individual hypothesis, we could reject the overarching null hypothesis. In the multiple testing framework, the control of the FWER under all possible configurations is called strong control whereas the control of the FWER under the complete null hypothesis (all individual null hypotheses are true) is called weak control (Bretz et al., 2016). We will use this terminology in the following even though the k-out-of-m-test is not a classical multiple testing approach since finally only one overarching test decision is made.

The k-adjustment controls the Type I error rate for the k-out-of-m-test if all individual null hypotheses are true, i.e., it offers weak control of the Type I error rate for the k-out-of-m-test. If strong control of the Type I error rate is required, the k-adjustment is too liberal. In the following, we discuss a multiplicity adjustment that gives strong Type I error rate control for the k-out-of-m-test.

More formally, we aim to control the Type I error rate at level α for the k-out-of-m-test. Let $I \subseteq \{1, \ldots, m\}$ be the set of true individual null hypotheses and $F \subseteq \{1, \ldots, m\}$ be the set of false individual null hypotheses with

$$I \cap F = \emptyset \text{ and } I \cup F = \{1, \ldots, m\}.$$

For controlling the Type I error rate with the k-out-of-m-test, it has to hold that

$$\begin{aligned} P(\text{Decide for } H_1 | H_0 \text{ true}) = &P(\text{Decide for } H_1 | \text{ at least } m-k+1 \text{ hypotheses are in } I) \\ &\leq \alpha. \end{aligned}$$

Therefore, it is necessary to control the Type I error rate for all situations in which at least $m-k+1$ individual null hypotheses are true, i.e., in situations in which $m-k+1, \ldots, m$ individual null hypotheses are true. Thus, it has to hold that

$$\begin{aligned} P(\text{Decide for } H_1 | \, |I| \geq m-k+1) = &P((\text{Decide for } H_1 | \, |I| = m-k+1, \, |F| = k-1) \\ &\cup (\text{Decide for } H_1 | \, |I| = m-k+2, \, |F| = k-2) \\ &\cup \ldots \cup (\text{Decide for } H_1 | \, |I| = m, \, |F| = 0)) \leq \alpha, \end{aligned}$$

where $|A|$ is the number of elements in a set A. Since one will be in exactly one of these situations (a specific set of hypotheses will be true or false), it is sufficient to guarantee that

$$\begin{aligned} &P(\text{Decide for } H_1 | \, |I| = m-k+1, \, |F| = k-1) \leq \alpha, \\ &P(\text{Decide for } H_1 | \, |I| = m-k+2, \, |F| = k-2) \leq \alpha, \\ &\qquad\qquad\qquad \vdots \\ &P(\text{Decide for } H_1 | \, |I| = m, \, |F| = 0) \leq \alpha. \end{aligned}$$

Therefore, it is necessary to identify a significance level α^* that can be used for testing the individual hypotheses $H^{(1)}, \ldots, H^{(m)}$ such that these equations hold true. As an example, let us again consider the situation with $m = 5$ and $k = 3$. Then, there are three different combinations of number of tests under the null and alternative hypotheses which could potentially lead to a false positive decision with the k-out-of-m-test: it could be that all five null hypotheses are true or that four null hypotheses are true or that three null hypotheses are true. Therefore, it is necessary to ensure that

$$\begin{aligned} P_1 &= P(\text{Decide for } H_1 | \, |I| = 3, \, |F| = 2) \leq \alpha, \\ P_2 &= P(\text{Decide for } H_1 | \, |I| = 4, \, |F| = 1) \leq \alpha, \\ P_3 &= P(\text{Decide for } H_1 | \, |I| = 5, \, |F| = 0) \leq \alpha. \end{aligned}$$

In the last scenario, all null hypotheses are true. That is why $k = 3$ wrong decisions on the individual hypotheses $H^{(1)}, \dots, H^{(m)}$ are necessary for making a wrong decision with the k-out-of-m-test. Thus, this is the situation for which the proposed k-adjustment is the correct adjustment because it guarantees the control of the k-FWER. The k-adjustment of the significance level α in this situation is given by

$$\alpha_1^* = \frac{k}{m} \cdot \alpha = \frac{3}{5} \cdot \alpha.$$

In the second situation, the rejection of two of the four true null hypotheses is required for making a wrong test decision with the k-out-of-m-test (since one test is already under the alternative), therefore by using the idea of the k-adjustment, we need to use

$$\alpha_2^* = \frac{k-1}{m-1} \cdot \alpha = \frac{2}{4} \cdot \alpha = \frac{1}{2} \cdot \alpha.$$

For the first situation, only one out of the three true null hypotheses has to be rejected for making a wrong test decision with the k-out-of-m-test (since two tests are already under the alternative). Therefore, the adjustment is given by

$$\alpha_3^* = \frac{k-2}{m-2} \cdot \alpha = \frac{1}{3} \cdot \alpha.$$

Since we do not know if we are in Situation 1, 2 or 3, we have to use the most conservative adjustment, i.e., α_3^* would be the correct multiplicity adjustment if the Type I error rate of the k-out-of-m-test is to be controlled.

Next, we consider the general case of m tests out of which k have to be successful. Then, the following potential multiplicity adjustments are obtained using the same strategy as above:

$$\begin{aligned}
\alpha_1^* =& \frac{k}{m} \cdot \alpha \text{ (for } |I| = m\text{)},\\
\alpha_2^* =& \frac{k-1}{m-1} \cdot \alpha \text{ (for } |I| = m-1\text{)},\\
&\vdots\\
\alpha_k^* =& \frac{1}{m-k+1} \cdot \alpha \text{ (for } |I| = m-k+1\text{)}.
\end{aligned}$$

Since

$$\frac{k-l}{m-l} \leq \frac{k}{m}$$

for $l = 0, \dots, k-1$, the adjustment of the significance level that offers strong Type I error rate control for the k-out-m-test is given by

$$\alpha^* = \frac{1}{m-k+1} \cdot \alpha.$$

We will call this adjustment the t-adjustment in the following. Compared to the k-adjustment which was proposed in the contributed paper and used

$$\alpha^* = \frac{k}{m} \cdot \alpha,$$

the t-adjustment is more stringent. The more stringent error control leads to a higher number of subjects which are to be included in the study and this will be explored in the following.

Figure 1 compares the minimal required sample size if the k-adjustment or the t-adjustment is used for a target power of 80% for different levels of correlation ρ between the tests. The results were generated assuming the same setting as the ones used for Figure 3 in the contributed paper (Mielke et al., 2018b), i.e., we assume that $m = 5$ tests are carried out and at least $k = 1, \ldots, 5$ out of these 5 tests have to be successful. For full details on the hypotheses which are to be tested and the assumed underlying distribution of the tests, we refer to the contributed paper. The figure shows that the increase in sample size is moderate if the t-adjustment is used instead of the k-adjustment.

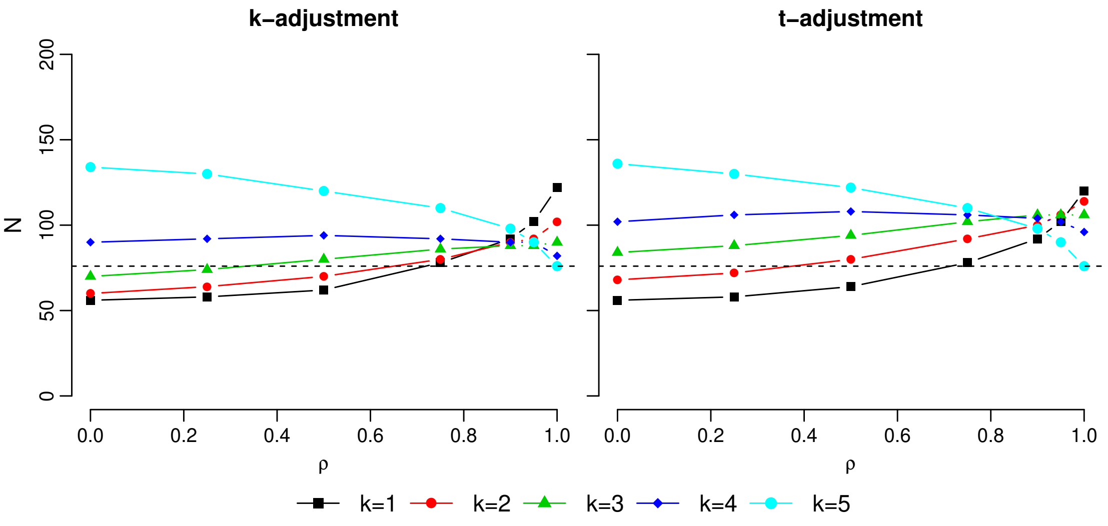

Figure 1: Comparison of the total required sample size for 80% power for the k-out-of-m-test using the k-adjustment and the t-adjustment. The required sample size is displayed for $m = 5$ and $k = 1, \ldots, 5$ for different levels of correlation ρ between the m tests. The dashed line indicates the required sample size for a single test ($m = k = 1$). Full information on the setting and the hypothesis testing can be found in Mielke et al. (2018b).

The final decision, which adjustment for multiplicity is to be used, depends on the required control of false positive decisions and has to be made taking into account the importance of the studies in the development programme in the context of the *totality of the data* (see Chapter 1). If, for example, the concerned analyses are supportive only, one might be willing to accept the weaker control of the k-FWER. On the other hand, if the concerned tests are considered the most important piece of evidence, a stricter control as achieved with the t-adjustment might be required.

Chapter 4

Incorporating historical information in biosimilar trials

4.1 Contributed material

Mielke, J., Schmidli, H. and Jones, B. (2018c): Incorporating historical information in biosimilar trials: challenges and a hybrid Bayesian-frequentist approach. *Biometrical Journal*, 60 (3), 564–582.

Co-authors' contribution: Byron Jones gave advice during the development of the proposed methodology and helped with the presentation of the material. Heinz Schmidli provided information and advice on Bayesian methodology and commented on the proposed approach.

4.2 Key results

Drug development is very expensive (DiMasi et al., 2003). Keeping the costs as low as possible is important and especially essential in biosimilar development because biosimilars are expected to be sold at a cheaper price than the reference product (Haustein et al., 2012). One way to reduce the costs of development is to make the best use of all available information. Since biosimilars are developed as copies of previously approved products and these products have already been studied several times both prior to market authorisation and in post-marketing trials, it seems natural to include the knowledge gathered in these studies into the showing of equivalent efficacy of the biosimilar and its reference product.

However, it is well-known that the incorporation of historical information can lead to an inflation of the Type I error rate in situations in which the data from the historical studies do not match the data from the new trial (the so-called prior-data conflict, see Schmidli et al., 2014). Since all studies in biosimilar development are confirmatory studies, we expect that an inflation of the Type I error rate will not be acceptable – especially if this inflation occurs in situations which are realistic in practice. In the contributed paper, we focus on binary endpoints (e.g., responders vs. non-responders). Let p_T and p_R be the true response rates of the biosimilar (test, T) and the reference product (R), respectively. We test the hypotheses

$$H_0 : |p_R - p_T| \geq \Delta \text{ vs. } H_1 : |p_R - p_T| < \Delta,$$

where $\Delta \in \mathbb{R}_+$ is a pre-specified value (the equivalence margin) and is the maximum difference such that the response rates are not considered different from a clinical point of view.

In the contributed paper (Mielke et al., 2018c), we assume a parallel groups design with n subjects per sequence. Then, we first show that the goals of complete control of the Type I error rate and a gain in power are incompatible if historical information is incorporated. We use the robustified Meta-Analytic-Predictive (MAP) approach which was introduced by Schmidli et al. (2014) as an example methodology. The MAP approach is one of the frequently used methodologies for incorporating historical data. It falls into the framework of Bayesian approaches and combines the information from the historical trials into a prior distribution by taking the between-trial variation into account. The main assumption for the approach is that the parameters of interest of the studies are not identical but similar and the degree of similarity is quantified with a random-effects meta-analytical model. The prior and the observed data in the study are combined using Bayes' theorem. As historical information is available for the reference product only, we use the informative prior derived with the MAP approach for the reference product, but assume a non-informative (uniform) distribution for the biosimilar. The decision whether the biosimilar and the reference product are equivalent is made using a Bayesian success criterion: let X_T and X_R be random variables that follow the posterior distributions of the test and the reference product, respectively. Equivalence is claimed if

$$B := P(|X_R - X_T| < \Delta) > c,$$

where $c \in [0,1]$ has to be chosen such that the desired Type I error rate profile is achieved. We first derive the Type I error rates of the MAP approach for several degrees of prior-data conflict ranging from a perfect match between historical data and data in the new trial to a complete mismatch. We show that, while the Type I error rate is controlled in the case of no prior-data conflict, the Type I error rate is substantially inflated if the mean value of the prior does not match the true response rate in the new study. It is most concerning that this inflation also occurs in situations with a minor prior-data conflict, i.e., in situations which are relevant in practice. That is why we conclude that the Type I error rate profile which we obtain with the MAP approach is unlikely to be accepted for regulatory approval in biosimilar development.

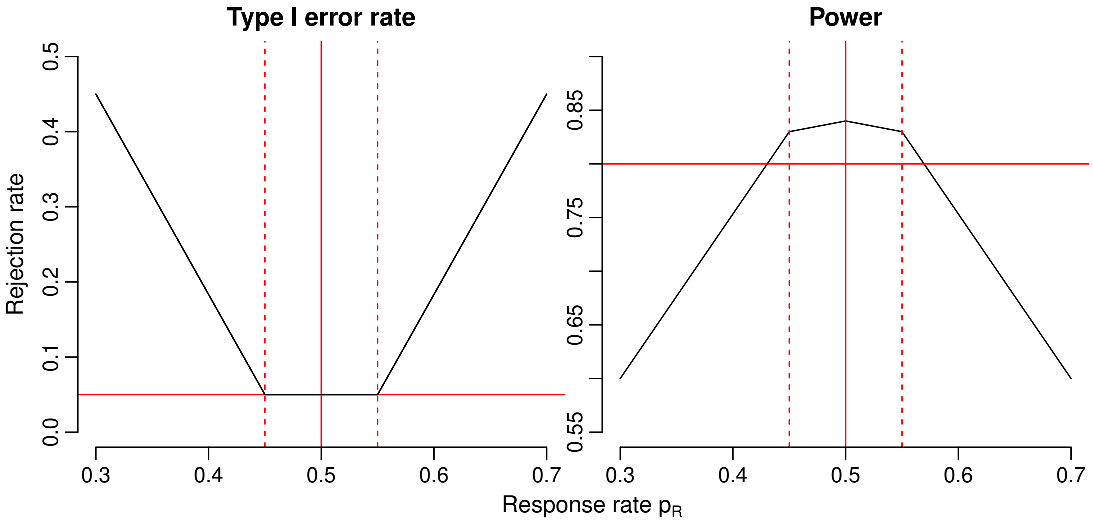

Figure 2: Desired Type I error rate and power profile (black curve). The operating characteristics of an (hypothetical) approach which is used as the benchmark are shown in red (horizontal lines). The solid vertical line gives the mean value of the prior distribution and the dotted vertical line indicate the boundaries of the interval C.

Figure 2 shows a Type I error rate and power profile that we would consider acceptable: the Type I error rate (left panel) is controlled in a neighbourhood of the mean value of the prior distribution. An inflation of the Type I error rate outside of this area is acceptable since we are highly confident that the true response rate will lie inside of the chosen neighbourhood of the mean value of the prior distribution. More formally, let $\bar{p}_H$ be the mean value of the prior distribution and $\delta \in \mathbb{R}_+$ be the parameter that defines the width of the controlled interval. Then, our aim is to control the Type I error rate for all response rates of the reference product p_R in the new study in the interval C:

$$C = [\bar{p}_H - \delta, \, \bar{p}_H + \delta].$$

A gain in power is essential in the interval C. Outside of the interval, we accept a lower power compared to the benchmark approach (see right panel in Figure 2).

We propose a novel hybrid Bayes-frequentist approach for incorporating the historical data into the test decision on equivalent efficacy. This approach has an advantage in terms of power against a standard frequentist approach that considers the data from the new trial only while controlling the Type I error rate in the interval C. This is achieved by the introduction of two switching rules and so-called response rate-dependent critical values. For the first switching rule, we check if the observed response rate of the reference product is *very* different from that of the historical data. As using historical data in the case of a strong prior-data conflict is not desirable both in terms of power and in terms of the Type I error rate, the historical data are ignored in this case. More formally, let $\hat{p}_R$ be the observed response rate of the reference product in the new study and $\bar{p}_H$ be the mean value of the prior distribution. If

$$|\bar{p}_H - \hat{p}_R| > \gamma_1,$$

we ignore the historical data. The tuning parameter $\gamma_1 \in [0{,}1]$ has to be pre-specified. The second switching rule aims to give an advantage in situations in which the response rates of the biosimilar and the reference product are *very* similar in the new study: in these situations, the estimate of the response rate of the reference product might be pulled away from the estimated response rate of the biosimilar by the historical data making the biosimilar and the reference product appear to be more different than they actually are. We compare the Bayesian success criterion with a *lower* critical value making it easier to reject in these situations. Let $\hat{p}_T$ be the observed response rate of the biosimilar in the new study and let $\gamma_2 \in [0{,}1]$ be a tuning parameter. Then, if

$$|\hat{p}_T - \hat{p}_R| < \gamma_2,$$

we compare the Bayesian success criterion with a fixed value $\bar{c} \in [0{,}1]$ which is also a tuning parameter that has to be pre-specified. The response rate-dependent critical values are motivated by the following: if an informative prior is used and combined with the observed data using Bayes' theorem, the Type I error rate is not constant for all response rates of the reference product in the new study. There are situations in which the approach is too liberal and situations in which the approach is too conservative. By using different critical values for different response rates, we aim to flatten the Type I error rate profile and to make it as constant as possible over the complete parameter space. The response rate-dependent critical values are determined by functions c_1 and c_2 which map the estimated response rate of the reference product to the critical value, i.e.,

$$c_1,\, c_2 : [0{,}1] \rightarrow [0{,}1].$$

Choosing the functions c_1 and c_2 without any assumptions on the functional form is not feasible. Therefore, we assume a logistic function with the parameters L (the minimal value of the function), U (the difference between the minimal and maximal value of the function), x_0 (the sigmoid's midpoint on the x-axis) and k (the steepness of the curve). In addition, we assume that c_1 and c_2 are complements of each other, i.e.,

$$c_1(x) = L + \frac{U}{1 + \exp(-k(x - x_0))} \text{ and } c_2(x) = L + \frac{U}{1 + \exp(k(x - x_0))}.$$

The main technical challenge of the proposed approach is the determination of the optimal parameters of the functions c_1 and c_2 (L, U, x_0, k) and the optimal tuning parameters γ_1,γ_2 and $\bar{c}$. A major simplification of this task is possible due to the discrete nature of the problem which allows the calculation of the exact Type I error rates and the exact values for power without using simulation. For that, we calculate the test decision for all combinations of numbers of responders under reference and test treatment, r_R and r_T, that can be observed in the new study. The test decision assuming the observed values r_R and r_T are denoted by d_{r_T,r_R} which is a binary variable with the value 1 if the test decision is for the alternative and 0 otherwise. For example, for $n = 150$ subjects per group in the new study, it is necessary to evaluate the test decision d_{r_T,r_R} for $151^2 = 22801$ scenarios. Finally, we combine the test decision with the probabilities that a specific number of responders under reference and test treatment is observed and these probabilities are denoted by $P(X = r_R)$ and $P(Y = r_T)$, respectively. A binomial distribution with the parameters p_R or p_T and the sample size of n subjects per group is used for calculating the probabilities $P(X = r_R)$ and $P(Y = r_T)$ for a specific scenario. This leads to the exact rejection rate:

$$r = \sum_{r_R=0}^{n} \sum_{r_T=0}^{n} P(X = r_R) P(Y = r_T) d_{r_T,r_R}.$$

However, the determination of the optimal parameters is challenging even if the exact rejection rates are used. In the contributed paper, we propose an algorithm that can be used for identifying a reasonable set of parameters. We also explain how the approach can be manually fine-tuned using a case study.

In a simulation study, we confirm that this approach leads to the desired profiles of the Type I error rate and power that are displayed in Figure 2. This is shown for one example in Figure 3: the Type I error rate is controlled within the interval C which is indicated by the vertical dotted lines. Outside of the interval, we observe an inflation of the Type I error rate, but this is acceptable. In terms of power, we gain in the interval C more than 5% power compared to the so-called two-one-sided-test (TOST) approach (Schuirmann, 1987) which is the standard frequentist approach in this setting and considers the data in the new study only (for details, see the contributed paper). Most power is gained in the situation with no prior-data conflict.

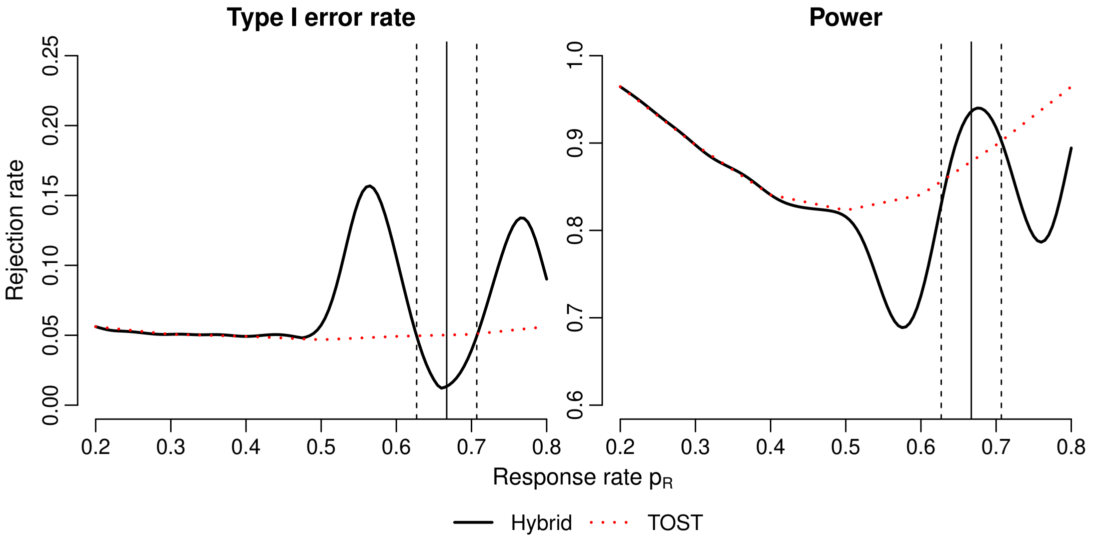

Figure 3: Operating characteristics for the novel hybrid approach and the TOST approach (the frequentist approach which considers the data in the new study only). The mean value of the prior distribution is indicated by the vertical solid line, the vertical dotted lines indicate the boundaries of the interval C. The displayed Type I error rate is the maximum of the two limiting scenarios of the null hypothesis: in one situation, the response rate of the reference product is larger than the response rate of the test product and in the other situation, it is the opposite. The absolute difference of the response rates under test and reference treatment is Δ (the equivalence margin) under the null hypothesis (left panel) and 0 under the alternative hypothesis (right panel).

Chapter 5

Efficient designs for the estimation of self and mixed-carryover effects

5.1 Contributed material

Mielke, J. and Kunert, J. (2018): Universally optimal crossover designs for the estimation of mixed-carryover effects with an application to biosimilar development. *SFB 823, Discussion paper*, 18 (3). DOI: 10.17877/DE290R-18786.

Kunert, J. and Mielke, J. (2018): Efficient designs for the estimation of mixed and self carryover effects. *SFB 823, Discussion paper,* 18 (8). DOI: 10.17877/DE290R-18820.

Co-author's contribution: Joachim Kunert supervised the research and gave advice on the presentation of the material. In Kunert and Mielke (2018), Joachim Kunert derived the adaptation of the Kushner method and the efficient designs for numbers of periods p which can be written as $p = 1 \bmod 4$. The derivation of the upper bound for the A-criterion was joint work.

5.2 Key results

This chapter deals with design considerations for the estimation of so-called mixed and self-carryover effects. This term was first introduced by Afsarinejad and Hedayat (2002): they suggested replacing the usual, first-order carryover effect which only depends on the treatment in the immediately previous period by two different carryover effects per treatment. These effects depend both on the treatment in the immediately previous period and on the treatment in the current period. A so-called mixed-carryover effect occurs if the treatments in periods k and $k-1$ differ and a self-carryover effect occurs if the treatments in periods k and $k-1$ are the same. The results presented in this chapter build up upon work of Kunert and Stufken (2002, 2008) who studied optimal designs for the estimation of the direct treatment effects in the presence of mixed and self-carryover effects which were considered by the authors to be nuisance parameters. Here, we aim to estimate mixed and self-carryover effects with the highest precision and now the treatment effect serves as one of the nuisance parameters.

For introducing the idea of self and mixed-carryover effects more formally, we assume that the response $y_{u,r}$ of subject u ($u = 1, \ldots ,n$) in period r ($r = 1, \ldots ,p$) can be written as (Kunert and Stufken, 2002)

$$y_{u,r} = \begin{cases} \alpha_u + \beta_r + \tau_{d(u,r)} + \rho_{d(u,r-1)} + e_{u,r} & \text{, if } d(u,r) \neq d(u,r-1) \\ \alpha_u + \beta_r + \tau_{d(u,r)} + \chi_{d(u,r-1)} + e_{u,r} & \text{, if } d(u,r) = d(u,r-1) \end{cases}, \tag{5.1}$$

where $d(u,r)$ gives the treatment applied to subject u in period r, α_u is the subject effect of subject u, β_r is the period effect in period r, τ_i is the direct treatment effect of treatment i ($i = 1, \ldots ,t$), ρ_i is the mixed-carryover effect of treatment i and χ_i is the self-carryover effect of treatment i. No carryover effect occurs in the first period, i.e., $\rho_{d(u,0)} = \chi_{d(u,0)} = 0$. The residual error $e_{u,r}$ is assumed to be independent and identically distributed with expectation 0 and variance σ^2. We focus on the case in which two treatments are considered (Test – T, Reference – R; $t = 2$) and assume at least 3 periods.

For determining efficient designs, we denote the set of all designs d with t treatments, n subjects and p periods as $\Omega_{t,n,p}$. Using the notation of Kunert and Stufken (2002), we define the matrices $\mathbf{U} = \mathbf{I}_n \otimes \mathbf{1}_p$ (subject effect), $\mathbf{P} = \mathbf{1}_n \otimes \mathbf{I}_p$ (period effect), $\mathbf{T}_d$ (treatment effect), $\mathbf{M}_d$ (mixed-carryover effect) and $\mathbf{S}_d$ (self-carryover effect), where $\otimes$ denotes the Kronecker product, $\mathbf{I}_m$ is the identity matrix of dimension m and $\mathbf{1}_m$ is a vector of length m that only contains the entries 1. The model in vector notation can be written as

$$\mathbf{y} = \mathbf{T_d}\tau + \mathbf{S_d}\chi + \mathbf{M_d}\rho + \mathbf{U}\alpha + \mathbf{P}\beta + \mathbf{e},$$

where τ is the vector of treatment effects, χ is the vector of self-carryover effects and ρ is the vector of mixed-carryover effects. Also, α, β and $\mathbf{e}$ are the vectors of subject effects, period effects and residual errors, respectively.

In the two contributed papers, we discuss efficient designs for the estimation of mixed and/or self-carryover effects. In Mielke and Kunert (2018) (Section 5.2.1), we focus on the determination of universally optimal designs for the estimation of mixed-carryover effects. Universally optimal is a term introduced by Kiefer (1975). If all information matrices $\mathbf{C}_d$ have row sums and column sums equal to 0, then a design d^* is universally optimal if its information matrix $\mathbf{C}_{d^*}$ is completely symmetric and the design d^* maximises the trace of $\mathbf{C}_d$ over all $d \in \Omega_{t,n,p}$. A matrix $\mathbf{A}$ is called completely symmetric if it can be written in the form

$$\mathbf{A} = a\mathbf{I} + b\mathbf{1}\mathbf{1}^T,$$

where a and b are real numbers. The information matrix for the estimation of the mixed-carryover effects, which uses, amongst others, the self-carryover effects as nuisance parameters, is given by

$$\tilde{\mathbf{C}}_d^{(1)} = \mathbf{M}_d^T \omega^\perp([\mathbf{P},\mathbf{U},\mathbf{T}_d,\mathbf{S}_d])\mathbf{M}_d,$$

where $\omega^\perp(\mathbf{A}) = \mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^-\mathbf{A}^T$ is the projection on the space of all vectors that are orthogonal to the columns of $\mathbf{A}$ and where $\mathbf{A}^T$ is the transpose and $\mathbf{A}^-$ is a g-inverse of $\mathbf{A}$.

In Kunert and Mielke (2018) (Section 5.2.2), we focus on the simultaneous estimation of *all* carryover effects,

$$\delta = \begin{pmatrix} \chi \\ \rho \end{pmatrix}.$$

The information matrix for the joint estimation of mixed and self-carryover effects is given by

$$\tilde{\mathbf{C}}_d^{(2)} = [\mathbf{S}_d,\mathbf{M}_d]^T \omega^\perp([\mathbf{P},\mathbf{U},\mathbf{T}_d])[\mathbf{S}_d,\mathbf{M}_d].$$

The performance of the designs for the joint estimation of self and mixed-carryover effects is assessed using the A-criterion. A design d^* is called A-optimal, if the trace of the inverse of the information matrix is minimal. An A-optimal design is therefore the design with the minimal average variance of the estimators (Rodrigues and Iemma, 2014). The trace of a matrix $\mathbf{A}$ is the sum of its eigenvalues. Therefore, if $\lambda_1,\ldots,\lambda_k$ are the k non-zero eigenvalues of the matrix $\tilde{\mathbf{C}}_d^{(2)}$, the A-criterion can be equivalently expressed as

$$\tilde{\phi}_A = \sum_{i=1}^{k} \frac{1}{\lambda_i},$$

or, as we prefer to maximise the criterion instead of minimising it, as

$$\phi_A = \frac{1}{\sum_{i=1}^{k} \frac{1}{\lambda_i}}. \tag{5.2}$$

In both contributed papers, we realise that the information matrices have row and column sums equal to 0 which allows the multiplication with a matrix $\mathbf{B_q} = \omega^{\perp}(\mathbf{1_q})$ without changing the result:

$$\begin{aligned}\tilde{\mathbf{C}}_d^{(1)} &= \mathbf{B}_2\tilde{\mathbf{C}}_d^{(1)}\mathbf{B}_2,\\ \tilde{\mathbf{C}}_d^{(2)} &= \mathbf{B}_4\tilde{\mathbf{C}}_d^{(2)}\mathbf{B}_4.\end{aligned}$$

5.2.1 Mielke and Kunert (2018)

Since the row and column sums of $\tilde{\mathbf{C}}_d^{(1)}$ are 0, the concept of universally optimal is applicable here. For identifying a design which is universally optimal, we need to maximise the trace of the information matrix $\tilde{\mathbf{C}}_d^{(1)}$. The strategy presented in this contributed paper for identifying the universally optimal designs follows the ideas of Kunert and Stufken (2002, 2008) and consists of two main steps: first, a matrix $\mathbf{C}_d^{(1)}$ that is larger in the Loewner sense than the information matrix $\tilde{\mathbf{C}}_d^{(1)}$ is derived. For that, we use Proposition 2 of Kunert (1983) which claims that

$$\tilde{\mathbf{C}}_d^{(1)} = \mathbf{B}_2\mathbf{M}_d^T\omega^{\perp}([\mathbf{P},\mathbf{U},\mathbf{T}_d,\mathbf{S}_d])\mathbf{M}_d\mathbf{B}_2 \leq \mathbf{B}_2\mathbf{M}_d^T\omega^{\perp}([\mathbf{U},\mathbf{T}_d,\mathbf{S}_d])\mathbf{M}_d\mathbf{B}_2 = \mathbf{C}_d^{(1)}, \text{ say,}$$

with equality if and only if

$$(\mathbf{M}_d\mathbf{B}_2)^T\omega^{\perp}([\mathbf{U},\mathbf{T}_d,\mathbf{S}_d])\mathbf{P} = 0. \tag{5.3}$$

Since we can show that the condition stated in Equation (5.3) holds, an upper bound for the trace of $\mathbf{C}_d^{(1)}$ is determined next and a class of designs is identified that reaches this bound. It is important to note that there is always a dual-balanced design among the optimal designs. A sequence s is called dual to a sequence s' if sequence s can be changed to sequence s' by interchanging the two treatments (e.g., TRTR, dual sequence: RTRT). A design d is dual-balanced if it uses sequence s exactly as often as sequence s'. Therefore, without loss of generality, we focus on dual-balanced designs. Let $x,y \in \mathbb{R}$ and l be an equivalence class of sequences which consists of a specific sequence s and its dual-balanced sequence s'. The set of all equivalence classes l is denoted by L. We are in the two-treatment case and we consider only dual-balanced designs and that is why there are exactly 2^{p-1} different equivalence classes. We define

$$\pi_d = (\pi_d(1),\ldots,\pi_d(2^{p-1}))^T$$

as a vector of length 2^{p-1} which gives the proportion of sequences of the design d that belong to the lth equivalence class with

$$\pi_d(l) \geq 0 \text{ and } \sum_{l=1}^{2^{p-1}} \pi_d(l) = 1.$$

Then, we use from Kunert and Stufken (2008) that for any design $d \in \Omega_{2,n,p}$,

$$tr\left(\mathbf{C}_d^{(1)}\right) \leq n \cdot \min_{x,y} \sum_{l=1}^{2^{p-1}} \pi_d(l) h_l(x,y),$$

with

$$h_l(x,y) = c_{11}(l) + 2xc_{12}(l) + x^2 c_{22}(l) + 2yc_{13}(l) + y^2 c_{33}(l) + 2xyc_{23}(l),$$

where the terms c_{ij} are derived in the contributed paper (Mielke and Kunert, 2018). Due to the properties of π_d (non-negative, $\sum_{l=1}^{2^{p-1}} \pi_d(l) = 1$), it is clear that

$$\sum_{l=1}^{2^{p-1}} \pi_d(l) h_l(x,y) \leq \max_{l \in L} h_l(x,y),$$

and therefore

$$tr\left(\mathbf{C}_d^{(1)}\right) \leq n \min_{x,y} \max_{l \in L} h_l(x,y).$$

The main technical difficulty is the identification of numbers x^* and y^* and optimal classes of sequences $l^* \in L$ such that

$$h_{l^*}(x^*,y^*) = \min_{x,y} \max_{l \in L} h_l(x,y).$$

Calculations shown in the contributed paper lead to the conclusion that for sequences with an even number of periods p, it is optimal to use different treatments in the first and in the last period and to switch after each period, e.g., for six periods, it is optimal to include the sequences TRTRTR and RTRTRT. For an odd number of periods, it is optimal to end with the same treatment that was used in the first period and to switch after each period, e.g., for five periods, it is optimal to include the sequences TRTRT and RTRTR. Since the study design has to be dual-balanced, the number of subjects in both sequences has to be the same.

We also discuss the inclusion of dummy treatments (periods with no treatment or placebo treatment). This investigation is motivated by a finding by Kunert and Stufken (2008) who showed that adding additional periods does not improve the design in a relevant way, but the inclusion of dummy treatments increases the precision of the estimators. We achieve comparable results also in our cases. However, it is important to emphasise that these findings highly depend on the model assumptions, e.g., the assumption that only the treatment in period $k-1$ is relevant in period k and earlier treatments have already been washed out.

5.2.2 Kunert and Mielke (2018)

For the joint estimation of self and mixed-carryover effects, it is not possible to determine universally optimal designs because the information matrix $\tilde{\mathbf{C}}_d^{(2)}$ is, even for efficient designs, not completely symmetric. That is why we focus on A-optimality instead. Since the optimality criterion used in Kunert and Stufken (2002) as well as in Kunert and Stufken (2008) is universal optimality and not A-optimality, it is not possible to use the same ideas and simply adjust their strategy to our scenario. This makes the task presented in Kunert and Mielke (2018) more challenging than the one in Mielke and Kunert (2018).

It is well-known that the information matrix $\tilde{\mathbf{C}}_d^{(2)}$ is in general not linear in its sequences, i.e.,

$$\tilde{\mathbf{C}}_d^{(2)} \neq n \sum_s \pi_d(s) \tilde{\mathbf{C}}_{d,s}^{(2)},$$

where $\pi_d(s)$ is the proportion of subjects allocated to sequence s and $\tilde{\mathbf{C}}_{d,s}^{(2)}$ is the design matrix of sequence s. Kushner (1997) proposed a methodology that splits the information matrix into matrices $\tilde{\mathbf{C}}_{dij}^{(2)}$ which are linear in the sequences and used this decomposition for deriving optimal designs. However, we note that the assumptions for the Kushner method are not fulfilled in our setting (not all matrices $\tilde{\mathbf{C}}_{dij}^{(2)}$ are square matrices). Therefore, in the first step of the paper, an adaptation of the Kushner method to our setting is derived. With this result, it is possible to prove that an upper bound for the A-criterion is given by

$$\phi_A(d) \leq n \frac{(p-1)(2p^2+2p-1)}{4p(2p^2+6p+3)}.$$

We determine efficient designs for $p=3$ and $p=1 \bmod 4$. For $p=3$, there are only eight possible sequences and since there is always a dual-balanced design among the optimal designs, we only need to consider four out of these eight sequences. Then, it is possible to determine terms $a, \dots, f$ such that

$$\frac{1}{n}\tilde{\mathbf{C}}_d^{(2)} = \begin{pmatrix} a & b & e & f \\ b & a & f & e \\ e & f & c & d \\ f & e & d & c \end{pmatrix},$$

and the non-zero eigenvalues of $\frac{1}{n}\tilde{\mathbf{C}}_d^{(2)}$ are then given by

$$\begin{aligned}
\lambda_1 &= \frac{a-b+c-d}{2} + \sqrt{(e-f)^2 + \left(\frac{c-d-a+b}{2}\right)^2}, \\
\lambda_2 &= \frac{a-b+c-d}{2} - \sqrt{(e-f)^2 + \left(\frac{c-d-a+b}{2}\right)^2}, \\
\lambda_3 &= \frac{a+b+c+d}{2} - e - f.
\end{aligned}$$

Using these results, it is possible to identify with a numerical search the optimal allocation ratios to the four sequences. We find that it is optimal to allocate 9.51% of the subjects to sequence TTT and its dual-balanced sequence RRR, 10.33% of the subjects to sequence RTT and TRR, 16.84% of the subjects to RTR and TRT and 13.32% of the subjects to RRT and TTR. The A-criterion for this design d_1 is for n subjects in the study given by

$$\phi_A(d_1) = 0.0636n.$$

In practice, design d_1 is not attractive because it takes a high number of subjects to construct a design with these proportions. A design d_2 with equal allocations to the sequences has an A-criterion of

$$\phi_A(d_2) = 0.0628n,$$

and is therefore close to the A-criterion of the optimal design and might be preferred in practice.

For $p = 1 \bmod 4$, we find that a study design with the sequences

$$\begin{aligned}
s_1 &= [T\ R\ R\ T\ T\ R\ R \dots], \\
s_2 &= [R\ T\ T\ R\ R\ T\ T \dots], \\
s_3 &= [T\ T\ R\ R\ T\ T\ R \dots], \\
s_4 &= [R\ R\ T\ T\ R\ R\ T \dots],
\end{aligned}$$

and an equal number of subjects allocated to each of these sequences achieves an efficiency of at least

$$E(p) = \frac{2p^3 + 6p^2 + 3p}{2p^3 + 8p^2 + 5p - 3},$$

where p is the number of periods. For $p = 5$, the efficiency is 0.88, for $p = 10$ it is already 0.92 and $E(p)$ converges to 1 if $p \to \infty$.

The analysis of mixed and self-carryover effects is closely related to the assessment of switchability (i.e., can patients switch between the biosimilar and its reference product without any impact on the treatment success) which is discussed in Chapter 6. In Section 6.3, we discuss in detail the performance of the study designs which were derived in Kunert and Mielke (2018) for one of the proposed methodologies for testing for switchability and confirm that the derived study designs have good characteristics for the assessment of switchability in practice.

Chapter 6

The assessment of switchability of biosimilars

6.1 Contributed Material

Mielke, J., Woehling, H. and Jones, B. (2018d): Longitudinal assessment of the impact of multiple switches between a biosimilar and its reference product on efficacy parameters. *Pharmaceutical Statistics*, 17 (3), 231–247.

Co-authors' contribution: Byron Jones gave advice during the development of the proposed methodologies and helped with the presentation of the material. Heike Woehling gave advice on practical considerations and on the case study.

6.2 Key results

Patients, physicians and health care providers in Europe have more than ten years of experience with the use of biosimilars in practice. Nonetheless, there is still uncertainty if patients who were already taking the reference product at the time when the biosimilar is approved should switch to the biosimilar or if even multiple switches between a biosimilar and its reference product are acceptable. That would allow a substitution of the reference product with the biosimilar at pharmacy level without the approval of the prescribing doctor which is accepted for generics in many countries already. The higher complexity of biologics and the limited experience with biosimilars raise doubts if this should be introduced for biosimilars as well. One way to reduce the uncertainty would be to conduct a study specifically focussing on the question if switching influences the efficacy. To date, there are only a few proposals for a statistical methodology for assessing switchability published (e.g., Zheng et al., 2017; Chow et al., 2013; Belleli et al., 2015) and none of these methodologies is tailored to assess the impact of multiple switches on normally distributed efficacy endpoints with a formal statistical testing procedure. In the contributed paper, we develop three statistical tests for switchability and assess their properties in simulation studies.

Before any statistical test can be developed, it is necessary to define the null and alternative hypothesis that are to be tested. In the considered setting, the null hypothesis refers to situations in which switching between the biosimilar and its reference product reduces the efficacy of the treatment whereas the alternative hypothesis relates to situations in which switching does not have any relevant negative impact on the treatment efficacy. For translating these ideas into statistical terms, we use a linear mixed-effects model

$$y_{i,j,k} = p_k + t_{a(i,j,k)} + I_{a(i,j,k-1),a(i,j,k)} + \kappa_{i,j} + \epsilon_{i,j,k}, \tag{6.1}$$

where the response of the ith subject ($i = 1,\dots,n$) in period k ($k = 1,\dots,p$) and sequence j ($j = 1,\dots,q$) is denoted by $y_{i,j,k}$. We assume that the response depends on the period effect p_k, the effect of the treatment $t_{a(i,j,k)}$ (with $a(i,j,k) = T$ if the test treatment, T, was given to subject i in sequence j and period k and $a(i,j,k) = R$ defined analogously for the reference treatment, R), a switching effect $I_{a(i,j,k-1),a(i,j,k)}$ (see below), the subject specific effect $\kappa_{i,j}$ which is constant over time and the residual error $\epsilon_{i,j,k}$. The subject effect and the residual error each follow a normal distribution with mean value equal to 0 and variance σ_s^2 and σ_e^2, respectively. The switching effect $I_{a(i,j,k-1),a(i,j,k)}$ depends only on the treatment in the immediately previous period and in the current period, i.e., we assume that the length of the periods is long enough such that the effects of earlier treatments have already been washed out. In addition, we assume that a switch from T to R leads to the same switching effect I_{TR} as a switch from R to T which is denoted by I_{RT}. If subjects do not switch, i.e., the treatment in period $k-1$ is the same as the treatment in period k, the switching effect is set to 0. We aim to test for switchability with the hypotheses

$$H_0 : |I_{TR}| = |I_{RT}| \geq \Delta \text{ vs } H_1 : |I_{TR}| = |I_{RT}| < \Delta,$$

where $\Delta \in \mathbb{R}_+$ is the pre-specified equivalence margin.

In the contributed paper, we propose three statistical tests with significance level α for assessing switchability. The first one (the estimation method) is based on estimating the effect of switching and is therefore related to the estimation of mixed and self-carryover effects that was discussed in Chapter 5. However, we make some changes so that the methodology is fully tailored for the test for switchability and discuss the differences between the models and its implications on the choice of the design in Section 6.3. For the estimation method, we fit the linear mixed-effects model

$$y_{i,j,k} = p_k + t_{a(i,j,k)} + c_{(j,k,k-1)} + \kappa_{i,j} + \epsilon_{i,j,k}, \tag{6.2}$$

which uses the same notation as the model stated in Equation (6.1). The only difference is that the switching effect in Equation (6.1), $I_{a(i,j,k-1),a(i,j,k)}$, is replaced by a carryover effect, $c_{(j,k,k-1)}$, which is a categorical factor with three levels (c_0, c_1, c_2): we use the first level c_0 for all observations in the first period in which no switching effect is expected and for subjects who do not change treatment from period $k-1$ to period k (continuous treatment with T or R). For subjects who are switching from T to R, the second level c_1 is used and for subjects switching from R to T, the third level c_2 is used. In total, the carryover effect in the proposed model distinguishes between "no switch", "switch R to T" and "switch T to R".

Compared to the model stated in Equation (6.1) which is used for defining the null and alternative hypothesis, we therefore allow in the estimation method for different effects for a switch from T to R and for a switch from R to T. Thus, we make the estimation method more flexible and in particular robust against deviation from the assumption that a switch from R to T leads to the same effect as a switch from T to R. We use c_0 as the reference level, i.e., c_1 represents the effect of a switch from T to R compared to continuous treatment and c_2 gives the effect of a switch from R to T also compared to continuous treatment (linear contrasts). As c_0 is used as the reference category, this effect is set to 0. The test decision is made based on comparing the estimated effects c_1 and c_2 to the α-quantiles of a multivariate normal distribution. The mean value and the variance-covariance matrix of the distribution under the null hypothesis are derived in the contributed paper.

The two other proposed methodologies are based on the idea of comparing predictions and observations. For that, we assume a study design in which the sequences can be split into switching sequences (subjects switch multiple times between the reference product (R) and the biosimilar (T), e.g., treatment sequence TRTRTR, denoted by s) and non-switching sequences (continuous treatment with the biosimilar or the reference product, denoted by ns). Then, the longitudinal observations are split into two datasets where the first dataset (the modelling dataset, M) consists only of one observation per subject which is known to be free of any switching effects and the second dataset (the evaluation dataset, E) consists of all other observations. We fit a linear model to the modelling dataset,

$$y_{i,j,k} = p_k + t_{a(i,j,k)} + \epsilon_{i,j,k},$$

where the same notation as introduced above is used. Next, the responses for the observations in the evaluation dataset are predicted using this model. The prediction error for each observation, which is given by the difference between the observed response, $y_{i,j,k}$, and the predicted response, $\hat{y}_{i,j,k}$, is calculated:

$$\tilde{y}_{i,j,k} = y_{i,j,k} - \hat{y}_{i,j,k},\ (i,j,k) \in E.$$

Both approaches compare the prediction errors of subjects in the switching sequences with the prediction errors of subjects in the non-switching sequences. The so-called quadratic prediction method is based on the mean squared differences (MSDs) of the prediction errors in the switching and non-switching sequences which is given by

$$MSD_{ns} = \frac{1}{q_{ns} \cdot n \cdot (p-1)} \sum_{(i,j,k) \in E,ns} \tilde{y}^2_{i,j,k} \text{ and } MSD_s = \frac{1}{q_s \cdot n \cdot (p-1)} \sum_{(i,j,k) \in E,s} \tilde{y}^2_{i,j,k}.$$

The parameters q_{ns} and q_s denote the number of non-switching and switching sequences, respectively. The test statistic uses the difference of the MSDs:

$$T_{MSD} := MSD_s - MSD_{ns}.$$

Small values indicate that the prediction errors are not larger for the switching sequences than for the non-switching sequences and this is the situation in which the null hypothesis is to be rejected and switchability can be claimed. Therefore, we reject the null hypothesis if t_{MSD}, the observed value of T_{MSD}, is smaller than the α-quantile of the distribution of T_{MSD} under the null hypothesis. The derivation of this distribution is discussed in the contributed paper.

The third proposed methodology (distribution prediction method) compares the distribution of the prediction errors in the switching and non-switching sequences with the Kolmogorov-Smirnov distance (Massey Jr., 1951) which is given for two empirical distribution functions $F_n^{(1)}$ and $F_n^{(2)}$ by

$$D = \max_{z \in \mathbb{R}} |F_n^{(1)}(z) - F_n^{(2)}(z)|.$$

For a vector of observations, $x = (x_1, \ldots, x_n)$, the empirical distribution function is defined as

$$F_n(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{x_i \leq z},$$

with **1** being the indicator function. For the distribution prediction method, we estimate the functions F_n for the switching and non-switching sequences. It is important to emphasise that these are not ordinary distribution functions in our setting because the prediction errors are not independent and also do not follow the same distribution. This is why we call them *estimated functions* in the following. Let the estimated function for the switching arms be denoted by $F_n^{(s)}$ and the estimated function for the non-switching arms be denoted by $F_n^{(ns)}$. Then, the test statistic is defined as

$$T_{DP} = \max_{z \in \mathbb{R}} |F_n^{(s)}(z) - F_n^{(ns)}(z)|.$$

If the biosimilar and its reference product are switchable, the observed test statistic will be small because the distributions of the prediction errors will be comparable (the alternative hypothesis). Therefore, the null hypothesis is rejected if the observed value t_{DP} of the test statistic T_{DP} is smaller than the α-quantile of the distribution of T_{DP} under the null hypothesis. The distribution of the test statistic T_{DP} is approximated by simulations and details are provided in the contributed paper.

We compare the three proposed methodologies in a simulation study assuming the model which is given in Equation (6.1) for the simulation of the datasets. It is shown that all three methods preserve the desired Type I error rate. The estimation method has the highest power in all settings. However, that was expected because the estimation method directly targets the change in mean and we use the same model for the simulation of the datasets that is also used for the estimation of the switching effects. The distribution prediction method has also high power, especially in situations in which the true effect of $I_{RT} = I_{TR}$ is small. These are the situations we consider most important in practice. The quadratic prediction method has substantially lower power if the variance of the subject effect and of the residual error is high and the sample size is small. Therefore, it would be necessary to enroll more subjects if this method is to be used.

As a sensitivity analysis, we also consider a situation in which the switching effects I_{RT} and I_{TR} are not fixed, but random effects with mean equal to 0 and variance σ_I^2. This setting refers to a situation in which switching has on average no impact on the efficacy of the treatment, but patients who are switching experience an unstable treatment response which might be disadvantageous for the patient and therefore a situation in which the null hypothesis should not be rejected (i.e., switchability should not be claimed). In simulations, we show that the estimation method is not sensitive to this deviation from switchability. The quadratic prediction method can detect this setting easily while the distribution prediction method is less sensitive than the quadratic prediction method but more sensitive than the estimation method. In total, we conclude that the preference for a specific method depends on the expectation and experience with the impact of switching. If, for example, it is clear that the only impact could be on the mean value, it would be best to choose the estimation method. However, if there is uncertainty whether also other deviations from switchability can occur, it might be better to choose the quadratic prediction method which can detect various deviations from the switchable setting. As the sample size for the quadratic prediction method has to be much higher compared to the other methods in some situations, and the high sample size might not be feasible due to practical reasons, the distribution prediction method might be a good compromise.

6.3 Design considerations for the estimation method and the performance of designs used in practice for the joint estimation of self and mixed-carryover effects

In Chapter 5, we derived efficient designs for the joint estimation of mixed and self-carryover effects (Kunert and Mielke, 2018) using the linear model (see Equation (5.1), Section 5.2)

$$y_{u,r} = \begin{cases} \alpha_u + \beta_r + \tau_{d(u,r)} + \rho_{d(u,r-1)} + e_{u,r} & \text{if } d(u,r) \neq d(u,r-1) \\ \alpha_u + \beta_r + \tau_{d(u,r)} + \chi_{d(u,r-1)} + e_{u,r} & \text{if } d(u,r) = d(u,r-1) \end{cases},$$

where $d(u,r)$ gives the treatment of subject u ($u = 1, \ldots, n$) in period r ($r = 1, \ldots, p$), α_u is the subject effect of subject u, β_r is the period effect in period r, τ_i is the treatment effect of treatment i ($i = 1, \ldots, t$), ρ_i is the mixed-carryover effect of treatment i, χ_i is the self-carryover effect of treatment i and $e_{u,r}$ is the residual error of subject u in period r. The estimates of the mixed and self-carryover effects can be used for the assessment of switchability since these quantities give the effect of continuous treatment with the biosimilar or the reference product (self-carryover effects) and the effect of switching from the biosimilar to its reference product or vice versa (mixed-carryover effects).

The estimation method proposed in Section 6.2 (Mielke et al., 2018d) implicitly uses the idea of comparing the mixed and the self-carryover effects by estimating the effect of switching (the mixed-carryover effects) compared to continuous treatment with the test or the reference product (the self-carryover effects). However, it is important to note that the model used for the estimation method (Equation (6.2), see Section 6.2) differs from the model with self and mixed-carryover effects (Equation (5.1), see Section 5.2) in several aspects: (1) the fixed subject effect is replaced with a random subject effect, (2) the two self-carryover effects are combined into one effect and there is no distinction between continuous treatment with T or with R, (3) it is assumed in the estimation method that the carryover effect in the first period is the same as the self-carryover effect of T or R. In order to show the close relationship between the two models, we first neglect the differences (1) and (3) and assume that all self and mixed-carryover effects appear equally often in the study design. Then, the coefficients c_1 and c_2 (the effect of switching from R to T and from T to R, respectively, compared to continuous treatment with T or R) which are used in the estimation method can be written as linear contrasts of the self-carryover effects χ_i and mixed-carryover effects ρ_i:

$$c_1 = \rho_T - \frac{1}{2}(\chi_R + \chi_T),$$
$$c_2 = \rho_R - \frac{1}{2}(\chi_R + \chi_T).$$

Therefore, due to the close relationship between the estimation method and the self and mixed-carryover effects, one might wonder if the efficient designs for the estimation of the mixed and self-carryover effects that were derived in Chapter 5 are also useful if the planned analysis of the study is the estimation method. This will be explored in the following. For that, we assume the model of the estimation method which is introduced in Equation (6.2) in Section 6.2 (Mielke et al., 2018d), but replace the random subject effects with fixed subject effects.

Using the notation introduced in Kunert and Mielke (2018), we define the design matrices $\mathbf{U}$ (fixed subject effect), $\mathbf{P}$ (period effect) and $\mathbf{T}_d$ (treatment effect). Instead of defining the mixed and self-carryover effects as in Chapter 5, we introduce the design matrix of a carryover effect $\tilde{c}_{(j,k,k-1)}$ and denote it with $\mathbf{C}_o$. This carryover effect has three levels, i.e.,

$$\tilde{c}_{(j,k,k-1)} = \begin{cases} \tilde{c}_0 & \text{, if the treatments in period } k \text{ and } k-1 \text{ is the same or } k=1, \\ \tilde{c}_1 & \text{, if the treatment in period } k \text{ was R and in period } k-1 \text{ was T,} \\ \tilde{c}_2 & \text{, if the treatment in period } k \text{ was T and in period } k-1 \text{ was R.} \end{cases}$$

Then, the information matrix of the carryover effect $\tilde{c}_{(j,k,k-1)}$, which is taking into account the nuisance parameters period, subject and treatment, is given by

$$\mathbf{C}_e = \mathbf{C}_o^T \omega^{\perp}([\mathbf{U},\mathbf{P},\mathbf{T}_d])\mathbf{C_o},$$

where $\omega^{\perp}(\mathbf{A})$ is a projection matrix as introduced in Chapter 5. Since the estimation method considers the linear contrasts between $\tilde{c}_1$ and $\tilde{c}_0$ and between $\tilde{c}_2$ and $\tilde{c}_0$, we define

$$c_1 = \tilde{c}_1 - \tilde{c}_0,$$
$$c_2 = \tilde{c}_2 - \tilde{c}_0.$$

For deriving the A-criterion of optimality (see Equation (5.2)) of the estimation method, it is necessary to define a matrix that gives these linear contrasts, i.e., a matrix $\mathbf{L} \in \mathbb{R}^{2\times 3}$ such that

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \mathbf{L} \begin{pmatrix} \tilde{c}_0 \\ \tilde{c}_1 \\ \tilde{c}_2 \end{pmatrix},$$

and this matrix $\mathbf{L}$ is given by

$$\mathbf{L} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

Then, the A-criterion can be calculated based on the information matrix

$$\mathbf{C}_d^* = (\mathbf{L}\mathbf{C}_e^- \mathbf{L}^T)^-,$$

where $\mathbf{A}^-$ is a g-inverse of a matrix $\mathbf{A}$.

As examples, we consider designs with five periods and four sequences and compare the A-criterion for the efficient design that was derived in Kunert and Mielke (2018) to the study design of the EGALITY study (Griffiths et al., 2017) and the design where the sequences can be split into switching and non-switching sequences (switching design) which is the study design that is the foundation for the work in this chapter. The sequences of the study designs are displayed in Table 1. The design used by Griffiths et al. (2017) differs from the switching design only in the last period where patients in Griffiths et al. (2017) stay on the previous treatment while patients in the switching design switch once more. Table 1 also give the A-criteria (see Equation (5.2)) assuming one subject per sequence. It is shown that the design proposed by Kunert and Mielke (2018) is not only efficient for the estimation of self and mixed-carryover effects but has also good properties for the estimation method. In contrast, the design used by Griffiths et al. (2017) and the switching design have a much lower A-criterion. Therefore, if a study is planned that is to be analysed with the estimation method, using the study design proposed by Kunert and Mielke (2018) is preferred over the study design by Griffiths et al. (2017) which is the design that was already used in practice. It should be noted that the results obtained in Mielke and Kunert (2018) are not directly applicable to this scenario because the optimal designs derived in Mielke and Kunert (2018) do not allow for the estimation of the self-carryover effects and this is required for the estimation method (see Section 6.2).

Table 1: Three study designs and their A-criteria (see Equation (5.2)) for the estimation method (Mielke et al., 2018d, assuming a fixed subject effect). A high value of the A-criterion is desirable.

Study design	Seq. 1	Seq. 2	Seq. 3	Seq. 4	A-criterion
Kunert and Mielke (2018)	RRTTR	TTRRT	TRRTT	RTTRR	1.0769
Griffiths et al. (2017)	RTRTT	TRTRR	TTTTT	RRRRR	0.3962
Switching design	RTRTR	TRTRT	TTTTT	RRRRR	0.25

Finally, one might wonder how the design proposed by Griffiths et al. (2017) performs in terms of the A-criterion if the goal of the study is to estimate the self and mixed-carryover effects in the model discussed in Kunert and Mielke (2018) (see Equation (5.1)). For that, we compare the A-criteria using the information matrix

$$\tilde{\mathbf{C}}_d^{(2)} = [\mathbf{S}_d,\mathbf{M}_d]^T \omega^{\perp}([\mathbf{P},\mathbf{U},\mathbf{T}_d])[\mathbf{S}_d,\mathbf{M}_d],$$

where $\mathbf{M}_d$ is the design matrix of the mixed-carryover effects, $\mathbf{S}_d$ is the design matrix of the self-carryover effects and all other notation is as introduced above. For further details, we refer to Chapter 5.

Table 2 shows the results: the design derived in Kunert and Mielke (2018) outperforms the design proposed by Griffiths et al. (2017) and the switching design. However, the absolute difference in the A-criterion between the designs is smaller for the estimation of the self and mixed-carryover effects (information matrix $\tilde{\mathbf{C}}_d^{(2)}$, see Table 2) than for the estimation method (information matrix $\mathbf{C}_d^*$, see Table 1).

It is important to acknowledge that the study design used by Griffiths et al. (2017) was not optimised for the estimation method or the estimation of self and mixed-carryover effects. In addition, study designs in practice are not purely chosen due to an optimality criterion, but also for operational reasons (e.g., is the required treatment sequence feasible in patients) or regulatory considerations (e.g., is there a recommended study design by regulatory authorities). However, if focussing on the A-criterion only, we point out that the design proposed in Kunert and Mielke (2018) has a higher efficiency than the design used by Griffiths et al. (2017) both if the estimation method is used or if the self and mixed-carryover effects are estimated. Therefore, the derived design might be a candidate which is worth considering if an analysis related to the estimation of self and mixed-carryover effects is planned.

Table 2: Three study designs and their A-criteria (see Equation (5.2)) for the estimation of self and mixed-carryover effects using the model which is stated in Equation (5.1) (Kunert and Mielke, 2018). A high value of the A-criterion is desirable.

Study design	Seq. 1	Seq. 2	Seq. 3	Seq. 4	A-criterion
Kunert and Mielke (2018)	RRTTR	TTRRT	TRRTT	RTTRR	0.5
Griffiths et al. (2017)	RTRTT	TRTRR	TTTTT	RRRRR	0.3434
Switching design	RTRTR	TRTRT	TTTTT	RRRRR	0.25

References

Afsarinejad, K. and Hedayat, A. S. (2002): Repeated measurements designs for a model with self and simple mixed carryover effects. *Journal of Statistical Planning and Inference*, 106 (1–2), 449–459.

Belleli, R., Fisch, R., Renard, D., Woehling, H. and Gsteiger, S. (2015): Assessing switchability for biosimilar products: modelling approaches applied to children's growth. *Pharmaceutical Statistics*, 14 (4), 341–349.

Blackstone, E. A. and Fuhr Jr, J. P. (2012): Innovation and competition: will biosimilars succeed? *Biotechnology Healthcare*, 9 (1), 24–27.

Blevins, T. C., Dahl, D., Rosenstock, J., Ilag, L. L., Huster, W. J., Zielonka, J. S., Pollom, R. K. and Prince, M. J. (2015): Efficacy and safety of LY2963016 insulin glargine compared with insulin glargine (Lantus®) in patients with type 1 diabetes in a randomized controlled trial: the ELEMENT 1 study. *Diabetes, Obesity and Metabolism*, 17 (8), 726–733.

Bretz, F., Hothorn, T. and Westfall, P. (2016): *Multiple comparisons using R*. CRC Press, Boca Raton.

CHMP (2008): Zarzio: EPAR - public assessment report. Available at `http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/000917/WC500046528.pdf` (accessed 22 Feb 2018).

CHMP (2014a): Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: non-clinical and clinical issues (revision 1). Available at `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/01/WC500180219.pdf` (accessed 22 Feb 2018).

CHMP (2014b): Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: quality issues (revision 1). Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/06/WC500167838.pdf (accessed 22 Feb 2018).

CHMP (2014c): Guideline on similar biological medicinal products (revision 1). Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/10/WC500176768.pdf (accessed 22 Feb 2018).

CHMP (2016): Lusduna: EPAR - public assessment report. Available at http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/004101/WC500219588.pdf (accessed 22 Feb 2018).

Chow, S.-C. (2013): *Biosimilars: design and analysis of follow-on biologics*. CRC Press, Boca Raton.

Chow, S.-C., Yang, L.-Y., Starr, A. and Chiu, S.-T. (2013): Statistical methods for assessing interchangeability of biosimilars. *Statistics in Medicine*, 32 (3), 442–448.

Crommelin, D., Bermejo, T., Bissig, M., Damiaans, J., Krämer, I., Rambourg, P., Scroccaro, G., Strukelj, B. and Tredree, R. (2005): Pharmaceutical evaluation of biosimilars: important differences from generic low-molecularweight pharmaceuticals. *The European Journal of Hospital Pharmacy Science*, 11 (1), 11–17.

DiMasi, J. A., Hansen, R. W. and Grabowski, H. G. (2003): The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 22 (2), 151–185.

EMA (2012a): Questions and answers on biosimilar medicines (similar biological medicinal products). Available at http://www.medicinesforeurope.com/2012/09/27/ema-questions-and-answers-on-biosimilar-medicines-similar-biological-medicinal-products/ (accessed 22 Feb 2018).

EMA (2012b): Questions and answers on generic medicines. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Medicine_QA/2009/11/WC500012382.pdf (accessed 19 Mar 2018).

EMA (2016): Tailored scientific advice to support step-by-step development of new biosimilars. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Other/2016/12/WC500218206.pdf (accessed 22 Feb 2018).

EMA (2018): European public assessment reports. Available at http://www.ema.europa.eu/ema/index.jsp?curl=pages%2Fmedicines%2Flanding%2Fepar_search.jsp&mid=WC0b01ac058001d124&searchTab=searchByAuthType&

alreadyLoaded=true&isNewQuery=true&status=Authorised&keyword=Enter+keywords&searchType=name&taxonomyPath=&treeNumber=&searchGenericType=biosimilars&genericsKeywordSearch=Submit (accessed 02 Jul 2018).

FDA (2009): Biologics Price Competition and Innovation Act. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/ucm216146.pdf (accessed 22 Feb 2018).

Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M. and Granger, C. B. (2015): *Fundamentals of clinical trials*. Springer, New York, 5th edition.

Griffiths, C. E. M., Thaçi, D., Gerdes, S., Arenberger, P., Pulka, G., Kingo, K., Weglowska, J., the EGALITY study group, Hattebuhr, N., Poetzl, J., Woehling, H., Wuerth, G. and Afonso, M. (2017): The EGALITY study: a confirmatory, randomized, double-blind study comparing the efficacy, safety and immunogenicity of GP2015, a proposed etanercept biosimilar, vs. the originator product in patients with moderate-to-severe chronic plaque-type psoriasis. *British Journal of Dermatology*, 176 (4), 928–938.

Haustein, R., de Millas, C., Höer, A. and Häussler, B. (2012): Saving money in the European healthcare systems with biosimilars. *Generics and Biosimilars Initiative Journal*, 1 (3–4), 120–126.

Hommel, G. and Hoffmann, T. (1988): Controlled uncertainty. In: *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, 154–161. Springer, Berlin.

Kiefer, J. (1975): Construction and optimality of generalized Youden designs. In: *A Survey of Statistical Designs and Linear Models*, 333–353. North Holland Pub. Co., Amsterdam.

Kunert, J. (1983): Optimal design and refinement of the linear model with applications to repeated measurements designs. *The Annals of Statistics*, 11 (1), 247–257.

Kunert, J. and Mielke, J. (2018): Efficient designs for the estimation of mixed and self carryover effects. *SFB 823, Discussion paper*, 18 (8). DOI: 10.17877/DE290R-18820.

Kunert, J. and Stufken, J. (2002): Optimal crossover designs in a model with self and mixed carryover effects. *Journal of the American Statistical Association*, 97 (459), 898–906.

Kunert, J. and Stufken, J. (2008): Optimal crossover designs for two treatments in the presence of mixed and self-carryover effects. *Journal of the American Statistical Association*, 103 (484), 1641–1647.

Kushner, H. B. (1997): Optimal repeated measurements designs: the linear optimality equations. *The Annals of Statistics*, 25 (6), 2328–2344.

Lehmann, E. L. and Romano, J. P. (2005): Generalization of the familywise error rate. *The Annals of Statistics*, 33 (3), 1138–1154.

Lubenau, H., Bias, P., Maly, A.-K., Siegler, K. E. and Mehltretter, K. (2009): Pharmacokinetic and pharmacodynamic profile of new biosimilar filgrastim XM02 equivalent to marketed filgrastim Neupogen®. *BioDrugs*, 23 (1), 43–51.

Massey Jr., F. J. (1951): The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46 (253), 68–78.

Mielke, J., Jilma, B., Jones, B. and Koenig, F. (2018a): An update on the clinical evidence that supports biosimilar approvals in Europe. *British Journal of Clinical Pharmacology*, 84 (7), 1415–1431.

Mielke, J., Jilma, B., Koenig, F. and Jones, B. (2016): Clinical trials for authorized biosimilars in the European Union: a systematic review. *British Journal of Clinical Pharmacology*, 82 (6), 1444–1457.

Mielke, J., Jones, B., Jilma, B. and König, F. (2018b): Sample size for multiple hypothesis testing in biosimilar development. *Statistics in Biopharmaceutical Research*, 10 (1), 39–49.

Mielke, J. and Kunert, J. (2018): Universally optimal crossover designs for the estimation of mixed-carryover effects with an application to biosimilar development. *SFB 823, Discussion paper*, 18 (3). DOI: 10.17877/DE290R-18786.

Mielke, J., Schmidli, H. and Jones, B. (2018c): Incorporating historical information in biosimilar trials: challenges and a hybrid Bayesian-frequentist approach. *Biometrical Journal*, 60 (3), 564–582.

Mielke, J., Woehling, H. and Jones, B. (2018d): Longitudinal assessment of the impact of multiple switches between a biosimilar and its reference product on efficacy parameters. *Pharmaceutical Statistics*, 17 (3), 231–247.

Ng, T.-H. (2014): *Noninferiority testing in clinical trials: issues and challenges*. CRC Press, Boca Raton.

Patterson, S. D. and Jones, B. (2017): *Bioequivalence and statistics in clinical pharmacology*. CRC Press, Boca Raton, 2nd edition.

Putrik, P., Ramiro, S., Kvien, T. K., Sokka, T., Pavlova, M., Uhlig, T., Boonen, A. and Working Group 'Equity in access to treatment of rheumatoid athrithis in Europe' (2014): Inequities in access to biologic and synthetic DMARDs across 46 European countries. *Annals of the Rheumatic Diseases*, 73 (1), 198–206.

Rodrigues, M. I. and Iemma, A. F. (2014): *Experimental design and process optimization*. CRC Press, Boca Raton.

Rosenstock, J., Hollander, P., Bhargava, A., Ilag, L. L., Pollom, R. K., Zielonka, J. S., Huster, W. J. and Prince, M. J. (2015): Similar efficacy and safety of LY2963016 insulin glargine and insulin glargine (Lantus®) in patients with type 2 diabetes who were insulin-naïve or previously treated with insulin glargine: a randomized, double-blind controlled trial (the ELEMENT 2 study). *Diabetes, Obesity and Metabolism*, 17 (8), 734–741.

Rüger, B. (1978): Das maximale Signifikanzniveau des Tests: "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen". *Metrika*, 25 (1), 171–178.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D. and Neuenschwander, B. (2014): Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70 (4), 1023–1032.

Schuirmann, D. J. (1987): A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15 (6), 657–680.

Tóthfalusi, L., Endrényi, L. and Chow, S.-C. (2014): Statistical and regulatory considerations in assessments of interchangeability of biological drug products. *The European Journal of Health Economics*, 15 (1), 5–11.

Weise, M., Kurki, P., Wolff-Holz, E., Bielsky, M.-C. and Schneider, C. K. (2014): Biosimilars: the science of extrapolation. *Blood*, 124 (22), 3191–3196.

Wellek, S. (2010): *Testing statistical hypotheses of equivalence and noninferiority*. CRC Press, London, 2nd edition.

Zheng, J., Chow, S.-C. and Song, F. (2017): On safety margin for drug interchangeability. *Journal of Biopharmaceutical Statistic*, 27 (2), 293–307.