Aus dem Institut für Biometrie und Klinische Epidemiologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Evaluating and improving sample size recalculation in adaptive clinical study designs / Evaluierung und Verbesserung von Fallzahlrekalkulation in adaptiven klinischen Studiendesigns

zur Erlangung des akademischen Grades

Doctor of Philosophy (PhD)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Carolin Herrmann
aus Bielefeld

Datum der Promotion: 04.03.2022

Preface

This cumulative dissertation is submitted to the Medical Faculty Charité – Universitätsmedizin Berlin within the PhD studies program “Health Data Sciences”.
The work is based on three *Thesis Articles* and five *Thesis Related Articles*. I was supervised by Prof. Dr. Geraldine Rauch (Berlin), Prof. Dr. Meinhard Kieser (Heidelberg) and Dr. Jochen Kruppa (Berlin). Moreover, I collaborated with Maximilian Pilz (Heidelberg) and Kevin Kunzmann (Cambridge, UK). My research was funded by the Deutsche Forschungsgesellschaft (grant RA 2347/4-1).

Drawn by Carolin Herrmann (2021).

Contents

List of Figures

List of Tables

List of Abbreviations

ADAS-Cog	Alzheimer’s Disease Assessment Scale-Cognitive Subscale
EMA	European Medicines Agency
FDA	Food and Drug Administration
ICH	International Conference on Harmonization
OCP	observed conditional power approach
OF	optimization function approach
PZ	promising zone approach
RA	recalculation area
ROCP	restricted observed conditional power approach

List of Symbols

General statistical quantities:

α	global one-sided significance level
$f(\cdot)$	probability density function
H_0, H_1	null and alternative hypothesis
$\Phi(\cdot)$	standard normal distribution function
$q_{1-\alpha}$	$(1-\alpha)$-quantile of standard normal distribution

Study design:

α_1, α_{1+2}	locally adjusted one-sided significance level after stage 1 and at the final analysis
α_0	futility stopping boundary
$1-\beta$	anticipated power (conditional or global depending on context)
$1-\beta_{min}$	lower conditional power boundary for sample size recalculation according to the restricted observed conditional power approach
C	index for control group
$c_2(\cdot)$	stage 2 critical value function
Δ	true standardized treatment effect
I	index for intervention group
μ_I, μ_C	population means in control and intervention group
n	sample size per group
n_1	stage 1 sample size per group
n_2	additional sample size per group in stage 2
n_{1+2}	overall sample size per group
n_{max}	maximally feasible sample size per group

n_{Δ}^{fix} fixed design's sample size per group for effect Δ and for a power of 80%

σ common standard deviation of the endpoint in both groups

w_1, w_2 weights for the inverse normal test for stage 1 and 2

$\overline{X}_{C,1}, \overline{X}_{C,2}$ observed means in control group after stage 1 or 2

$\overline{X}_{I,1}, \overline{X}_{I,2}$ observed means in intervention group after stage 1 or 2

z_1 observed value of interim test statistic

Z_1, Z_2 Z-test statistic including data of stage 1 and 2

Z_{1+2} combined test statistic for stages 1 and 2

Conditional performance score:

CN conditional sample size

CP conditional power

CS conditional performance score

e_{CN}, e_{CP} location components of the conditional performance score for conditional sample size and conditional power

$\gamma_{e,CN}, \gamma_{v,CN}$ customized component weights in the conditional performance score for conditional sample size

$\gamma_{e,CP}, \gamma_{v,CP}$ customized component weights in the conditional performance score for conditional power

S_{CN}, S_{CP} conditional sample size and conditional power sub-score

v_{CN}, v_{CP} variation components of the conditional performance score for conditional sample size and conditional power

$Var^{max}(CN)$ maximally possible variance of conditional sample size as element of the conditional performance score

$Var^{max}(CP)$ maximally possible variance of conditional power as element of the conditional performance score

$CN^{max} - CN^{min}$ maximally possible deviation of conditional sample size as element of the conditional performance score

$CP^{max} - CP^{min}$ maximally possible deviation of conditional power as element of the conditional performance score

$CN_{\Delta}^{target}, CP_{\Delta}^{target}$ conditional sample size or conditional power target value as element of the conditional performance score for an effect Δ

Smoothing correction:

c_{incr} smallest interim test statistic suggesting the maximally feasible sample size

Optimization approach:

$\mathcal{D}(\cdot,\cdot,\cdot,\cdot,\cdot)$	five-tuple describing an adaptive two-stage design
k	number of pivot points in optimization approach
$\tilde{S}(\cdot,\cdot,\cdot,\cdot,\cdot)$	customized score in optimization approach
$z_1^{(i)}, i \in \{1,...,k\}$	discrete set of pivot points describing possible values of the interim test statistic in the recalculation area
$\omega_i, i \in \{1,...,k\}$	weights in Gaussian quadrature rule

Abstract

Background

A valid sample size calculation is a key aspect for ethical medical research. While the sample size must be large enough to detect an existing relevant effect with sufficient power, it is at the same time crucial to include as few patients as possible to minimize exposure to study related risks and time to potential market approval. Different parameter assumptions, like the expected effect size and the outcome's variance, are required to calculate the sample size. However, even with high medical knowledge it is often not easy to make reasonable assumptions on these parameters. Published results from the literature may vary or may not be comparable to the current situation. Adaptive designs offer a possible solution to deal with those planning difficulties. At an interim analysis, the standardized treatment effect is estimated and used to adapt the sample size. In the literature, there exists a variety of strategies for recalculating the sample size. However, the definition of performance criteria for those strategies is complex since the second stage sample size is a random variable. It is also known since long that most existing sample size recalculation strategies have major shortcomings, such as a high variability in the recalculated sample size.

Methods

Within *Thesis Article 1*, me and my coauthors developed a new performance score for comparing different sample size recalculation rules in a fair and transparent manner. This performance score is the basis to develop improved sample size recalculation strategies in a second step. In *Thesis Article 2*, me and my supervisor propose smoothing corrections to be combined with existing sample size recalculation rules to reduce the variability. *Thesis Article 3* deals with the determination of the second stage sample size as the numerical solution of a constrained optimization problem, which is solved by a new `R`-package named `adoptr`. To illustrate the relation of the three *Thesis Articles*, all new approaches are applied to a clinical trial example to show the methods' benefits in comparison to an established sample size recalculation strategy.

Results

The global aim of defining high-performance sample size recalculation rules was approached considerably by my work. The performance of adaptive designs with sample size recalculation can now be compared by means of a single comprehensive score. Moreover, our new smoothing corrections define one possibility to improve an existing sample size recalculation rule with respect to this new performance score. The new software further allows to directly determine an optimal second stage sample size with respect to predefined optimality criteria.

Conclusions

I was able to reduce methodological shortcomings in sample size recalculation by four aspects: providing new methods for 1) performance evaluation, 2) performance improvement, 3) performance optimization and 4) software solutions. In addition, I illustrate how these methods can be combined and applied to a clinical trial example.

Zusammenfassung

Hintergrund

Eine valide Fallzahlberechnung ist ein zentraler Aspekt für ethische medizinsiche Forschung. Während die Fallzahl groß genug sein muss, um einen vorliegenden relevanten Effekt mit genügend großer Power zu entdecken, ist es gleichzeitig wichtig, so wenig Patient*innen wie möglich einzuschließen, um studienbezogene Risiken sowie die Zeit bis zur Marktzulassung zu minimieren. Verschiedene Parameterannahmen, wie die erwartete Effektgröße und die Varianz des Endpunktes, werden benötigt, um die Fallzahl zu berechnen. Auch mit hoher medizinischer Expertise ist es häufig nicht einfach, die zugrundliegenden Parameterannahmen zu treffen. Publizierte Ergebnisse aus der Literatur können variieren oder auf die aktuelle Situation nicht übertragbar sein. Adaptive Designs sind eine Möglichkeit, um mit diesen Planungsunsicherheiten umzugehen. Zur Zwischenanalyse wird der Behandlungseffekt geschätzt und genutzt, um die Fallzahl anzupassen. In der Literatur gibt es eine Vielzahl an Strategien die Fallzahl anzupassen. Die Definition von Beurteilungskriterien dieser Strategien ist jedoch komplex, da die Fallzahl der zweiten Stufe eine Zufallsvariable ist. Hinzu kommt, dass viele existierende Fallzahlrekalkulations-Strategien Defizite haben, beispielsweise eine hohe Variabilität in der rekalkulierten Fallzahl.

Methoden

Im *Promotionsartikel 1* entwickelten meine Koautor*innen und ich einen neuen Performance-Score für einen fairen und transparenten Vergleich von Fallzahlrekalkulations-Strategien. Dieser Performance-Score diente im zweiten Schritt als Basis, um verbesserte Fallzahlrekalkulations-Strategien zu entwickeln. Hierfür schlugen meine Betreuerin und ich im *Promotionsartikel 2* Smoothing-Korrekturen zur Varianzreduktion vor, die mit bereits existierenden Fallzahlrekalkulations-Strategien kombiniert werden können. Im *Promotionsartikel 3* wurde die Fallzahl der zweiten Stufe als numerische Lösung eines Optimierungsproblems aufgefasst, welche durch das neue R-Paket `adoptr` berechnet wird. Um den Zusammenhang der drei zugrundeliegenden Artikel zu illustrieren, wurden die neuen Methoden auf ein klinisches Studienbeispiel angewandt und ihre Vorteile gegenüber einer etablierten Fallzahlrekalkulations-Strategie erläutert.

Ergebnisse
Das übergeordnete Ziel qualitativ hochwertige Fallzahlrekalkulations-Strategien zu definieren, wurde durch meine Arbeit beträchtlich vorangetrieben. Die Performance von adaptiven Designs mit Fallzahlrekalkulation kann nun durch einen umfassenden Score beurteilt werden. Darüberhinaus stellen die neuen Smoothing-Korrekturen eine Möglichkeit dar, um Fallzahlrekalkulations-Strategien hinsichtlich des neuen Performance-Scores zu verbessern. Die neue Software erlaubt darüber hinaus, eine optimale Fallzahl der zweiten Stufe in Bezug auf vorab definierte Optimalitätskriterien zu bestimmen.

Schlussfolgerungen
Im Rahmen dieser Arbeit habe ich durch vier Aspekte dazu beigetragen, methodische Defizite im Bereich der Fallzahlrekalkulation zu reduzieren: 1) Performance-Bewertung, 2) Performance-Verbesserung, 3) Performance-Optimierung und 4) Software-Lösungen. Zusätzlich wird illustriert wie diese Methoden kombiniert und auf ein klinisches Studienbeispiel angewandt werden können.

1 Introduction

It is the common interest of patients, medical professionals, authorities, and pharmaceutical companies to demonstrate "efficacy and safety of new treatments" in clinical trials (Bretz et al., 2017). After the clinical goal is defined, it is the biostatistician's responsibility to suggest an appropriate statistical design (Dragalin, 2006). One major statistical design aspect is the sample size determination. Sample size calculation is one of the most common topics in biostatistical consulting and an important criterion for judging medical research projects by ethics committees (Kieser, 2018). The sample size is calculated to detect a relevant treatment effect with sufficient power while the type I error rate is protected. In the ICH-E9 guideline it is stated on page 19 that the

> "*number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed*" (ICH, 1998).

Indeed, in a study with too small sample size, time and money would be wasted without the chance of reaching the study aim. On the other hand, if more patients than necessary are recruited, later recruited patients in the inferior arm are unnecessarily treated with a less effective treatment, although evidence for the other arm is already sufficient and the market approval is unnecessarily prolonged. Determining the "correct" sample size is therefore a critical task. To perform sample size calculation, several parameter assumptions are required which are based on medical knowledge and on published findings from the literature. While the power and significance level leave only little room for discussion, it is a much more difficult task to determine the targeted treatment effect and the outcome's variance, as historical studies might not be comparable to the current situation, deviate in their findings or even do not exist at all. Hence, the sample size may be calculated based on vague assumptions and is therefore ethically questionable. In addition, Koch (2006) points out the difficult prediction of placebo responses and therefore misguidingly calculated sample sizes as one reason for study failures.

To deal with planning uncertainties related to sample size determination, the idea of updating the sample size during an ongoing trial seems appealing. Adaptive group sequential designs present the framework for this. While in a classical study design, the sample size of the trial is fixed before the recruitment of patients starts, adaptive designs allow for an update of the sample size during an ongoing trial. Especially for studies that recruit over a longer period of time, interim looks seem appealing. Jennison and Turnbull (1999) state that sequential methods often reduce time and costs compared to a fixed sample size design. In an adaptive design, one or more interim analyses are integrated. Those interim analyses include the opportunity of stopping early due to either futility or early proof of efficacy, or due to safety reasons. Otherwise, the study continues with the subsequent stage and further patients are included. The innovative aspect is that adaptive designs allow for trial design modifications based on the results of the interim analysis or new external information. Bauer et al. (2016) state on page 339:

> "*Adaptive confirmatory designs have shaken the classical design paradigm that the details of the design and statistical analysis all have to be laid down in advance.*"

Whereas in principle, adaptive designs allow various design modifications such as a change of the target population, a switch in endpoints or drop of study arms, sample size recalculation can be seen as the most prominent adaptive design element (Bauer et al., 2016). Thereby, sample size recalculation in adaptive designs is not restricted to a particular medical research area. Applications of such designs can be found in research on Alzheimer's disease (Wang et al., 2015), knee osteoarthritis (Bowden and Mander, 2014), schizophrenia (Mehta and Pocock, 2011), cancer (Mauer et al., 2012), depression (Fedgchin et al., 2019), sepsis (Hager et al., 2019) or multiple sclerosis (Zajicek et al., 2012). Adaptive study designs are the subject of regulatory guidance documents in Europe and the United States and there is still active methodological research ongoing, also with respect to other study types than classical randomized, controlled trials, e.g., for single-arm phase IIa trials (Schmidt et al., 2018) or diagnostic studies (Stark and Zapf, 2020). The European Medicines Agency (EMA) through its Committee for Medicinal Products for Human Use (2007) provided the first regulatory guidance on adaptive designs with a clear focus on caution. In line with this, the Food and Drug Administration (FDA) (2019) recently published their finalized guidance on adaptive designs for medical device clinical studies including principles for the design, performance and report.

There exists a variety of so called sample size recalculation rules in the literature. In here, the focus is on sample size recalculation after unblinded interim analyses, which allows to re-estimate the treatment effect. The treatment effect is a main parameter and often subject to uncertainty (Wassmer and Brannath, 2016, Kieser, 2020). In contrast, blinded interim analyses can only be used to re-estimate nuisance parameters such as the outcome's variance (Kieser and Friede, 2003, Proschan, 2009). Blinded interim analyses will not be the topic in the remainder of this thesis. Many of the published sample size recalculation rules rely on conditional power arguments, e.g., Proschan and Hunsberger (1995), Cui et al. (1999), Denne (2001), Shun et al. (2001), Posch et al. (2003), Chen et al. (2004), Gao et al. (2008), Mehta and Pocock (2011). The conditional power depends on the true treatment effect and different approaches exist to deal with this unknown parameter. For example, the assumed conditional power describes the probability of correctly rejecting the null hypothesis at the end of the trial based on the assumed treatment effect in the planning stage, compare e.g. Bauer and König (2006). The observed conditional power describes the probability of correctly rejecting the null hypothesis at the end of the trial given the data observed at interim. Extended approaches are given by Mehta and Patel (2006) who suggested to also take economic considerations into account and Mehta and Pocock (2011) who proposed a decision rule approach to update the sample size. This approach was extended by requiring a minimum conditional power for an increase in sample size (Hsiao et al., 2019). Jennison and Turnbull (2015) suggested a cost benefit approach where the number of patients is outweighed against an increased conditional power. Pilz et al. (2019) proposed a general formulation of an optimization approach for a two-stage design taking first and second stage parameters into account. These are just some examples of the broad literature published on sample size recalculation.

Despite the great variety of sample size recalculation approaches, there exists no clear guidance on how to choose and compare a specific design. Neither the EMA nor the FDA give comments on this aspect in their related guidelines (European Medicines Agency (EMA) through its Committee for Medicinal Products for Human Use, 2007, Food and Drug Administration (FDA), 2019). Whereas for a standard fixed design, performance assessment is well established with respect to evaluating sample size and power, the performance assessment of adaptive designs is more complex, as the conditional power and second stage sample size are random variables in this case. Sample sizes are usually updated based on the observed interim data in adaptive designs (Kieser, 2020). The observed interim effect, however, is usually based on rather small interim sample sizes and is thus subject to a considerable random error. This may "*lead to highly variable second-stage sample sizes*" (Dragalin, 2006). Therefore, Wassmer and Brannath (2016) suggest to consider and calculate the conditional power for a range of effect sizes. Another strategy to account for the variability of the interim effect estimate is to use Bayesian conditional power concepts for sample size recalculation, which rely on a prior distribution for the interim effect (Spiegelhalter and Freedman, 1986, Dmitrienko and Wang, 2006). Besides the variability in sample size, other points of criticism are that recalculated sample sizes are very large and the fact that the target power is often not met (Bauer and Köhne, 1994, Levin et al., 2013). Therefore, to fully profit from the benefits of adaptive sample size recalculation, there remain open methodological tasks to be solved. Another important aspect is the availability of software, which is essential for the application and thus also recognition of statistical methodology (Bauer et al., 2016), since adaptive designs are often related to high computational requirements. Software reviews on adaptive designs can be found in Wassmer and Vandemeulebroecke (2006), Tymofyeyev (2014), Wassmer and Brannath (2016), Grayling and Wheeler (2020). Commercial software for sample size recalculation is provided by, e.g., East® (Cytel, 2020) with ADAPT® and SURVADAPT®. Within the open source software R (R Core Team, 2021), there also exist several related packages, compare e.g., Vandemeulebroecke (2009), Wassmer and Pahlke (2021). Nevertheless, Bauer et al. (2016) emphasize the need for software. Pallmann et al. (2018) even go further and mention unavailable software as time-limiting factor when considering the application of adaptive study designs.

This thesis is structured as follows: After this introduction, in the methods section, I describe the underlying study setting, the basic mathematical framework and the methodology developed within this thesis. In the results section, I summarize the results from the three *Thesis Articles* and illustrate their relation by additional results of a comparative clinical trial example. In the discussion section, I highlight the achievements and the limitations of my work in the context of the existing literature, describe briefly the main aspects of the *Thesis Related Articles* and deduce practical recommendations for applications. Moreover, I draw the final conclusions and give an outlook for planned future work.

Aim of this thesis

My PhD position was funded by the German Research Foundation by the joint Berlin and Heidelberg project called "**O**ptimal **R**ules for **A**daptive re**C**alculation of samp**LE** size in clinical trials (ORACLE)" (grant RA 2347/4-1). For this reason, the Heidelberg project partners define the collaborators and coauthors of my work.

The global aim of this thesis is to overcome existing problems of adaptive designs with sample size recalculation and to make a contribution to the optimization of such designs. This overall aim is addressed in four steps: First, I will suggest a new performance score for adaptive study designs which allows for a fair and comprehensive judgment and comparison of existing recalculation rules. Second, using this score, I will suggest methods to improve existing sample size recalculation rules with respect to variability reduction. Third, optimal sample size recalculation rules can be directly deduced by defining an optimization problem with predefined optimality criteria, where the new performance score defines an attractive option. Finally, I will also fill the gap of missing software.

These aspects are addressed in the following three *Thesis Articles*, which comprise my cumulative thesis:

1) **C. Herrmann**, M. Pilz, M. Kieser, G. Rauch (2020). "A new conditional performance score for the evaluation of adaptive group sequential designs with sample size recalculation." *Statistics in Medicine*, **39**, 2067–2100. doi: 10.1002/sim.8534.

2) **C. Herrmann**, G. Rauch (2021). "Smoothing corrections for improving sample size recalculation rules in adaptive group sequential study designs." *Methods of Information in Medicine*, **60**, 1-8. doi: 10.1055/s-0040-1721727.

3) K. Kunzmann[1], M. Pilz[1], **C. Herrmann**, G. Rauch, M. Kieser (2021). "The adoptr package:

[1] Authors contributed equally

adaptive optimal designs for clinical trials in R." *Journal of Statistical Software*, **98**, 1-21. doi: 10.18637/jss.v098.i09.

Apart from these three articles, during my thesis work, my collaborators and I wrote five other *Thesis Related Articles*, which are in different publications stages:

I) M. Pilz, K. Kunzmann, **C. Herrmann**, G. Rauch, M. Kieser (2019). "A variational approach to optimal two-stage designs." *Statistics in Medicine*, **38**:4159-4171. doi: 10.1002/sim.8291.

II) X. Li, **C. Herrmann**, G. Rauch (2020). "Optimality criteria for futility stopping boundaries for group sequential designs with a continuous endpoint." *BMC Medical Research Methodology*, **20**, 274. doi: 10.1186/s12874-020-01141-5.

III) **C. Herrmann**[1], C. Kluge[1], M. Pilz, M. Kieser, G. Rauch (2021). "Improving sample size recalculation rules in adaptive clinical trials by resampling." *Pharmaceutical Statistics*. doi: 10.1002/pst.2122.

IV) M. Pilz, K. Kunzmann, **C. Herrmann**, G. Rauch, M. Kieser (2021). "Optimal planning of adaptive two-stage designs." *Statistics in Medicine*, **40**, 3196-3213. doi: 10.1002/sim.8953.

V) M. Pilz, **C. Herrmann**, G. Rauch, M. Kieser. "Optimal unplanned design modification in adaptive two-stage trials." (*submitted in 2020*)

[1] Authors contributed equally

2 Methods

2.1 The underlying study setting

Within this thesis, a two-armed clinical trial with a normally distributed endpoint is considered. The one-sided superiority test problem is formulated as

$$H_0 : \mu_I - \mu_C \leq 0 \quad \text{versus} \quad H_1 : \mu_I - \mu_C > 0, \tag{2.1}$$

where μ_I and μ_C refer to the population means in the intervention and the control group, respectively. The study is planned with a two-stage adaptive group sequential study design with a single interim analysis, following the recommendation of the European Medicines Agency (EMA) through its Committee for Medicinal Products for Human Use (2007) to keep the number of design modifications low. The global one-sided significance level is set to α. In the following, the fundamentals of the considered adaptive design will be outlined. Details on the theory of adaptive designs are provided, e.g., by Bretz et al. (2009), Wassmer and Brannath (2016). The test statistic for the interim analysis, which is exclusively based on the data of the first stage, is the standard test statistic of the two-sample Z-test given by

$$Z_1 = \frac{\overline{X}_{I,1} - \overline{X}_{C,1}}{\sigma} \cdot \sqrt{\frac{n_1}{2}}, \tag{2.2}$$

where $\overline{X}_{I,1}$ and $\overline{X}_{C,1}$ are the observed means at interim per group. The variance σ^2 is supposed to be known and equal in both groups. Note that it can also be considered as the Z-approximation of the t-test if the standard deviations are not known and σ is replaced by the pooled standard deviation.

At the interim analysis after having observed n_1 patients per group, the trial can be stopped early or continued with the possibility to adapt the sample size:

- The trial is stopped after the first stage with the rejection of the null hypothesis if $Z_1 \geq q_{1-\alpha_1}$, where α_1 is the local, adjusted one-sided significance level with $\alpha_1 < \alpha$ and $q_{1-\alpha_1}$ is the corresponding standard normal quantile.

- The trial is stopped for futility after the first stage if $Z_1 < q_{1-\alpha_0}$, where $q_{1-\alpha_0}$ with $\alpha_0 > \alpha$ is the futility stopping boundary.
- If $q_{1-\alpha_0} \leq Z_1 < q_{1-\alpha_1}$, the trial continues with the second stage and further patients are recruited. The number of the additional patients $n_2(z_1)$ is usually calculated based on the observed interim test statistic z_1 and can therefore be written as a function of z_1.

If the trial continues to the second stage, the normally distributed, independent incremental test statistic Z_2 is calculated exclusively based on the data of the second stage as

$$Z_2 = \frac{\overline{X}_{I,2} - \overline{X}_{C,2}}{\sigma} \cdot \sqrt{\frac{n_2}{2}}, \tag{2.3}$$

where the index 2 refers to the data of the second stage. Note that Z_1 and Z_2 are stochastically independent by construction and can be combined by the inverse normal combination test (Lehmacher and Wassmer, 1999) as follows,

$$Z_{1+2} = \frac{w_1 \cdot Z_1 + w_2 \cdot Z_2}{\sqrt{w_1^2 + w_2^2}}, \tag{2.4}$$

with predefined weights w_1, w_2 such that $w_1^2 + w_2^2 = 1$. The test statistic Z_{1+2} defines the test statistic for the final analysis. Weights are often chosen as $w_1 = \sqrt{n_1}/\sqrt{n_1 + n_2}$ and $w_2 = \sqrt{n_2}/\sqrt{n_1 + n_2}$, which corresponds to an optimal choice if the sample size is not adapted. Note that apart from the inverse normal combination test, there also exist other combination tests, e.g., as proposed by Bauer and Köhne (1994), or the conditional error function approach (Proschan and Hunsberger, 1995, Müller and Schäfer, 2004) for combining the stage-wise data. The methods proposed in the *Thesis Articles* and presented in here are not restricted to a specific combination function. If $Z_{1+2} \geq q_{1-\alpha_{1+2}}$ with a local, adjusted one-sided significance level $\alpha_{1+2} < \alpha$, the null hypothesis is rejected at the final analysis. Otherwise, the null hypothesis cannot be rejected. The different options are visualized in Figure 2.1.

Note that due to the interim analysis, a multiple testing problem arises and therefore the stage-wise local significance levels α_1 and α_{1+2} must be adjusted. Within this thesis, a global one-sided significance level of $\alpha = 0.025$ was chosen and local critical values $q_{1-\alpha_1}$, $q_{1-\alpha_{1+2}}$ according to the most simple adjustment strategy proposed by Pocock (1977), given by equal local significance levels $\alpha_1 = \alpha_{1+2} = 0.0147$. Moreover, a binding futility boundary $q_{1-\alpha_0}$ was applied which means that whenever $Z_1 < q_{1-\alpha_0}$, a trial stop at interim is mandatory. Note that by applying a binding futility boundary, the local significance levels might be increased to fully exhaust the global level. Apart from *Thesis Article 2* (Herrmann and Rauch, 2021), we neglected this possible modification in the work presented in this thesis. The futility stopping boundary was set to $q_{1-\alpha_0} = 0$ which is equivalent to stopping the trial when the effect size points in the wrong direction. Whenever the observed interim test statistic suggests a second stage of

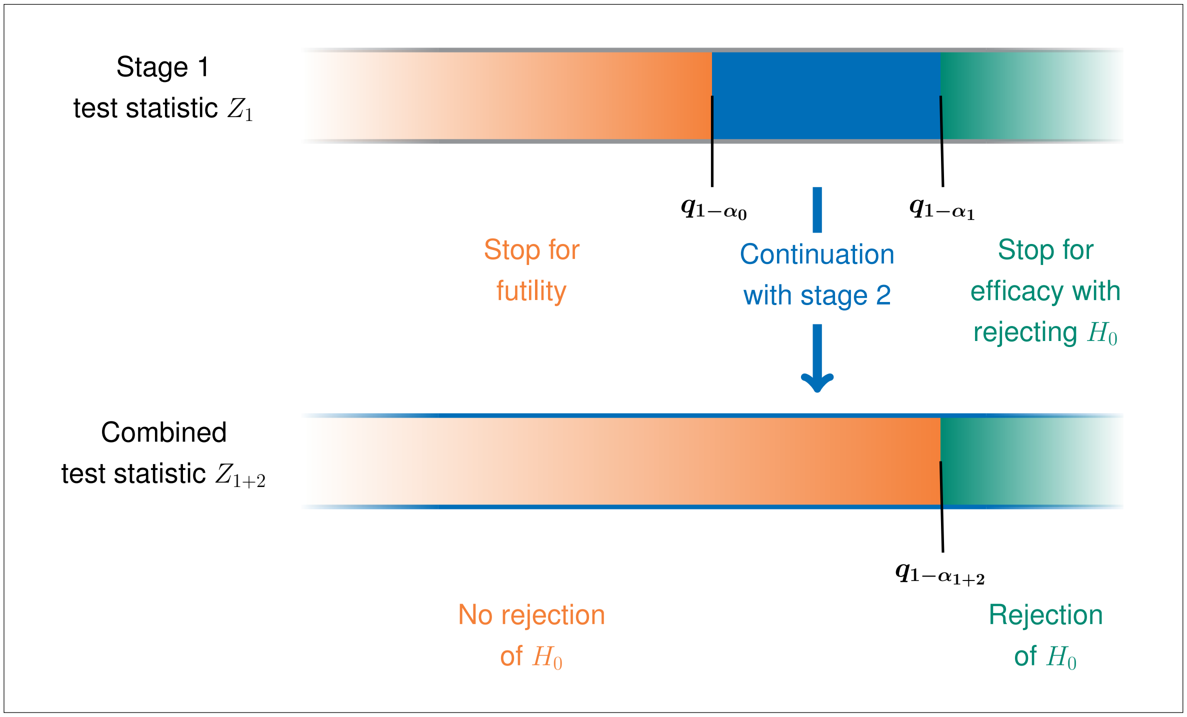


Figure 2.1 Schematic scheme of options in a two-stage adaptive design with binding futility stopping bound α_0, local significance levels α_1 and α_{1+2}, respective quantiles q and test statistics Z_1 and Z_{1+2}.

the trial, i.e., $q_{1-\alpha_0} \leq Z_1 < q_{1-\alpha_1}$, further patients may be recruited. The corresponding interval $RA := [q_{1-\alpha_0}; q_{1-\alpha_1})$ is called the recalculation area and for $\alpha_0 = 0.5$ and $\alpha_1 = 0.0147$ given by $[0; 2.178)$. If the trial continues to the second stage, there exists a large variety of sample size recalculation strategies of which some of the most prominent ones based on conditional power arguments are summarized in detail in *Thesis Article 1* (Herrmann et al., 2020). In here, the "restricted observed conditional power approach" is exemplary described as one of the most common options. This approach was also investigated and adapted in *Thesis Articles 1 and 2* (Herrmann et al., 2020, Herrmann and Rauch, 2021). The restricted observed conditional power approach mimics the principles of sample size calculation in a fixed design. The approach determines the second stage sample size such that a certain conditional power for the final analysis is attained where the condition refers to the observed interim outcome. The conditional power describes the probability of correctly rejecting the null hypothesis after having observed n patients under knowledge of the interim test statistic z_1,

$$CP_\Delta(z_1, n) \quad := \begin{cases} 0, \text{ if } z_1 < q_{1-\alpha_0}, \\ 1 - \Phi\left(q_{1-\alpha_{1+2}} \cdot \frac{\sqrt{w_1^2+w_2^2}}{w_2} - z_1 \cdot \frac{w_1}{w_2} - \Delta \cdot \sqrt{\frac{n_1}{2}} \cdot \sqrt{\frac{n-n_1}{n_1}}\right), \\ \quad \text{if } z_1 \in [q_{1-\alpha_0}; q_{1-\alpha_1}), \\ 1, \text{ if } z_1 \geq q_{1-\alpha_1}, \end{cases} \tag{2.5}$$

where $\Phi(\cdot)$ stands for the standard normal distribution function. Since the conditional power depends on the unknown true standardized treatment effect size $\Delta = (\mu_I - \mu_C)/\sigma$ with known common standard deviation σ, there exist different strategies in the literature for approaching the true effect size, e.g., by the observed interim effect, by the initially assumed effect, or by a prior distribution referred to as predictive power (Spiegelhalter et al., 1986, Dmitrienko and Wang, 2006, Lan et al., 2009). The restricted observed conditional power approach makes use of the observed interim treatment effect. This approach is used here for illustration without the intention to recommend this option. In a first step, the sample size in the recalculation area is determined such that a pre-specified conditional power value of $1 - \beta$ is attained if the interim test statistic falls within the recalculation area by using formula (2.5). If the recalculated sample size exceeds a maximally feasible sample size n_{max}, this maximum sample size is chosen instead. Additionally, the approach requires a minimal conditional power for the updated sample size. If for a certain effect size, a conditional power of $1 - \beta_{\min} \leq 1 - \beta$ cannot be achieved with the maximally feasible sample size, the sample size remains at n_1 such that the trial is stopped after the first stage. Note that in that case, the recalculation area RA, also consists of "recalculated" second stage sample sizes equal to zero. Without loss of generality, the power parameters were set as $1 - \beta = 0.8$ and $1 - \beta_{\min} = 0.6$ within the remainder of this thesis. A more detailed description of the restricted observed conditional power approach can be found in *Thesis Article 1* (Herrmann et al., 2020). The approach is graphically illustrated in Figure 2.2.

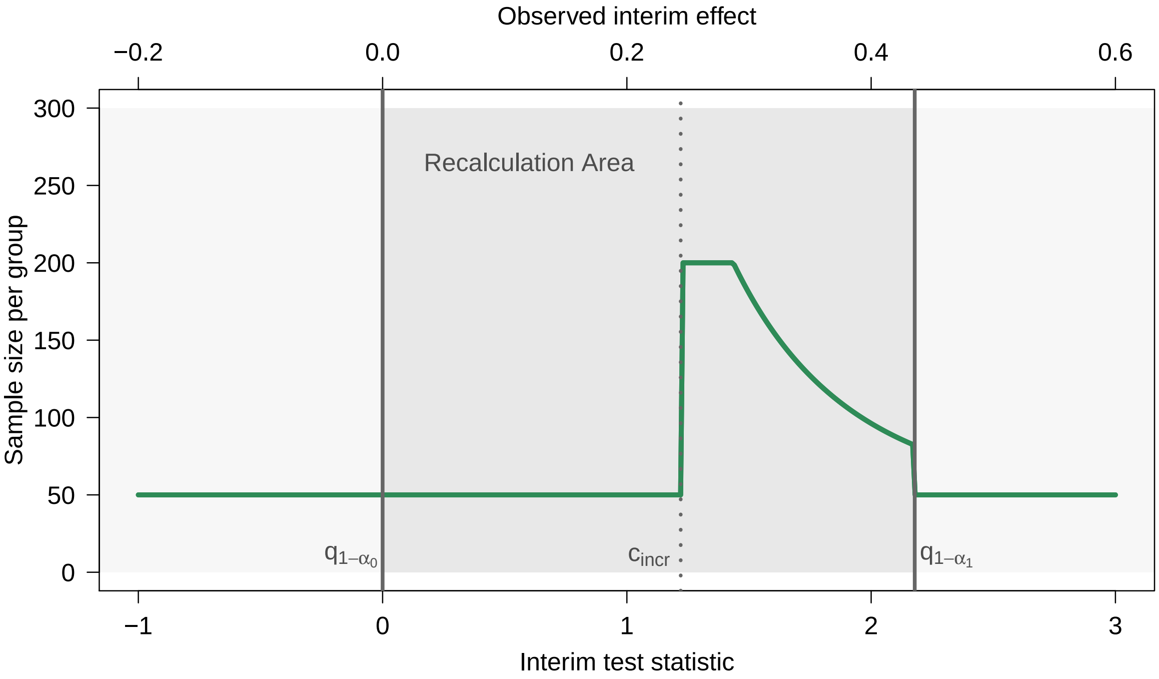


Figure 2.2 Sample size recalculation according to the restricted observed conditional power approach for $n_1 = 50$, $n_{max} = 200$, $\alpha_1 = \alpha_{1+2} = 0.0147$, $\alpha_0 = 0.5$ and with recalculation area $[q_{1-\alpha_0}; q_{1-\alpha_1})$, where c_{incr} describes the smallest value of the interim test statistic that suggests n_{max}. The graph is similar to Figure 1 in *Thesis Article 1* (Herrmann et al., 2020).

2.2 New methodological approaches

2.2.1 *Thesis Article 1*: Evaluating sample size recalculation with a new conditional performance score

To improve sample size recalculation in general, one needs an instrument to judge the performance of sample size recalculation approaches. Zhang et al. (2016) state on page 3388:

> "*In the area of flexible sample size designs, performance assessment and comparison perhaps are the most confusing issue.*"

While in a fixed sample size design a good performance is naturally classified by a design meeting a certain power given an assumed effect, there are different perspectives to be considered for judging a sample size recalculation rule in an adaptive study design. One perspective is to evaluate the global (or overall) performance that is looking at the properties of the adaptive design *before the study starts*. For example, the global power as one global performance measure describes the probability of correctly rejecting the null hypothesis either at interim or at the final stage. Thus, global power and global average sample size are often reported as performance measures for adaptive sample size designs, compare e.g., Lehmacher and Wassmer (1999), Liu et al. (2008), Mehta and Pocock (2011). Another point of view is the conditional perspective assessing the properties of the adaptive design *when the interim results suggest a trial continuation*. Both perspectives are required simultaneously and valid, however when raising the question how a good approach to calculate the *second stage sample size* would look like, one naturally falls within the second perspective as sample size recalculation is only done if the interim effect falls within the recalculation area. In principle, for the conditional perspective one is also interested in sample size and power, however both – conditional power and sample size – are random variables now, which are best described by location *and* variation measures. The term *conditional sample size* is used in this thesis to describe the total sample size $n_1 + n_2(z_1)$, under the condition that the observed interim test statistic z_1 falls within the recalculation area. Although the literature often focuses on the global perspective, global performance measures alone are not sufficient. Even if a design performs well with respect to overall power, the recalculation strategy can still not be considered as satisfactory if the average conditional power is low and the high global power is due to a good power at interim. Moreover, a high average conditional power might not be sufficient if it is subject to high random variability.

Several authors have published on global performance measures and scores for adaptive designs, compare e.g., Liu et al. (2008), Wu and Cui (2012), Fang et al. (2018). However, until now there was no performance score measuring the conditional perspective which seems very natural in an adaptive setting.

Thesis Article 1 (Herrmann et al., 2020) presents a new performance score for the evaluation of sample size recalculation rules in adaptive study designs from the conditional perspective. The formal definition of the performance score is rather technical and requires several mathematical definition steps. In here, only a brief description of the construction will be provided. However, a full understanding will be easier when reading the complete *Thesis Article 1* (Herrmann et al., 2020). The conditional performance score takes values between 0 and 1, where higher values indicate a better performance. Furthermore, the conditional performance score also includes measures of variation next to measures of location. More precisely, the conditional performance score CS consists of a conditional power and a conditional sample size (as defined above) sub-score, S_{CP} and S_{CN}, and is given by

$$CS(\Delta) = 0.5 \cdot (S_{CP}(\Delta) + S_{CN}(\Delta)) \,. \tag{2.6}$$

Note that in general also a different weighting of the two sub-scores is possible but for simplicity reasons I rely on an equal weighting in here. Irrespectively of the chosen weights, both sub-scores consist of a location and variation component, e_{CN} and v_{CN}, for the the conditional sample size and a location and variation component, e_{CP} and v_{CP}, for the conditional power. The components are averaged according to customized weights $\gamma_{e,CN} + \gamma_{v,CN} = 1$ and $\gamma_{e,CP} + \gamma_{v,CP} = 1$ as follows:

$$S_{CN}(\Delta) = \gamma_{e,CN} \cdot e_{CN}(\Delta) + \gamma_{v,CN} \cdot v_{CN}(\Delta), \tag{2.7}$$

$$S_{CP}(\Delta) = \gamma_{e,CP} \cdot e_{CP}(\Delta) + \gamma_{v,CP} \cdot v_{CP}(\Delta). \tag{2.8}$$

In the remainder of this thesis, I assume $\gamma_{e,CN} = \gamma_{v,CN} = \gamma_{e,CP} = \gamma_{v,CP} = 0.5$. The score components are graphically illustrated in Figure 2.3. By construction, the sub-scores as well as the location and variation components values also range between 0 and 1. The sub-scores as well as its variation and location components can thus be reported and interpreted separately, which allows investigating the reasons for a high or low total score.

The idea of the conditional sample size location component is to evaluate the relation of the difference of the expected conditional sample size and a predefined target value CN_{Δ}^{target} to the maximally possible deviation of the conditional sample size $CN^{max} - CN^{min}$, i.e.,

$$e_{CN}(\Delta) = 1 - \frac{\left|\mathbb{E}[CN(Z_1)] - CN_{\Delta}^{target}\right|}{CN^{max} - CN^{min}}. \tag{2.9}$$

The location component for the conditional power is given accordingly by

$$e_{CP}(\Delta) = 1 - \frac{\left|\mathbb{E}[CP(Z_1)] - CP_{\Delta}^{target}\right|}{CP^{max} - CP^{min}}. \tag{2.10}$$

In case $\Delta = 0$ or the related fixed sample size n_{Δ}^{fix} exceeds the maximally feasible sample size n_{max}, an increase in sample size is considered as not meaningful. Thus, the predefined target values depend on the treatment effect Δ and are given explicitly in Table 2.1.

The variation components evaluate the ratio of the variance of CN or CP to the maximally

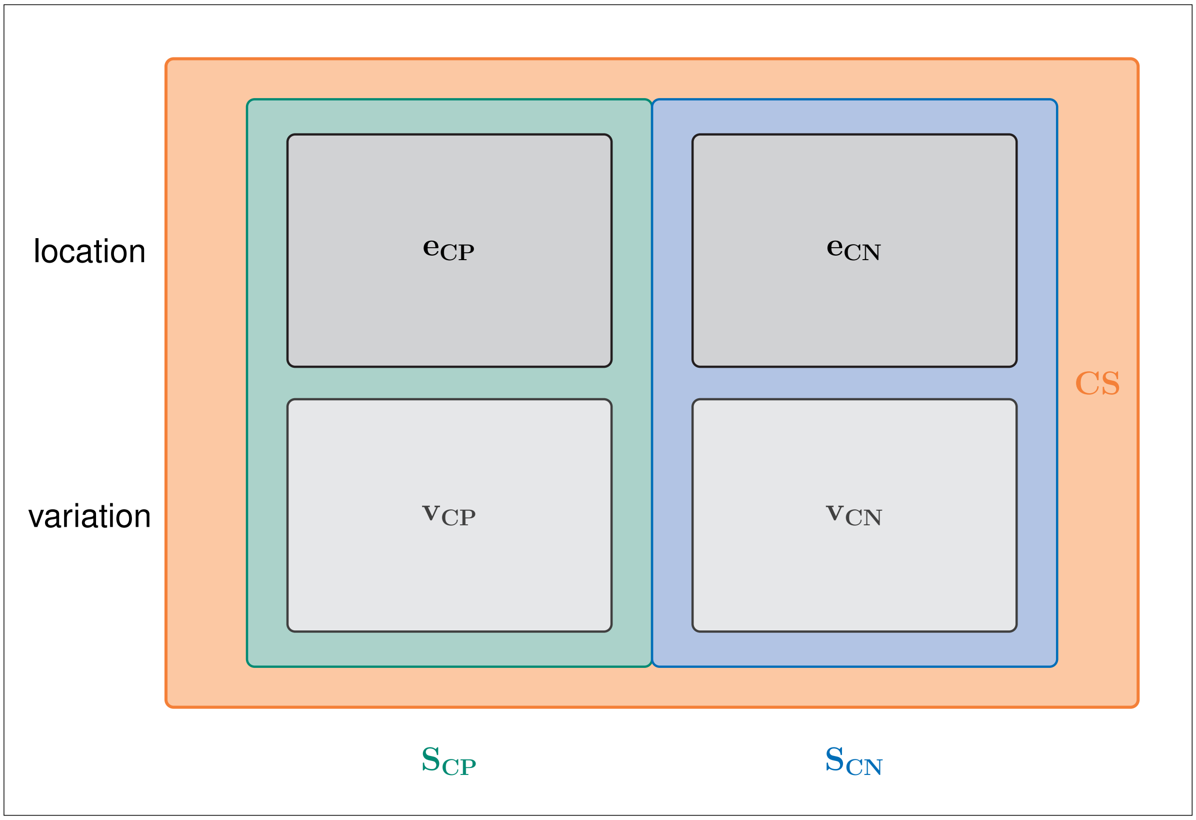


Figure 2.3 Illustration of the conditional performance score's composition with conditional power and conditional sample size sub-scores S_{CP} and S_{CN}, location components e_{CP}, e_{CN} and variation components v_{CP}, v_{CN}.

possible variance $Var^{max}(CN)$ or $Var^{max}(CP)$ and take the square root thereof,

$$v_{CN}(\Delta) = 1 - \sqrt{\frac{Var(CN(Z_1))}{Var^{max}(CN)}}, \tag{2.11}$$

as well as

$$v_{CP}(\Delta) = 1 - \sqrt{\frac{Var(CP(Z_1))}{Var^{max}(CP)}}. \tag{2.12}$$

The maximal deviations and maximally possible variances are defined independently of the true standardized treatment effect and are also given in Table 2.1. Since the score is also supposed to be applied to group sequential study designs with constant stage 2 sample sizes, the variation component for the conditional sample size v_{CN} is defined as 1 in that case. In accordance with the global performance score by Liu et al. (2008), the conditional performance score can be evaluated pointwise or averaged over a range of plausible effect sizes.

A rule of thumb for differentiating low, moderate and high performance score values is given in *Thesis Article 1* (Herrmann et al., 2020, Chapter 4.9).

Table 2.1 Values underlying the conditional performance score.

Performance measure	Target value for $n_\Delta^{fix} \leq n_{max}$ and $\Delta \neq 0$	Target value for $n_\Delta^{fix} > n_{max}$ or $\Delta = 0$	Maximally possible deviation	Maximally possible variance
CN	n_Δ^{fix}	n_1	$n_{max} - n_1$	$((n_{max} - n_1)/2)^2$
CP	$1 - \beta$	α	$1 - \alpha$	$((1 - 0)/2)^2$

CN: Conditional sample size; CP: Conditional power; n_Δ^{fix}: Sample size for standardized effect size Δ in fixed sample size design; n_{max}: maximally feasible sample size; n_1: first stage sample size; α: global significance level; $1 - \beta$: anticipated conditional power value.

With this new performance score, it is now possible to evaluate and compare different sample size recalculation rules. More specifically, the power and sample size sub-scores as well as their location and variation components allow identifying sources for good and bad performance. By this new methodological research, sample size recalculation rules can be developed with a clear focus on performance improvement.

2.2.2 *Thesis Article 2*: Improving sample size recalculation with smoothing corrections

With the new performance score, there exists now an option to judge and evaluate sample size recalculation rules. It can provide a basis to improve existing approaches. One common point of criticism with respect to many published sample size recalculation rules is the high variability in the recalculated sample size (Dragalin, 2006, Bauer et al., 2016). One reason for this feature is a sudden and abrupt increase from the first stage's to the maximum sample size (cf. also green line in Figure 2.2). This "jump" in sample size is part of many sample size recalculation functions, as for very small interim effect sizes a continuation of the trial to the second stage might not seem justified (second stage sample size equals 0), whereas starting from some minimally relevant interim effect, the second stage sample size is increased to meet a certain conditional power value which results in the maximum sample size n_{max}. The lowest interim effect starting from which a large second stage sample size seems appropriate is, however, rather arbitrary. A specific rule might for example suggest a trial stop for an observed interim effect of 0.22 but the same rule can suggest the continuation of the trial with the maximum sample size for an observed interim effect of 0.23 (Herrmann and Rauch, 2021). The idea of *Thesis Article 2* (Herrmann and Rauch, 2021) was therefore to evaluate whether a smoothed increase in sample size from n_1 to n_{max} improves the performance with respect to a reduced random variability. In the article, several smoothing corrections were proposed that can be combined with established recalculation approaches including such a jump in sample size, like the restricted observed conditional power approach (Figure 2.2). More precisely, five possible smoothing corrections were suggested in *Thesis Article 2* (Herrmann and Rauch, 2021). Figure 2.4 shows the shape of two exemplary smoothing functions when applied to the established restricted observed conditional

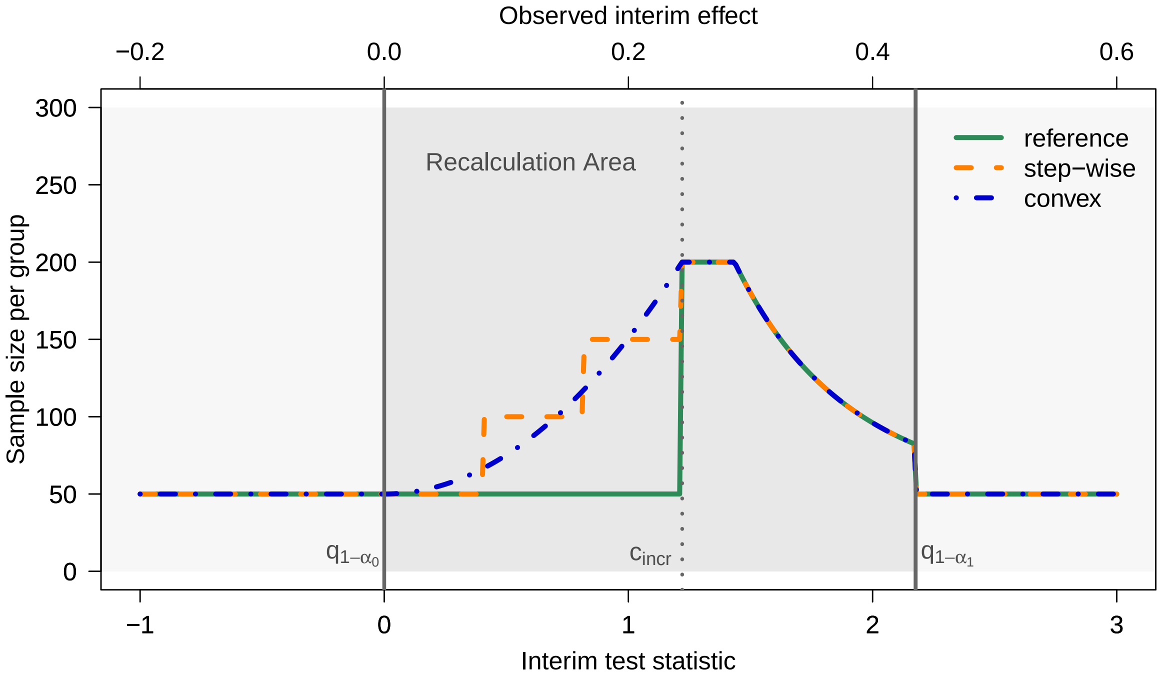


Figure 2.4 Sample size recalculation according to the restricted observed conditional power approach (reference) and combined with step-wise and convex smoothing correction for $n_1 = 50$, $n_{max} = 200$, $\alpha_1 = \alpha_{1+2} = 0.01476$, $\alpha_0 = 0.5$ and with recalculation area $[q_{1-\alpha_0}; q_{1-\alpha_1})$, where c_{incr} describes the smallest value of the interim test statistic that suggests n_{max}. The graphs are similar to Figure 1 in *Thesis Article 2* (Herrmann and Rauch, 2021).

power approach. One can see that the sample size is increased from the stage 1 sample size n_1 to the maximum sample size n_{max} within the interval $[q_{1-\alpha_0}; c_{incr})$, where c_{incr} is the smallest interim test statistic that suggests the maximum sample size. Of course, the shape of the smoothing function can also be formulated analytically as done in *Thesis Article 2*. In the following, the formulas for the total sample size n_{1+2} per group for the so-called step-wise and the so called convex smoothing corrections are exemplary presented. Step-wise smoothing as plotted in Figure 2.4 can be expressed as

$$n_{1+2,step-wise}(z_1) = \begin{cases} n_1 & \text{for} \quad z_1 \in [q_{1-\alpha_0}; \frac{c_{incr}}{3}), \\ n_1 + \frac{n_{max}-n_1}{3} & \text{for} \quad z_1 \in [\frac{c_{incr}}{3}; 2 \cdot \frac{c_{incr}}{3}), \\ n_1 + 2 \cdot \frac{n_{max}-n_1}{3} & \text{for} \quad z_1 \in [2 \cdot \frac{c_{incr}}{3}; c_{incr}), \end{cases} \tag{2.13}$$

whereas convex smoothing as plotted in Figure 2.4 can be described by

$$n_{1+2,convex}(z_1) = n_1 + \frac{n_{max} - n_1}{{c_{incr}}^2} \cdot {z_1}^2 \qquad \text{for} \quad \mathrm{z}_1 \in [\mathrm{q}_{1-\alpha_0}; \mathrm{c}_{\mathrm{incr}}). \tag{2.14}$$

The sample size in the remaining recalculation area, $[c_{incr}; q_{1-\alpha_1})$, is given by the reference sample size function, thus in here by the restricted observed conditional power approach. The smoothing corrections were proposed to improve the performance of existing sample size recalculation rules with respect to the variability of the conditional sample size and the conditional power. *Thesis Article 2* shows that the different smoothing corrections perform differently well with respect to this aim as described in more detail in the following Results Section 3.1.2. Note that another possible approach towards variability reduction in conditional sample size is to recalculate the sample size based on a resampled observed interim test statistic as suggested and evaluated in *Thesis Related Article III* (Herrmann et al., 2021).

2.2.3 *Thesis Article 3*: Optimizing sample size recalculation by constrained optimization in the R-package `adoptr`

Instead of improving existing sample size recalculation rules, an alternative appealing option is to directly determine an "optimal" design with respect to predefined optimality or performance criteria. To follow this approach, an adaptive two-stage design can be interpreted as a five-tuple of parameters

$$\mathcal{D} = (n_1, q_{1-\alpha_0}, q_{1-\alpha_1}, n_2(\cdot), c_2(\cdot)), \tag{2.15}$$

where $n_2(\cdot)$ describes the stage 2 sample size function and $c_2(\cdot)$ the stage 2 critical value function, where the null hypothesis is rejected at the final stage if $Z_2 > c_2(z_1)$ for an observed test statistic $Z_1 = z_1$. In the additional *Thesis Related Article I* of Pilz et al. (2019), we introduced the optimization of the complete adaptive two-stage design $\mathcal{D}$ under a customized score $\tilde{S}$, where the expected overall sample size was chosen as score-example. The mathematical description of the optimization problem to determine the optimal parameters $\mathcal{D}$ with respect to minimization of a customized score $\tilde{S}$ subject to constraints on the type I error rate and power at a specific standardized effect size $\Delta > 0$ is given by

$$\begin{aligned} \text{minimize} \quad & \tilde{S}_\Delta(\mathcal{D}) \\ \text{subject to} \quad & \text{type I error rate} \leq \alpha \\ & \text{power} \geq 1 - \beta. \end{aligned} \tag{2.16}$$

By applying Lagrange multipliers, the constrained variational problem (2.16) can be transformed into an unconstrained problem, which in turn can be solved by the help of Euler-Lagrange equations as shown in the additional *Thesis Related Article I* (Pilz et al., 2019).

An alternative to the analytic solution strategy of using Lagrange multipliers and the corresponding Euler-Lagrange equation is to translate the variational problem into a multivariate optimization problem embedded into a finite parameter space, which is related to the approach by Englert and Kieser (2013), and to solve it directly over all five parameters simultaneously by numerical integration. The latter approach allows a convenient evaluation and comparison of different optimality criteria, while the variational approach with the Euler-Lagrange equations can be very complex and time-consuming. The multivariate optimization problem was realized in the newly developed R-package `adoptr`, which is presented in *Thesis Article 3* (Kunzmann et al., 2021). The core idea of the approach is as follows:

- First, it is distinguished whether the observed interim test statistic z_1 falls within the recalculation area $RA = [q_{1-\alpha_0}; q_{1-\alpha_1})$ or not. Note that in a general optimization setting, the values of α_0 and α_1 need not to be pre-specified. If $z_1 \notin RA$, the functions $n_2(\cdot)$ and $c_2(\cdot)$ are constant. If $z_1 \in RA$, a discrete set of pivot points $z_1^{(i)}, i \in \{1, \ldots, k\}$, is required to determine the two functions $n_2(\cdot)$ and $c_2(\cdot)$ by cubic Hermite splines. Hence, an optimization problem of dimension $3 + 2k$ with the parameters

$$\left(n_1, q_{1-\alpha_0}, q_{1-\alpha_1}, n_2^{(1)}, ..., n_2^{(k)}, c_2^{(1)}, ..., c_2^{(k)}\right), \tag{2.17}$$

  has to be solved.

- Since the score function of the optimization problem is often an integral, e.g., the expected overall sample size or the expected conditional performance score, the pivot points are also used as the nodes of the Gauss-Legendre quadrature rule. Using the expected overall sample size $\mathbb{E}[n_{1+2}(Z_1)]$ as score function for illustration, this reads as

$$\begin{aligned} \mathbb{E}[n_{1+2}(Z_1)] &= \mathbb{E}[n_1 + n_2(Z_1)] \\ &= n_1 + \int n_2(z_1) f(z_1) dz_1 \approx n_1 + \sum_{i=1}^{k} \omega_i \cdot n_2(z_1^{(i)}) f(z_1^{(i)}), \end{aligned} \tag{2.18}$$

  where f is the probability density function of Z_1 and $\omega_i \neq 0, i = 1, ..., k$, are the respective weights of the Gauss-Legendre quadrature rule (Rannacher, 2017).

- For solving the optimization problem of dimension $3 + 2k$, the package `nloptr` (Johnson, 2018) is used.

Apart from a fully flexible optimization, it is furthermore possible to fix certain design parameters of the five-tuple $\mathcal{D}$. Moreover, the user may define customized scores where the conditional performance score proposed in *Thesis Article 1* (Herrmann et al., 2020) defines an attractive possibility. The R-package `adoptr` (Kunzmann et al., 2020) supports normally and binary distributed endpoints. So far, we have looked at the performance of a design at predefined point priors under the null or alternative hypothesis. Furthermore, the `adoptr` package also supports continuous prior distributions for the standardized treatment effect, e.g., a truncated normal distribution.

Using the constrained optimization framework of *Thesis Related Article I* (Pilz et al., 2019) and the corresponding software solution proposed in *Thesis Article 3* (Kunzmann et al., 2021), it is now possible to directly determine an optimal design with sample size recalculation with respect to arbitrary constraints and optimization criteria.

3 Results

The three approaches developed within this thesis, which were presented above, were evaluated by means of Monte-Carlo simulation studies and/or clinical trial examples to illustrate how the methods behave in different realistic data situations. In the underlying *Thesis Articles* (Herrmann et al., 2020, Herrmann and Rauch, 2021, Kunzmann et al., 2021), extensive results for the new performance score, the smoothing corrections and the R-package adoptr are provided and discussed. In here, the main results of the *Thesis Articles* are briefly summarized. To illustrate the connection of the three *Thesis Articles*, additional results not published in the related articles are presented for a clinical trial example to which all three new methods were applied. All results were generated by using the software R (R Core Team, 2021, Version 4.0.3). The simulation and parameter settings for the underlying Monte-Carlo simulation studies which allow to reproduce the results can be deduced from the corresponding *Thesis Articles*. These rather technical settings are omitted here for the sake of simplicity.

3.1 Results of Thesis Articles

3.1.1 *Thesis Article 1*: Results for the conditional performance score

For the evaluation of the conditional performance score, we applied the score to four established sample size recalculation approaches, namely

1) the observed conditional power approach (OCP), e.g., described in Posch et al. (2003),

2) the restricted observed conditional power approach (ROCP), e.g., Chen et al. (2004) and also compare Figure 2.2,

3) the promising zone approach (PZ) (Mehta and Pocock, 2011), and

4) the optimization function approach (OF) (Jennison and Turnbull, 2015),

each for a range of different effect sizes. We judged the plausibility of the conditional performance score by comparing its sub-scores for conditional power and sample size (S_{CN} and S_{CP})

and location and variance components (e_{CN}, e_{CP}, v_{CN}, v_{CP}) across the different recalculation approaches and across different effect sizes. Higher variabilities in conditional sample size (cf. definition in Section 2.2.1) or conditional power were reflected by lower variance components (v_{CN} or v_{CP}) and thus resulted in higher sub-scores for conditional sample size and power (S_{CN} and S_{CP}), compare Tables 1 and 2 in *Thesis Article 1* (Herrmann et al., 2020). Similarly, larger deviations of the observed mean values from the target values as depicted in Figure 2 of *Thesis Article 1* (Herrmann et al., 2020) resulted in worse location components (e_{CN} or e_{CP}), which impacts the respective sub-score performance, cf. Table 2 in *Thesis Article 1*. Table 3.1 displays the main results of *Thesis Article 1* given by the point-wise score and its component-values for varying effect sizes.

Table 3.1 Performance score (CS), sub-scores (S_{CN}, S_{CP}) and score components (e_{CN}, v_{CN}, e_{CP}, v_{CP}).

Δ	Approach	e_{CN}	v_{CN}	S_{CN}	e_{CP}	v_{CP}	S_{CP}	CS
0.0	OCP	0.053	0.680	0.366	0.762	0.412	0.587	0.477
(–)	ROCP	0.851	0.359	0.605	0.868	0.385	0.626	0.615
	PZ	0.617	0.699	0.658	0.843	0.448	0.646	0.652
	OF	0.438	0.391	0.414	0.773	0.378	0.576	0.495
0.1	OCP	0.090	0.592	0.341	0.680	0.361	0.520	0.431
(1571)	ROCP	0.790	0.291	0.540	0.789	0.292	0.541	0.541
	PZ	0.595	0.651	0.623	0.766	0.361	0.564	0.593
	OF	0.396	0.383	0.389	0.686	0.320	0.503	0.446
0.2	OCP	0.138	0.512	0.325	0.593	0.349	0.471	0.398
(395)	ROCP	0.728	0.255	0.491	0.701	0.236	0.468	0.480
	PZ	0.576	0.622	0.599	0.681	0.309	0.495	0.547
	OF	0.362	0.388	0.375	0.594	0.302	0.448	0.411
0.3	OCP	0.965	0.451	0.708	0.705	0.376	0.540	0.624
(177)	ROCP	0.493	0.247	0.370	0.601	0.220	0.410	0.390
	PZ	0.605	0.593	0.599	0.619	0.292	0.456	0.527
	OF	0.822	0.403	0.612	0.710	0.325	0.518	0.565
0.4	OCP	0.598	0.407	0.502	0.787	0.427	0.607	0.555
(101)	ROCP	0.949	0.278	0.613	0.701	0.249	0.475	0.544
	PZ	0.869	0.579	0.724	0.718	0.311	0.515	0.619
	OF	0.679	0.421	0.550	0.800	0.376	0.588	0.569
0.5	OCP	0.433	0.386	0.410	0.850	0.503	0.676	0.543
(65)	ROCP	0.686	0.329	0.508	0.782	0.313	0.547	0.527
	PZ	0.635	0.594	0.614	0.795	0.362	0.579	0.597
	OF	0.466	0.435	0.450	0.870	0.452	0.661	0.556

Δ: standardized effect size; OCP: observed conditional power approach; ROCP: restricted observed conditional power approach; PZ: promising zone approach; OF: optimization function approach; numbers in brackets present the required sample sizes in the fixed design. The table corresponds to an excerpt of Tables 1 and 2 in *Thesis Article 1* (Herrmann et al., 2020). For a detailed parameter description, see *Thesis Article 1* (Herrmann et al., 2020).

Under the null hypothesis, the promising zone approach performs best with respect to the conditional sample size and power sub-scores (cf. Table 3.1, Columns 5 and 8) and therefore also has the best total conditional performance score compared to the other three recalculation approaches (cf. Table 3.1, Column 9). For $\Delta = 0.1$ and $\Delta = 0.2$, the ranking with respect to the conditional performance score remains the same with the promising zone approach performing best (cf. Table 3.1, Column 9). While the observed conditional power approach was the worst performing approach for $\Delta \leq 0.2$, it is the best performing approach for $\Delta = 0.3$. This stems from the fact that the recalculated sample sizes for $\Delta = 0.2$ and $\Delta = 0.3$ are similar but the target sample size values change from $n_1 = 50$ to 177 such that the conditional sample size sub-score value is enormously increased (cf. Table 3.1, Column 3). The ranking for $\Delta = 0.4$ and $\Delta = 0.5$ is again the same with the promising zone approach performing best due the good conditional sample size performance (cf. Table 3.1, Column 5).

Note that the primary goal of *Thesis Article 1* was to illustrate the application of the conditional performance score and not to find a definite ranking of different recalculation approaches. The published recalculation rules all include different parameters and possibilities to tune them, so the constellations considered in here might not have been the optimal ones with respect to the new score.

3.1.2 *Thesis Article 2*: Results for sample size recalculation with smoothing corrections

In *Thesis Article 2* (Herrmann and Rauch, 2021), we have evaluated how to overcome abrupt jumps in sample size in standard sample size recalculation rules to reduce the variability in conditional sample sizes and conditional power. For the sake of illustration, we have chosen the restricted observed conditional power approach as a reference recalculation rule, combined it with five different newly proposed smoothing corrections, and compared the resulting new approaches. Comparison was performed with respect to conditional power, conditional sample size and the conditional performance score from *Thesis Article 1* (Herrmann et al., 2020) for a range of true underlying effect sizes, stage 1 and 2 sample size constellations as well as different adjustment strategies for multiple testing. Table 3.2 displays the main results as published in Supplementary Table 4 of *Thesis Article 2* (Herrmann and Rauch, 2021) for a multiplicity adjustment according to Pocock (1977) and a range of effect sizes. Note that unlike the other results presented in this thesis, the global significance level was fully exhausted in *Thesis Article 2* by choosing $\alpha_1 = \alpha_{1+2} = 0.01476$ instead of $\alpha_1 = \alpha_{1+2} = 0.0147$. However, the conditional performance score results from Table 3.2 differ only by at most 0.001 from choosing significance levels $\alpha_1 = \alpha_{1+2} = 0.0147$ (data not shown).

The results for other sample size constellations and other multiplicity adjustments were comparable and can be deduced from Table 1 and Supplementary Tables 2, 3 and 5 in *Thesis Article 2* (Herrmann and Rauch, 2021). For all smoothing corrections, the expected second stage sample size naturally increases as the smoothing causes a sample size increase in an area where the sample size usually jumps from the interim to the maximum sample size, compare Figure 2.4. As a consequence, also the average conditional power is naturally increased (cf. Table 3.2, Column 5). The smoothing approaches cause a variance reduction in the conditional sample size compared to the standard sample size recalculation approach for most considered effect

Table 3.2 Performance evaluation of smoothing corrections.

Δ	Approach	$\mathbb{E}[CN]$	$\mathrm{Var(CN)}$	S_{CN}	$\mathbb{E}[CP]$	$Var(CP)$	S_{CP}	CS
0.0	Reference	72.403	2315.261	0.605	0.154	0.094	0.627	0.616
(–)	Linear	126.096	2039.918	0.445	0.236	0.089	0.594	0.519
	Step-wise	106.867	2331.292	0.489	0.218	0.087	0.606	0.547
	Sigmoid	116.892	3561.255	0.379	0.241	0.091	0.588	0.483
	Concave	146.366	2003.897	0.380	0.249	0.088	0.588	0.484
	Convex	105.826	2478.141	0.482	0.222	0.090	0.598	0.540
0.1	Reference	81.548	2836.295	0.540	0.231	0.125	0.540	0.540
(1571)	Linear	132.947	1986.652	0.426	0.317	0.108	0.522	0.474
	Step-wise	115.584	2318.376	0.460	0.298	0.108	0.531	0.496
	Sigmoid	126.247	3403.540	0.357	0.323	0.108	0.518	0.437
	Concave	150.776	1884.502	0.375	0.331	0.105	0.519	0.447
	Convex	115.117	2534.299	0.447	0.304	0.111	0.524	0.486
0.2	Reference	90.800	3125.985	0.491	0.317	0.146	0.468	0.480
(395)	Linear	138.420	1838.551	0.419	0.403	0.114	0.469	0.444
	Step-wise	123.562	2130.691	0.447	0.385	0.117	0.474	0.460
	Sigmoid	134.347	3045.228	0.351	0.410	0.113	0.466	0.409
	Concave	153.599	1769.403	0.374	0.416	0.109	0.469	0.422
	Convex	123.242	2369.011	0.431	0.390	0.119	0.468	0.450
0.3	Reference	100.195	3191.827	0.370	0.411	0.152	0.410	0.390
(177)	Linear	142.079	1744.103	0.608	0.495	0.107	0.517	0.563
	Step-wise	129.620	1929.064	0.552	0.477	0.112	0.500	0.526
	Sigmoid	140.454	2656.991	0.537	0.502	0.105	0.523	0.530
	Concave	153.886	1757.232	0.646	0.507	0.101	0.532	0.589
	Convex	130.271	2158.494	0.537	0.484	0.113	0.502	0.520
0.4	Reference	107.786	2932.545	0.613	0.509	0.141	0.475	0.544
(101)	Linear	140.663	1703.894	0.589	0.579	0.091	0.585	0.587
	Step-wise	131.213	1741.666	0.618	0.564	0.098	0.566	0.592
	Sigmoid	140.152	2332.493	0.544	0.585	0.089	0.592	0.568
	Concave	149.193	1816.565	0.552	0.588	0.085	0.600	0.576
	Convex	132.134	1952.366	0.599	0.571	0.097	0.571	0.585
0.5	Reference	111.499	2534.837	0.507	0.587	0.118	0.547	0.527
(65)	Linear	136.913	1662.355	0.486	0.644	0.070	0.656	0.571
	Step-wise	129.872	1576.203	0.517	0.632	0.077	0.637	0.577
	Sigmoid	137.159	2084.516	0.453	0.649	0.067	0.663	0.558
	Concave	142.997	1856.021	0.451	0.651	0.064	0.670	0.560
	Convex	130.829	1754.830	0.499	0.637	0.075	0.642	0.571

Δ: standardized effect size; CS: conditional performance score; $\mathbb{E}[CN]$: expected conditional sample size per group; $Var(CN)$: variance of conditional sample size; $\mathbb{E}[CP]$: expected conditional power; $Var(CP)$: variance of expected conditional power; Reference: restricted observed conditional power approach (ROCP) without smoothing correction; Linear / Step-wise / Sigmoid / Concave / Convex: ROCP with five different candidate smoothing corrections; numbers in brackets present the required sample sizes in the fixed design. The table corresponds to an excerpt of Supplementary Table 4 in *Thesis Article 2* (Herrmann and Rauch, 2021). For a detailed parameter description, see also *Thesis Article 2* (Herrmann and Rauch, 2021).

sizes (Column 4). Additionally, the variance in conditional power is reduced by the smoothing corrections for all considered effect sizes (Column 6). When considering the conditional performance score, which summarizes variance and location aspects of sample size and conditional power, the smoothing corrections only provide a net benefit for underlying standardized effect sizes $\Delta \geq 0.3$ (Column 7). Overall, no smoothing correction always outperforms the other smoothing corrections. However, the step-wise followed by the convex smoothing approach show a reasonable or good performance for most considered effect sizes such that these two smoothing corrections are recommended from the five candidates investigated.

3.1.3 *Thesis Article 3*: Results for the optimization approach implemented in the R-package `adoptr`

In *Thesis Article 3* (Kunzmann et al., 2021), we present the new R-package `adoptr` (Kunzmann et al., 2020) implementing a numerical approach to find an optimal design with sample size recalculation. The article provides several examples how to apply this approach in practice. By these examples, we see, e.g., that the optimization under uncertainty presented by continuous priors, comes at the price of a larger expected sample size as well as a higher number of iterations for the optimization procedure. Similarly, adding constraints, e.g., a conditional power constraint, or fixing certain parameters also increases the expected sample size. However, *Thesis Article 3* puts less focus on the results of specific optimization problems but on the software implementation of their solution.

The key design principles of `adoptr` presented in *Thesis Article 3*, which can be interpreted as the main results of the underlying programming task, are interactivity, reliability and extensibility (Kunzmann et al., 2021). Considering interactivity, `adoptr` consists of a step-wise problem formulation since it is supposed to motivate the investigation of different adaptive design optimization scenarios. This goes in hand with the software being open-source and available on CRAN (2021) for transparency reasons. Furthermore, there exists a broad online documentation for the functions as well as vignettes to support the exploration of different optimization possibilities. The second design principle is reliability of software for clinical trials. Therefore, `adoptr` uses the R-package `testthat` (Wickham, 2020) for conducting an extended test suite to detect and localize errors during development as described in *Thesis Article 3* (Kunzmann et al., 2021). In addition, there is an extensive validation report that compares results obtained by numerical integration with simulated results as well as with the corresponding results of other available R-packages, e.g., `rpact` (Wassmer and Pahlke, 2021). Extensibility is the third design principle of `adoptr`. An optimal trial design highly depends on the pre-defined optimality criteria, which is the score and the constraints. Within *Thesis Article 3* (Kunzmann et al., 2021), we present some exemplary suggestions but `adoptr` provides explicitly room for defining customized scores and constraints.

3.2 Clinical trial example

To illustrate how the new methods are related, the methodology of *Thesis Articles 1, 2 and 3* is applied to a clinical trial example in the field of Alzheimer's disease. The aim is to compare a therapy of a cholinesterase inhibitor and an add-on therapy (intervention group I) with the monotherapy of receiving only the cholinesterase inhibitor (control group C). The primary endpoint is given by the change in the ADAS-Cog from baseline to month 6. The ADAS-Cog score was initially proposed by Rosen et al. (1984) and is a score ranging between 0 and 70 where higher values indicate a worse outcome. For the sake of simplicity, the score values are assumed to be approximately normally distributed. However, this normal assumption is less important if the stage 1 sample size is sufficiently large. The one-sided hypotheses are formulated as

$$\begin{gathered} H_0 : (\mu_{I,baseline} - \mu_{I,6months}) - (\mu_{C,baseline} - \mu_{C,6months}) \leq 0 \\ \text{and} \\ H_1 : (\mu_{I,baseline} - \mu_{I,6months}) - (\mu_{C,baseline} - \mu_{C,6months}) > 0. \end{gathered} \tag{3.1}$$

Wang et al. (2015) describe common shortcomings of clinical trials on Alzheimer's disease, such as a high pre-trial insecurity about the treatment effect. Cummings et al. (2012) mention the use of adaptive designs in Alzheimer's disease studies to use accumulating data for making trial modifications. For the sake of illustration, an adaptive design with the possibility to recalculate the sample size at interim and the possibility to stop early for efficacy or futility will be applied. To test the above null hypothesis, a one-sided approximate Z-test (cf. Section 2.1) with a global one-sided type I error of $\alpha = 0.025$ is applied. The maximum sample size per group is restricted to $n_{max} = 450$, similar to Wang et al. (2015, Supplementary Table 1). For the first stage, a sample size of $n_1 = 70$ per group was chosen. Note that by this sample size setting, the performance of our methods is illustrated in an alternative setting compared to the simulations provided above, where $n_1 = 50$ and $n_{max} = 200$. With this maximum sample size per group, the standard group sequential approach with a constant stage 2 sample size $n_2 = 450 - 70 = 380$ and the above specified parameters yields approximately 67% power for a standardized treatment effect of $\Delta = 0.2$ and approximately 94% power for $\Delta = 0.3$. This can serve as a reference in the remainder. A binding futility stopping boundary of $\alpha_0 = 0.5$ was chosen such that the trial is stopped at interim if the effect size points in the opposite direction. The local significance levels were chosen as $\alpha_1 = \alpha_{1+2} = 0.0147$ according to Pocock (1977). Moreover, the combined test statistic of both stages is defined by the inverse normal combination test (Lehmacher and Wassmer, 1999). The above described adaptive design in principle allows incorporating an arbitrary rule for sample size recalculation. When applying sample size recalculation, the simulation of different adaptive design options and different data scenarios in the planning stage is inevitable (Mayer et al., 2019). As the aim of this example is to illustrate the relation of the *Thesis Articles'* methodology, the following four sample size recalculation approaches are compared:

1) Restricted observed conditional power approach as an established reference design (cf. Figure 2.2),

2) Restricted observed conditional power approach with the newly developed step-wise smoothing correction (cf. Equation (2.13)), *Thesis Article 2*,

3) Restricted observed conditional power approach with the newly developed convex smoothing correction (cf. Equation (2.14)), *Thesis Article 2*, and

4) Newly developed optimization approach with the conditional performance score as the optimization function with fixed critical values and futility bound, *Thesis Articles 1 and 3*.

Performance results for approaches 1) to 3) are obtained by a Monte-Carlo simulation study with 10'000 simulation runs. Performance results for approach 4) are achieved by the R-package `adoptr` presented in *Thesis Article 3* (Kunzmann et al., 2021) with numerical integration using $k = 7$ interpolation points and cubic Hermite splines. For determining the corresponding optimal sample size curve as given in Figure 3.1, we thereby rely on a point prior at $\Delta = 0.3$, which is also the effect that can be detected with more than 90% power. Note that due to the inverse normal combination approach, the c_2-function in the optimization approach is fixed by

$$\begin{aligned} & Z_{1+2} = \frac{w_1 Z_1 + w_2 Z_2}{\sqrt{w_1^2 + w_2^2}} > q_{1-\alpha_{1+2}} \\ \Leftrightarrow \quad & Z_2 > \frac{q_{1-\alpha_{1+2}}\sqrt{w_1^2 + w_2^2} - w_1 Z_1}{w_2} =: c_2(Z_1). \end{aligned} \tag{3.2}$$

The four approaches are evaluated and compared with respect to conditional performance measures, i.e., expected conditional sample size and its variation, expected conditional power and its variation, as well as the new conditional performance score with its sub-scores for standar-

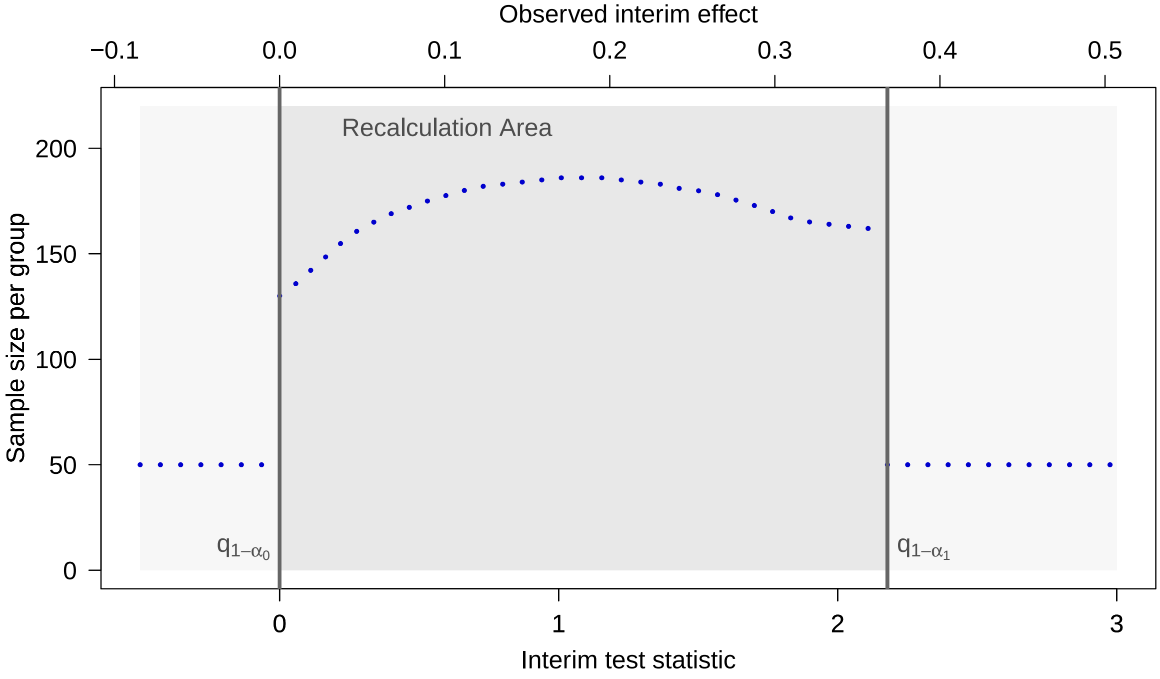


Figure 3.1 Sample size curve for the optimization approach described in Section 3.2 with a point prior at the standardized treatment effect $\Delta = 0.3$, stage 1 sample size $n_1 = 70$ and maximum sample size $n_{max} = 450$ per group, adjusted significance levels $\alpha_1 = \alpha_{1+2} = 0.0147$ and binding futility stopping bound $\alpha_0 = 0.5$.

Table 3.3 Performance results for the clinical trial example in Section 3.2.

Δ	Approach	$\mathbb{E}[CN]$	$Var(CN)$	S_{CN}	$\mathbb{E}[CP]$	$Var(CP)$	S_{CP}	CS
0.1	1) Reference	165.402	17931.270	0.522	0.337	0.150	0.453	0.488
(1571)	2) Step	237.384	14088.970	0.467	0.391	0.124	0.459	0.463
	3) Convex	236.804	15215.172	0.456	0.397	0.126	0.455	0.455
	4) Optimal	172.833	190.514	0.828	0.274	0.088	0.433	0.631
0.2	1) Reference	185.212	17816.854	0.374	0.450	0.151	0.432	0.403
(395)	2) Step	247.462	13039.132	0.506	0.500	0.117	0.505	0.505
	3) Convex	248.146	14152.321	0.495	0.506	0.117	0.507	0.501
	4) Optimal	174.263	142.507	0.682	0.377	0.104	0.460	0.571
0.3	1) Reference	197.675	15675.805	0.642	0.562	0.129	0.519	0.581
(177)	2) Step	243.885	11960.453	0.623	0.602	0.093	0.593	0.608
	3) Convex	245.288	13010.846	0.609	0.607	0.092	0.597	0.603
	4) Optimal	174.377	109.963	0.972	0.484	0.105	0.514	0.743
0.4	1) Reference	196.729	12844.165	0.575	0.644	0.096	0.611	0.593
(101)	2) Step	229.266	10836.494	0.556	0.674	0.066	0.679	0.618
	3) Convex	230.834	11709.808	0.543	0.678	0.064	0.684	0.614
	4) Optimal	173.477	92.101	0.876	0.582	0.093	0.583	0.729
0.5	1) Reference	194.327	10840.641	0.555	0.705	0.063	0.700	0.628
(65)	2) Step	215.179	10026.244	0.538	0.726	0.040	0.762	0.650
	3) Convex	216.827	10834.320	0.526	0.729	0.038	0.768	0.647
	4) Optimal	172.106	81.217	0.832	0.664	0.074	0.657	0.745

Δ: standardized effect size; $\mathbb{E}[CN]$: expected conditional sample size per group; $Var(CN)$: variance of conditional sample size; $\mathbb{E}[CP]$: expected conditional power; $Var(CP)$: variance of expected conditional power; S_{CN}, S_{CP}: conditional performance sub-scores for sample size and power; CS: conditional performance score; 1) Reference: restricted observed conditional power approach as reference; 2) Step: restricted observed conditional power approach combined with step-wise smoothing; 3) Convex: restricted observed conditional power approach combined with convex smoothing; 4) Optimization: optimization approach with R-package `adoptr`; numbers in brackets present the required sample sizes in the fixed design. The underlying parameter values for this additional simulation study are given in the description above in this Section 3.2.

dized treatment effects Δ ranging between 0.1 and 0.5 by steps of 0.1. The results are given in Table 3.3.

All the newly developed sample size approaches counteract steeply increasing sample size curves and therefore reduce the variability in conditional sample size (cf. Column 4).
More precisely, the smoothing corrections 2) and 3) reduce the sample size variation by increasing the conditional sample size (cf. Column 3) and therefore also increase the conditional power (cf. Column 6) compared to reference approach 1). The optimization approach 4) reduces the variability very strongly by smaller expected sample sizes (cf. Column 3) and therefore decreased conditional power values (cf. Column 6) compared to approach 1).
The conditional performance sub-scores also reflect these observations. The variance reduction in conditional sample size for approaches 2) and 3) achieves only sometimes an improvement in the sample size sub-score compared to the reference approach 1) (cf. Column 5), due to a lower performance in the location component of the performance score. Approach 4) reduces the variance of the conditional sample size tremendously, such that it always leads to high con-

ditional sample size sub-score values (cf. Column 5). The conditional power sub-scores for approaches 2) and 3) outperform reference approach 1) (cf. Column 8) due to higher conditional power values combined with smaller variances. Approach 4), however, almost always has a worse conditional power sub-score compared to the other three approaches 1) to 3) (cf. Column 8) due to a worse performance with respect to location. As the differences of the four approaches are stronger for the conditional sample size sub-score than for the conditional power sub-score values, approach 4) is the best performing approach with respect to the overall conditional performance score (cf. Column 9). However, it is also valid to judge the sub-scores separately, which allows different conclusions.

4 Discussion

The **aim of this thesis** was to overcome shortcomings related to sample size recalculation in adaptive study designs, which was addressed by the following contributions:

1. Comprehensive performance evaluation and comparison of sample size recalculation rules in adaptive designs by a **newly developed performance score** in *Thesis Article 1* (Herrmann et al., 2020),

2. Improving established sample size recalculation rules by reducing their variability in conditional sample size and conditional power with **newly developed smoothing corrections** in *Thesis Article 2* (Herrmann and Rauch, 2021),

3. Optimizing adaptive two-stage designs with sample size recalculation by a **newly developed constrained optimization framework** in *Thesis Article 3* (Kunzmann et al., 2021),

   and

4. **Newly developed software** for all methods described in all *Thesis Articles*.

Additional methods related to these four aspects were also proposed in the *Thesis Related Articles I to V*.

Due to the lack of a clear guidance on how to choose and compare specific sample size recalculation rules, we suggested a **conditional performance score** for a fair and comprehensive judgment and comparison of sample size recalculation rules. As described in Section 2, there exist the global and the conditional perspective for evaluating adaptive designs with sample size recalculation. Since existing performance scores refer only to the global evaluation perspective (Liu et al., 2008, Wu and Cui, 2012, Fang et al., 2018), we filled the remaining gap and introduced a performance score from the conditional perspective in *Thesis Article 1* (Herrmann et al., 2020). Within the score, we incorporated components evaluating the variance of the conditional power and sample size, which was not addressed in the literature so far, and stressed the importance of these variance components since the interim effect is a random variable. The conditional performance score can be reported as an average performance measure over a range of reasonable effect sizes or separately. Moreover, the sub-scores for conditional sample size and conditional power as well as the location and variation components can be reported and interpreted separately.

To improve existing sample size recalculation rules, we introduced **smoothing corrections** to be combined with standard recalculation rules in *Thesis Article 2* (Herrmann and Rauch, 2021) to reduce the high variability in the conditional sample size (Dragalin, 2006, Bauer et al., 2016) and conditional power, and therefore to indirectly address the randomness of the observed interim effect. We evaluated a selection of smoothing classes exemplary combined with the restricted observed conditional power approach. We observed a reduced variance in conditional power and in conditional sample size. Moreover, the smoothing corrections increased the overall conditional performance for medium and large observed standardized effect sizes. Therefore, they present a possible improvement strategy for sample size recalculation strategies based on conditional power arguments and are easy to apply.

In *Thesis Article 3* (Kunzmann et al., 2021), we addressed the **optimization** of adaptive two-stage designs with sample size recalculation. The underlying `R`-package `adoptr` (Kunzmann et al., 2020) provides a tool with great flexibility. There exists a range of software for designing group sequential trials, e.g., ADDPLAN® (ICON plc, 2019), East® (Cytel, 2020), PASS® (NCSS, 2021) and SAS® (SAS Institute Inc, Cary, NC, 2020) and software related to their optimization, e.g., the `R`-packages `OptGS` (Wason and Burkardt, 2015) and `rpact` (Wassmer and Pahlke, 2021). Software for adaptive study designs is rather rare, e.g., the `R`-package `adaptTest` (Vandemeulebroecke, 2009). The `R`-package `adoptr` is a software solution solving constrained optimization problems. This approach can be interpreted as an extension of the approach by Jennison and Turnbull (2015) since it provides room for choosing arbitrary objective functions. With the possibility of formulating specific target power and/or maximum sample size constraints, the newly suggested optimization approach can address common points of criticism of adaptive designs with sample size recalculation as formulated by Bauer and Köhne (1994) and Levin et al. (2013). Moreover, the `R`-package `adoptr` is an elegant solution if one does not want to base sample size recalculation rules on conditional power arguments taking into account that their use is controversially discussed in the literature (Bartroff and Lai, 2008, Jennison and Turnbull, 2015, Levin et al., 2013).

The `R`-package `adoptr` underlying the optimization approach is available on CRAN (2021) and therefore addresses the lack of **software** for adaptive designs (Bauer et al., 2016). In addition, the `R`-code underlying the other *Thesis Articles* is also publicly available on GitHub (`https://github.com/shareCH/SSR-conditional-score`; `https://github.com/shareCH/SSR-smoothing-corrections`).

To outline the relation of the three *Thesis Articles*, a **clinical trial example** was included in this thesis for illustration. We have seen that both the smoothing corrections combined with a standard sample size recalculation rule and the optimization approach optimizing the stage 2 sample size curve performed better than the reference sample size recalculation rule when assessed with the new conditional performance score. In particular, we could see that the resulting “optimal” stage 2 sample size curve mimics the smoothing corrections as the sample size curve is first increasing and then decreasing. One might argue that a sample size curve should intuitively decrease with increasing observed interim effect. However, the smoothing corrections and the optimization overcome a steep and arbitrary “jump” in the sample size curve, which is also not intuitive. Moreover, concave sample size functions were declared as optimal for specific settings in the literature (Pilz et al., 2020). This can be explained by the influence of the variation components in the performance score. When deciding on a specific sample size recalculation approach, one generally has to choose between the simple, less complex established rules with potential smoothing corrections or other improvement tools, e.g., as proposed in *Thesis Related Article III* (Herrmann et al., 2021), and a more complex numerical solution optimizing specific parameters. Both perspectives seem valid and may even come to similar results in specific settings.

The presented new methodology also comes along with some **limitations** as well as open questions to be addressed in future work. While the new conditional performance score presented in *Thesis Article 1* was primarily developed for normally distributed endpoints, extensions to binary and time-to-event endpoints seem attractive but need some formula adjustments. Furthermore, instead of an equal weighting, it might also be interesting to give a different weight to sample size compared to power or a different weight to location compared to variation. A potential limitation of *Thesis Article 3* is that optimization problems in general come along with the problem of the definition of optimality criteria and prior assumptions. The solutions are optimal with respect to the pre-defined criteria but their definition is the crucial point. Therefore, optimized sample size recalculation rules are complex and can never be considered independently of suitable scoring criteria. In general, it is therefore worth to consider a selection of parameter constellations as well as different constraints. Another limiting factor is that not every constraint combination leads to a solvable optimization problem.

The simulations shown in this work are based on specific design settings and assumptions. However, the underlying methodology could also be applied if the underlying study design is varied. For example, we could also base the conditional power on the assumed treatment effect instead of the observed effect at interim or on a prior distribution, which corresponds to the so-called predictive power (Spiegelhalter et al., 1986) that may be more efficient in some circumstances. Similarly, a different adjustment for multiple testing, e.g., as proposed by O’Brien and Fleming (1979) or Lan and DeMets (1983), might be applied. In the *Thesis Related Article II* (Li et al., 2020), we also proposed alternative views to define futility stopping criteria. Moreover, other combination functions for the stage-wise test statistics, e.g. by Bauer and Köhne (1994), may be used. Note that the general comparison of group sequential study designs with a constant stage 2 sample size and adaptive study designs can be found, e.g., in Tsiatis and Mehta (2003), Kelly et al. (2005), Levin et al. (2013) and is omitted in here.

Within this thesis, I did not further present the results of the **Thesis Related Articles**. *Thesis Related Article I* (Pilz et al., 2019) served as the theoretical foundation for the numerical optimization approach of *Thesis Article 3* (Kunzmann et al., 2021). The application of the new constrained optimization approach in clinical practice was moreover described in *Thesis Related Article IV* (Pilz et al., 2021). In addition, the theory is also extended to unplanned interim analyses in *Thesis Related Article V* (Pilz et al., Submitted). The specific aspect of choosing futility stopping boundaries in an optimal way is addressed in *Thesis Related Article II* by Li et al. (2020). Finally, in addition to smoothing corrections, we also suggested another way of reducing the variability of the recalculated sample sizes by resampling the observed interim effect size. This approach is published in *Thesis Related Article III* (Herrmann et al., 2021).

Regarding **future work**, the open aspects mentioned above will be addressed by me and my colleagues in an already accepted follow-up project funded by the German Research Foundation (grant number RA 2347/4-2).

Bibliography

Bartroff, J. and Lai, T. L. (2008), ‘Efficient adaptive designs with mid-course sample size adjustment in clinical trials’, *Stat Med* **27**(10), 1593–1611. doi: 10.1002/sim.3201.

Bauer, P., Bretz, F., Dragalin, V., König, F. and Wassmer, G. (2016), ‘Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls’, *Stat Med* **35**(3), 325–347. doi: 10.1002/sim.6472.

Bauer, P. and Köhne, K. (1994), ‘Evaluation of experiments with adaptive interim analyses’, *Biometrics* **50**(4), 1029–1041. doi: 10.2307/2533441.

Bauer, P. and König, F. (2006), ‘The reassessment of trial perspectives from interim data – a critical view’, *Stat Med* **25**(1), 23–36. doi: 10.1002/sim.2180.

Bowden, J. and Mander, A. (2014), ‘A review and re-interpretation of a group-sequential approach to sample size re-estimation in two-stage trials’, *Pharm Stat* **13**(3), 163–172. doi: 10.1002/pst.1613.

Bretz, F., Gallo, P. and Maurer, W. (2017), ‘Adaptive designs: The swiss army knife among clinical trial designs?’, *Clin Trials* **14**(5), 417–424. doi: 10.1177/1740774517699406.

Bretz, F., Koenig, F., Brannath, W., Glimm, E. and Posch, M. (2009), ‘Adaptive designs for confirmatory clinical trials’, *Stat Med* **28**(8), 1181–1217. doi: 10.1002/sim.3538.

Chen, Y. J., DeMets, D. L. and Lan, K. G. (2004), ‘Increasing the sample size when the unblinded interim result is promising’, *Stat Med* **23**(7), 1023–1038. doi: 10.1002/sim.1688.

CRAN (2021), *Comprehensive R Archive Network*. `https://cran.r-project.org/` (last accessed: 28/03/2021, 4:02 pm).

Cui, L., Hung, H. J. and Wang, S.-J. (1999), ‘Modification of sample size in group sequential clinical trials’, *Biometrics* **55**(3), 853–857. doi: 10.1111/j.0006-341x.1999.00853.x.

Cummings, J., Gould, H. and Zhong, K. (2012), ‘Advances in designs for alzheimer’s disease clinical trials’, *Am J Neurodegener Dis* **1**(3), 205–216. PMCID: PMC3560467.

Cytel (2020), *East®*. `https://www.cytel.com/software/east` (last accessed: 28/03/2021, 4:03 pm).

Denne, J. S. (2001), ‘Sample size recalculation using conditional power’, *Stat Med* **20**(17-18), 2645–2660. doi: 10.1002/sim.734.

Dmitrienko, A. and Wang, M.-D. (2006), ‘Bayesian predictive approach to interim monitoring in clinical trials’, *Stat Med* **25**(13), 2178–2195. doi: 10.1002/sim.2204.

Dragalin, V. (2006), ‘Adaptive designs: terminology and classification’, *Drug Inf J* **40**(4), 425–435. doi: 10.1177/216847900604000408.

Englert, S. and Kieser, M. (2013), ‘Optimal adaptive two-stage designs for phase ii cancer clinical trials’, *Biom J* **55**(6), 955–968. doi: 10.1002/bimj.201200220.

European Medicines Agency (EMA) through its Committee for Medicinal Products for Human Use (2007), ‘Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design’. `https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf` (last accessed: 28/03/2021, 3:55 pm).

Fang, F., Lin, Y., Shih, W., Lu, S. and Zhu, G. (2018), ‘Evaluation of performance of adaptive designs based on treatment effect intervals’, *Int J Stat Prob* **7**(6), 81–93. doi: 10.5539/ijsp.v7n6p81.

Fedgchin, M., Trivedi, M., Daly, E. J., Melkote, R., Lane, R., Lim, P., Vitagliano, D., Blier, P., Fava, M., Liebowitz, M., Ravindran, A., Gaillard, R., Van Den Ameele, H., Preskorn, S., Manji, H., Hough, D., Drevets, W. C. and Singh, J. B. (2019), ‘Efficacy and safety of fixed-dose esketamine nasal spray combined with a new oral antidepressant in treatment-resistant depression: results of a randomized, double-blind, active-controlled study (TRANSFORM-1)’, *Int J Neuropsychopharmacol* **22**(10), 616–630. doi: 10.1093/ijnp/pyz039.

Food and Drug Administration (FDA) (2019), ‘Guidance for industry: Adaptive design for clinical trials of drugs and biologics’. `https://www.fda.gov/media/78495/download` (last accessed: 28/03/2021, 3:55 pm).

Gao, P., Ware, J. H. and Mehta, C. (2008), ‘Sample size re-estimation for adaptive sequential design in clinical trials’, *J Biopharm Stat* **18**(6), 1184–1196. doi: 0.1080/10543400802369053.

Grayling, M. J. and Wheeler, G. M. (2020), ‘A review of available software for adaptive clinical trial design’, *Clin Trials* **17**(3), 323–331. doi: 10.1177/1740774520906398.

Hager, D. N., Hooper, M. H., Bernard, G. R., Busse, L. W., Ely, E. W., Gaieski, D. F., Hall, A., Hinson, J. S., Jackson, J. C., Kelen, G. D., Levine, M., Lindsell, C. J., Malone, R. E., McGlothlin, A., Rothman, R. E., Viele, K., Wright, D. W., E., S. J. and Martin, G. S. (2019), 'The vitamin C, thiamine and steroids in sepsis (VICTAS) protocol: a prospective, multi-center, double-blind, adaptive sample size, randomized, placebo-controlled, clinical trial', *Trials* **20**(1), 197. doi: 10.1186/s13063-019-3254-2.

Herrmann, C., Kluge, C., Pilz, M., Kieser, M. and Rauch, G. (2021), 'Improving sample size recalculation in adaptive clinical trials by resampling', *Pharm Stat* . doi: 10.1002/pst.2122.

Herrmann, C., Pilz, M., Kieser, M. and Rauch, G. (2020), 'A new conditional performance score for the evaluation of adaptive group sequential designs with sample size recalculation', *Stat Med* **39**(15), 2067–2100. doi: 10.1002/sim.8534.

Herrmann, C. and Rauch, G. (2021), 'Smoothing corrections for improving sample size recalculation rules in adaptive group sequential study designs', *Methods Inf Med* **60**, 1–8. doi: 10.1055/s-0040-1721727.

Hsiao, S. T., Liu, L. and Mehta, C. R. (2019), 'Optimal promising zone designs', *Biom J* **61**(5), 1175–1186. doi: 10.1002/bimj.201700308.

ICH (1998), 'Topic E9: Statistical principles for clinical trials'. European Agency for the Evaluation of Medicinal Products. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf (last accessed: 28/03/2021, 3:55 pm).

ICON plc (2019), *ADDPLAN®*. https://www.iconplc.com/innovation/addplan/ (last accessed: 28/03/2021, 4:00 pm).

Jennison, C. and Turnbull, B. W. (1999), *Group Sequential Methods with Applications to Clinical Trials*, 1st ed., Boca Raton, Chapman and Hall/CRC. doi: 10.1201/9780367805326.

Jennison, C. and Turnbull, B. W. (2015), 'Adaptive sample size modification in clinical trials: start small then ask for more?', *Stat Med* **34**(29), 3793–3810. doi: 10.1002/sim.6575.

Johnson, S. G. (2018), *The NLopt Nonlinear-Optimization Package*. http://ab-initio.mit.edu/nlopt (last accessed: 28/03/2021, 4:04 pm).

Kelly, P. J., Roshini Sooriyarachchi, M., Stallard, N. and Todd, S. (2005), 'A practical comparison of group-sequential and adaptive designs', *J Biopharm Stat* **15**(4), 719–738. doi: 10.1081/BIP-200062859.

Kieser, M. (2018), *Fallzahlberechnung in der medizinischen Forschung: Eine Einführung für Mediziner und Biostatistiker*, 1st ed., Wiesbaden, Springer. doi: 10.1007/978-3-658-20740-3.

Kieser, M. (2020), *Methods and Applications of Sample Size Calculation and Recalculation in Clinical Trials*, 1st ed., Cham, Springer Nature Switzerland. doi: 10.1007/978-3-030-49528-2.

Kieser, M. and Friede, T. (2003), 'Simple procedures for blinded sample size adjustment that do not affect the type I error rate', *Stat Med* **22**(23), 3571–3581. doi: 10.1002/sim.1585.

Koch, A. (2006), 'Confirmatory clinical trials with an adaptive design', *Biom J* **48**(4), 574–585. doi: 10.1002/bimj.200510239.

Kunzmann, K., Pilz, M., Herrmann, C., Rauch, G. and Kieser, M. (2021), 'The adoptr package: Adaptive optimal designs for clinical trials in R', *J Stat Softw* **98**(9), 1–21. doi: 10.18637/jss.v098.i09.

Kunzmann, K., Pilz, M. and Herrmann, C. (2020), *adoptr: Adaptive Optimal Two-Stage Designs in R*. R package version 0.4.1, `https://doi.org/10.5281/zenodo.4110773` (last accessed: 28/03/2021, 4:01 pm).

Lan, K. G. and DeMets, D. L. (1983), 'Discrete sequential boundaries for clinical trials', *Biometrika* **70**(3), 659–663.

Lan, K. G., Hu, P. and Proschan, M. A. (2009), 'A conditional power approach to the evaluation of predictive power', *Stat Biopharm Res* **1**(2), 131–136. doi: 10.1198/sbr.2009.0035.

Lehmacher, W. and Wassmer, G. (1999), 'Adaptive sample size calculations in group sequential trials', *Biometrics* **55**(4), 1286–1290. doi: 10.1111/j.0006-341x.1999.01286.x.

Levin, G. P., Emerson, S. C. and Emerson, S. S. (2013), 'Adaptive clinical trial designs with prespecified rules for modifying the sample size: understanding efficient types of adaptation', *Stat Med* **32**(8), 1259–1275. doi: 10.1002/sim.5662.

Li, X., Herrmann, C. and Rauch, G. (2020), 'Optimality criteria for futility stopping boundaries for group sequential designs with a continuous endpoint', *BMC Med Res Methodol* **20**(1), 1–8. doi: 10.1186/s12874-020-01141-5.

Liu, G. F., Zhu, G. R. and Cui, L. (2008), 'Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval', *Stat Med* **27**(4), 584–596. doi: 10.1002/sim.2998.

Mauer, M., Collette, L., Bogaerts, J. and European Organisation for Research and Treatment of Cancer (EORTC) Statistics Department (2012), 'Adaptive designs at european organisation for research and treatment of cancer (EORTC) with a focus on adaptive sample size re-estimation based on interim-effect size', *Eur J Cancer* **48**(9), 1386–1391. doi: 10.1016/j.ejca.2011.12.024.

Mayer, C., Perevozskaya, I., Leonov, S., Dragalin, V., Pritchett, Y., Bedding, A., Hartford, A., Fardipour, P. and Cicconetti, G. (2019), 'Simulation practices for adaptive trial designs in drug and device development', *Stat Biopharm Res* **11**(4), 325–335. doi: 10.1080/19466315.2018.1560359.

Mehta, C. R. and Patel, N. R. (2006), 'Adaptive, group sequential and decision theoretic approaches to sample size determination', *Stat Med* **25**(19), 3250–3269. doi: 10.1002/sim.2638.

Mehta, C. R. and Pocock, S. J. (2011), 'Adaptive increase in sample size when interim results are promising: a practical guide with examples', *Stat Med* **30**(28), 3267–3284. doi: 10.1002/sim.4102.

Müller, H.-H. and Schäfer, H. (2004), 'A general statistical principle for changing a design any time during the course of a trial', *Stat Med* **23**(16), 2497–2508. doi: 10.1002/sim.1852.

NCSS (2021), *PASS Sample Size 2019®*. `https://www.ncss.com/software/pass/` (last accessed: 28/03/2021, 4:05 pm).

O'Brien, P. C. and Fleming, T. R. (1979), 'A multiple testing procedure for clinical trials', *Biometrics* **35**(3), 549–556. doi: 10.2307/2530245.

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Sydes, M. R., Villar, S. S., Wason, J. M. S., Weir, C. J., Wheeler, G. M., Yap, C. and Jaki, T. (2018), 'Adaptive designs in clinical trials: why use them, and how to run and report them', *BMC Med* **16**(1), 1–15. doi: 10.1186/s12916-018-1017-7.

Pilz, M., Herrmann, C., Rauch, G. and Kieser, M. (Submitted), 'Optimal unplanned design modification in adaptive two-stage trials'.

Pilz, M., Kilian, S. and Kieser, M. (2020), 'A note on the shape of sample size functions of optimal adaptive two-stage designs', *Commun. Stat. – Theory Methods* pp. 1–8. doi: 10.1080/03610926.2020.1776875.

Pilz, M., Kunzmann, K., Herrmann, C., Rauch, G. and Kieser, M. (2019), 'A variational approach to optimal two-stage designs', *Stat Med* **38**(21), 4159–4171. doi: 10.1002/sim.8291.

Pilz, M., Kunzmann, K., Herrmann, C., Rauch, G. and Kieser, M. (2021), 'Optimal planning of adaptive two-stage designs', *Stat Med* **40**, 3196–3213. doi: 10.1002/sim.8953.

Pocock, S. J. (1977), 'Group sequential methods in the design and analysis of clinical trials', *Biometrika* **64**(2), 191–199. doi: 10.1093/biomet/64.2.191.

Posch, M., Bauer, P. and Brannath, W. (2003), 'Issues in designing flexible trials', *Stat Med* **22**(6), 953–969. doi: 10.1002/sim.1455.

Proschan, M. A. (2009), ‘Sample size re-estimation in clinical trials’, *Biom J* **51**(2), 348–357. doi: 10.1002/bimj.200800266.

Proschan, M. A. and Hunsberger, S. A. (1995), ‘Designed extension of studies based on conditional power’, *Biometrics* **51**(4), 1315–1324. doi: 10.2307/2533262.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*. The R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/ (last accessed: 28/03/2021, 4:05 pm).

Rannacher, R. (2017), ‘Numerik 0: Einführung in die Numerische Mathematik’. doi: 10.17885/heiup.206.281 (last accessed: 28/03/2021, 4:00 pm).

Rosen, W. G., Mohs, R. C. and Davis, K. L. (1984), ‘A new rating scale for alzheimer’s disease.’, *Am J Psychiatry* **141**(11). doi: 10.1176/ajp.141.11.1356.

SAS Institute Inc, Cary, NC (2020), *SAS®*. https://www.jmp.com/en_gb/software/clinical-data-analysis-software.html (last accessed: 28/03/2021, 4:06 pm).

Schmidt, R., Faldum, A. and Kwiecien, R. (2018), ‘Adaptive designs for the one-sample log-rank test’, *Biometrics* **74**(2), 529–537.

Shun, Z., Yuan, W., Brady, W. E. and Hsu, H. (2001), ‘Type I error in sample size re-estimations based on observed treatment difference’, *Stat Med* **20**(4), 497–513. doi: 10.1002/sim.531.

Spiegelhalter, D. J. and Freedman, L. S. (1986), ‘A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion’, *Stat Med* **5**(1), 1–13. doi: 10.1002/sim.4780050103.

Spiegelhalter, D. J., Freedman, L. S. and Blackburn, P. R. (1986), ‘Monitoring clinical trials: conditional or predictive power?’, *Control Clin Trials* **7**(1), 8–17. doi: 10.1016/0197-2456(86)90003-6.

Stark, M. and Zapf, A. (2020), ‘Sample size calculation and re-estimation based on the prevalence in a single-arm confirmatory diagnostic accuracy study’, *Stat Methods Med Res* **29**(10), 2958–2971. doi: 10.1177/0962280220913588.

Tsiatis, A. A. and Mehta, C. (2003), ‘On the inefficiency of the adaptive design for monitoring clinical trials’, *Biometrika* **90**(2), 367–378. doi: 10.1093/biomet/90.2.367.

Tymofyeyev, Y. (2014), ‘A review of available software and capabilities for adaptive designs’, *in: Practical Considerations for Adaptive Trial Design and Implementation* . 1st ed, New York, Springer, 139–155. doi: 10.1007/978-1-4939-1100-4_8.

Vandemeulebroecke, M. (2009), *adaptTest: Adaptive Two-Stage Tests.* R package version 1.0, https://CRAN.R-project.org/package=adaptTest (last accessed: 28/03/2021, 4:00 pm).

Wang, G., Kennedy, R. E., Cutter, G. R. and Schneider, L. S. (2015), ‘Effect of sample size re-estimation in adaptive clinical trials for alzheimer’s disease and mild cognitive impairment’, *Alzheimer’s & Dementia: TRCI* **1**(1), 63–71. doi: 10.1016/j.trci.2015.03.002.

Wason, J. M. S. and Burkardt, J. (2015), *OptGS: Near-Optimal and Balanced Group-Sequential Designs for Clinical Trials with Continuous Outcomes*. R package version 1.1.1, `https://CRAN.R-project.org/package=OptGS` (last accessed: 28/03/2021, 4:05 pm).

Wassmer, G. and Brannath, W. (2016), *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, 1st ed., Heidelberg, Springer International Publishing. doi: 10.1007/978-3-319-32562-0.

Wassmer, G. and Pahlke, F. (2021), *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis*. R package version 3.0.4, `https://CRAN.R-project.org/package=rpact` (last accessed: 28/03/2021, 4:06 pm).

Wassmer, G. and Vandemeulebroecke, M. (2006), ‘A brief review on software developments for group sequential and adaptive designs’, *Biom J* **48**(4), 732–737.

Wickham, H. (2020), *testthat: Get Started with Testing*. R package version 3.0.2, `https://cran.r-project.org/web/packages/testthat/` (last accessed: 28/03/2021, 4:07 pm).

Wu, X. and Cui, L. (2012), ‘Group sequential and discretized sample size re-estimation designs: a comparison of flexibility’, *Stat Med* **31**(24), 2844–2857. doi: 10.1002/sim.5395.

Zajicek, J. P., Hobart, J. C., Slade, A., Barnes, D., Mattison, P. G. and MUSEC Research Group (2012), ‘Multiple sclerosis and extract of cannabis: results of the MUSEC trial’, *JNNP* **83**(11), 1125–1132. doi: 10.1136/jnnp-2012-302468.

Zhang, L., Cui, L. and Yang, B. (2016), ‘Optimal flexible sample size design with robust power’, *Stat Med* **35**(19), 3385–3396. doi: 10.1002/sim.6931.

Appendix

Statutory Declaration

I, Carolin Herrmann, by personally signing this document in lieu of an oath, hereby affirm that I prepared the submitted dissertation on the topic “Evaluating and improving sample size recalculation in adaptive clinical study designs” (German title: “Evaluierung und Verbesserung von Fallzahlrekalkulation in adaptiven klinischen Studiendesigns”), independently and without the support of third parties, and that I used no other sources and aids than those stated. All parts which are based on the publications or presentations of other authors, either in letter or in spirit, are specified as such in accordance with the citing guidelines. The sections on methodology and results are exclusively my responsibility. Furthermore, I declare that I have correctly marked all of the methods, the analyses, and the conclusions generated in collaboration with other persons, and that I have correctly marked my own contribution and the contributions of other persons (cf. declaration of contribution). I have correctly marked all texts or parts of texts that were generated in collaboration with other persons. Figures were generated by myself. The opening and closing drawing were painted by myself.

My contributions to any publications to this dissertation correspond to those stated in the below joint declaration made together with the first supervisor. All publications created within the scope of the dissertation comply with the guidelines of the ICMJE (International Committee of Medical Journal Editors; www.icmje.org) on authorship. In addition, I declare that I shall comply with the regulations of Charité – Universitätsmedizin Berlin on ensuring good scientific practice. I declare that I have not yet submitted this dissertation in identical or similar form to another Faculty.

The significance of this statutory declaration and the consequences of a false statutory declaration under criminal law (Sections 156, 161 of the German Criminal Code) are known to me.

Date and signature of doctoral candidate

Declaration of own contribution to the publications

Carolin Herrmann contributed the following to the below listed publications:

***Thesis Article 1*: Carolin Herrmann, Maximilian Pilz, Meinhard Kieser, Geraldine (2020). "A new conditional performance score for the evaluation of adaptive group sequential designs with sample size recalculation." *Statistics in Medicine*, 39(15), 2067–2100. doi: 10.1002/sim.8534.**

Carolin Herrmann developed and tested the score components and wrote more than 60% of the manuscript independently (i.e., the introduction, simulation and discussion section as well as parts 4.7 and 4.9 of the methods section). She conducted the literature search, developed the new methodology, conducted the related simulation studies and analyzed the respective results. Moreover, she generated all figures and tables within the manuscript, i.e., Figures 1 and 2, Tables 1-4, and Tables A1-A4. Her supervisor, Geraldine Rauch, was steadily supporting in advisory capacity. By regular discussions with the colleagues in Heidelberg (Maximilian Pilz and Meinhard Kieser), the ideas and concepts could be fine-tuned steadily during the whole process.

***Thesis Article 2*: Carolin Herrmann, Geraldine Rauch (2021). "Smoothing corrections for improving sample size recalculation rules in adaptive group sequential study designs." *Methods of Information in Medicine*, 60, 1-8. doi: 10.1055/s-0040-1721727.**

Carolin Herrmann developed and tested the different smoothing approaches and wrote approximately 90% of the manuscript independently (i.e., the introduction, methods, results and discussion section). She conducted the literature search, developed the new methodology, conducted the related simulation studies and analyzed the respective results. Moreover, she generated all figures and tables within the manuscript, i.e., Figure 1, Table 1, and Supplementary Tables S1-S4 in the online supplementary material (https://www.thieme-connect.de/media/10.1055-s-00035037/EFirst/supmat/10-1055-s-0040-1721727-s20050009.pdf). Her supervisor, Geraldine Rauch, was steadily supporting in advisory capacity.

***Thesis Article 3*: Kevin Kunzmann, Maximilian Pilz, Carolin Herrmann, Geraldine Rauch, Meinhard Kieser (2021). "The adoptr package: adaptive optimal designs for clinical trials in R." *Journal of Statistical Software*. 98, 1-21. doi: 10.18637/jss.v098.i09.**

Carolin Herrmann was essentially involved in the extensive documentation of the R-package `adoptr`, which is available online for supporting the use of the package. Furthermore, she discussed the vignettes, examined the first draft of the article and was closely involved in the revision process of the article. Particularly, she suggested a visualization of the package's most important classes and methods, cf. Figure 1 in Kunzmann et al. (2021). She approved the final version of the article for publication and together with the other authors she bears the responsibility for the content.

Date, signature and stamp of first supervising university professor

Date and signature of doctoral candidate

Reprints of articles

In the electronic version of my work, DOI-routed reference links are included instead of my three thesis articles.

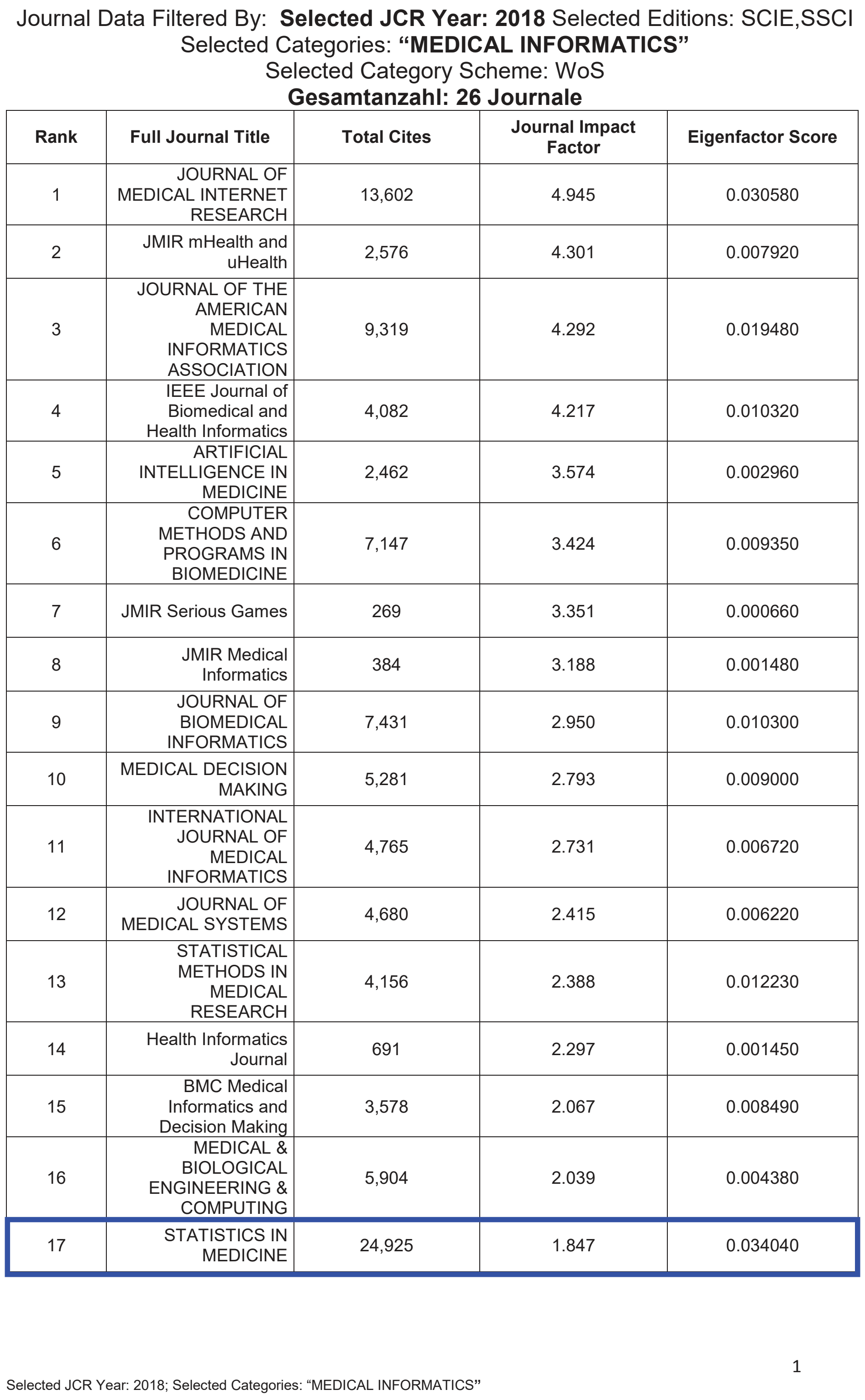
Thesis Article 1:
C. Herrmann, M. Pilz, M. Kieser, G. Rauch (2020). “A new conditional performance score for the evaluation of adaptive group sequential designs with sample size recalculation.” *Statistics in Medicine*, **39**, 2067–2100. doi: 10.1002/sim.8534.
Impact factor: 1.847

Thesis Article 2:
C. Herrmann, G. Rauch (2021). “Smoothing corrections for improving sample size recalculation rules in adaptive group sequential study designs.” *Methods of Information in Medicine*, *60*, 1-8. doi: 10.1055/s-0040-1721727.
Note that *Thesis Article 2* also consists of online supplementary material which can be accessed at `https://www.thieme-connect.de/media/10.1055-s-00035037/EFirst/supmat/10-1055-s-0040-1721727-s20050009.pdf`. This supplementary PDF-file is attached in here as well right after the main text of *Thesis Article 2*.
Impact factor: 1.574

Thesis Article 3:
K. Kunzmann, M. Pilz, **C. Herrmann**, G. Rauch, M. Kieser (2021). “The adoptr package: adaptive optimal designs for clinical trials in R.” *Journal of Statistical Software*, **98**, 1-21. doi: 10.18637/jss.v098.i09.
Impact factor: 13.642

Thesis Article 1

https://doi.org/10.1002/sim.8534



Journal Data Filtered By: **Selected JCR Year: 2018** Selected Editions: SCIE,SSCI
Selected Categories: **“MEDICAL INFORMATICS”**
Selected Category Scheme: WoS
Gesamtanzahl: 26 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	JOURNAL OF MEDICAL INTERNET RESEARCH	13,602	4.945	0.030580
2	JMIR mHealth and uHealth	2,576	4.301	0.007920
3	JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	9,319	4.292	0.019480
4	IEEE Journal of Biomedical and Health Informatics	4,082	4.217	0.010320
5	ARTIFICIAL INTELLIGENCE IN MEDICINE	2,462	3.574	0.002960
6	COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE	7,147	3.424	0.009350
7	JMIR Serious Games	269	3.351	0.000660
8	JMIR Medical Informatics	384	3.188	0.001480
9	JOURNAL OF BIOMEDICAL INFORMATICS	7,431	2.950	0.010300
10	MEDICAL DECISION MAKING	5,281	2.793	0.009000
11	INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS	4,765	2.731	0.006720
12	JOURNAL OF MEDICAL SYSTEMS	4,680	2.415	0.006220
13	STATISTICAL METHODS IN MEDICAL RESEARCH	4,156	2.388	0.012230
14	Health Informatics Journal	691	2.297	0.001450
15	BMC Medical Informatics and Decision Making	3,578	2.067	0.008490
16	MEDICAL & BIOLOGICAL ENGINEERING & COMPUTING	5,904	2.039	0.004380
17	STATISTICS IN MEDICINE	24,925	1.847	0.034040

1

Selected JCR Year: 2018; Selected Categories: “MEDICAL INFORMATICS”

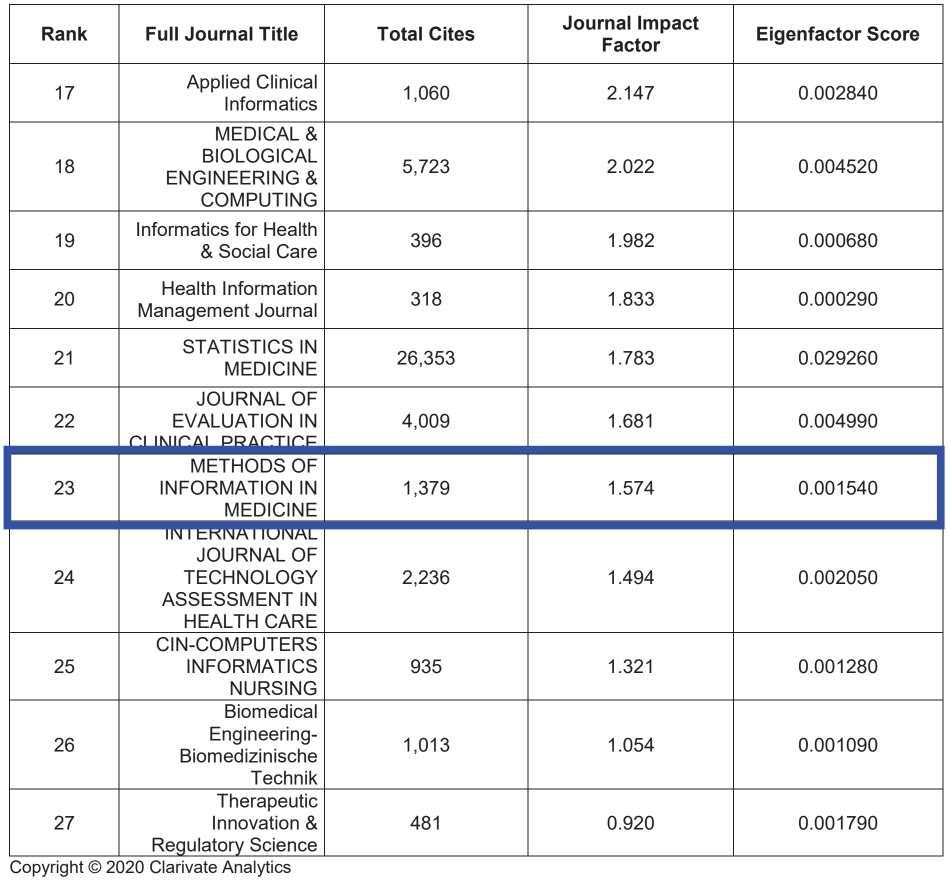
Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
18	Health Information Management Journal	320	1.742	0.000390
19	JOURNAL OF EVALUATION IN CLINICAL PRACTICE	4,039	1.536	0.005120
20	INTERNATIONAL JOURNAL OF TECHNOLOGY ASSESSMENT IN HEALTH CARE	2,143	1.418	0.002140
21	Applied Clinical Informatics	664	1.306	0.002050
22	Informatics for Health & Social Care	285	1.218	0.000470
23	CIN-COMPUTERS INFORMATICS NURSING	836	1.029	0.001120
24	METHODS OF INFORMATION IN MEDICINE	1,330	1.024	0.001760
25	Biomedical Engineering-Biomedizinische Technik	1,007	1.007	0.001320
26	Therapeutic Innovation & Regulatory Science	371	0.901	0.001600

Thesis Article 2

https://doi.org/10.1055/s-0040-1721727

Journal Data Filtered By: **Selected JCR Year: 2019** Selected Editions: SCIE,SSCI
Selected Categories: **“MEDICAL INFORMATICS”**
Selected Category Scheme: WoS
Gesamtanzahl: 27 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	IEEE Journal of Biomedical and Health Informatics	5,472	5.223	0.012910
2	JOURNAL OF MEDICAL INTERNET RESEARCH	16,349	5.034	0.029410
3	ARTIFICIAL INTELLIGENCE IN MEDICINE	2,953	4.383	0.003370
4	JMIR mHealth and uHealth	4,226	4.313	0.010020
5	JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	9,959	4.112	0.017380
6	COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE	8,014	3.632	0.011370
7	JMIR Serious Games	350	3.526	0.000660
7	JOURNAL OF BIOMEDICAL INFORMATICS	8,253	3.526	0.011190
9	Internet Interventions-The Application of Information Technology in Mental and Behavioural Health	996	3.513	0.002720
10	JOURNAL OF MEDICAL SYSTEMS	5,695	3.058	0.007050
11	INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS	5,368	3.025	0.007110
12	Health Informatics Journal	981	2.932	0.001530
13	JMIR Medical Informatics	650	2.577	0.002340
14	BMC Medical Informatics and Decision Making	4,117	2.317	0.007370
15	MEDICAL DECISION MAKING	5,291	2.309	0.007670
16	STATISTICAL METHODS IN MEDICAL RESEARCH	4,647	2.291	0.011850



Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
17	Applied Clinical Informatics	1,060	2.147	0.002840
18	MEDICAL & BIOLOGICAL ENGINEERING & COMPUTING	5,723	2.022	0.004520
19	Informatics for Health & Social Care	396	1.982	0.000680
20	Health Information Management Journal	318	1.833	0.000290
21	STATISTICS IN MEDICINE	26,353	1.783	0.029260
22	JOURNAL OF EVALUATION IN CLINICAL PRACTICE	4,009	1.681	0.004990
23	METHODS OF INFORMATION IN MEDICINE	1,379	1.574	0.001540
24	INTERNATIONAL JOURNAL OF TECHNOLOGY ASSESSMENT IN HEALTH CARE	2,236	1.494	0.002050
25	CIN-COMPUTERS INFORMATICS NURSING	935	1.321	0.001280
26	Biomedical Engineering-Biomedizinische Technik	1,013	1.054	0.001090
27	Therapeutic Innovation & Regulatory Science	481	0.920	0.001790

Thesis Article 3

https://doi.org/10.18637/jss.v098.i09

The *Journal of Statistical Software* is not listed at https://intranet.charite.de/medbib/zugangsdaten_fuer_zeitschriften/.

In the list of journals “Statistics & Probability” at https://jcr.clarivate.com/, the *Journal of Statistical Software* takes rank 1 of 124 (https://jcr.clarivate.com/JCRJournalHomeAction.action?pg=JRNLHOME&year=2019&edition=SCIE&categories=XY&newApplicationFlag=Y, accessed 28/03/2021, 3:50 pm) with a **Journal Impact Factor of 13.642**.

Curriculum Vitæ

My curriculum vitæ does not appear in the electronic version of my work for reasons of data protection.

List of publications

Number of first authorships[1]: 4
Number of co-authorships[1]: 8

The following journal impact factors are based on `https://jcr.clarivate.com/` (accessed February 05, 2021, 6:45 pm).

1) K. Kunzmann[2], M. Pilz[2], **C. Herrmann**, G. Rauch, M. Kieser. "The adoptr Package: Adaptive Optimal Designs for Clinical Trials in R." *Journal of Statistical Software*. **98**, 1-21. doi: 10.18637/jss.v098.i09.
Impact factor: 13.642

2) **C. Herrmann**[2], C. Kluge[2], M. Pilz, M. Kieser, G. Rauch. "Improving sample size recalculation rules in adaptive clinical trials by resampling." *Pharmaceutical Statistics*. doi: 10.1002/pst.2122.
Impact factor: 1.374

3) **C. Herrmann**, G. Rauch. "Smoothing corrections for improving sample size recalculation rules in adaptive group sequential study designs." *Methods of Information in Medicine*. **60**, 1-8. doi: 10.1055/s-0040-1721727.
Impact factor: 1.574

4) M. Pilz, K. Kunzmann, **C. Herrmann**, G. Rauch, M. Kieser. "Optimal planning of adaptive two-stage designs." *Statistics in Medicine*. **40**, 3196-3213. doi: 10.1002/sim.8953.
Impact factor: 1.783

5) T. Maleitzke, M. Pumberger, U. A. Gerlach, **C. Herrmann**, A. Slagman, L. Scheutz Henriksen, F. von Mauchenheim, N. Hüttermann, A. N. Santos, F. Fleckenstein, G. Rauch, S. Märdian, C. Perka, U. Stöckle, M. Möckel, T. Lindner, T. Winkler (2021). "Impact of the COVID-19 Shutdown on Trauma-Patterns and Numbers in an Academic Level-I Trauma Center Emergency Department in Berlin, Germany." *PLoS One*. doi: 10.1371/journal.pone-.0246956.
Impact factor: 2.740

6) J. Kruppa, J. Rohmann, **C. Herrmann**, M. Sieg, K. Rubarth, S. Piper (2021). "What statistics instructors need to know about concept acquisition to make statistics stick." *Journal of University Teaching and Learning Practice*, **18**, 2. Available at: `https://ro.uow.edu.au/jutlp/vol18/iss2/02`
Impact factor: –

[1] Including book contributions
[2] Authors contributed equally

7) **C. Herrmann** (2021). “Methoden zur Abwechslung, Auflockerung und Aktivierung in der (Biometrie-)Lehre.” Published in: C. Herrmann, U. Berger, C. Weiß, I. Burkholder, G. Rauch, J. Kruppa (eds) Zeig mir Health Data Science!. Berlin, Heidelberg, Springer Spektrum, pp. 81–92. doi: 10.1007/978-3-662-62193-6_7.
Impact factor: –

8) X. Li, **C. Herrmann**, G. Rauch (2020). “Optimality constraints for futility stopping boundaries in group-sequential designs with a continuous endpoint.” *BMC Medical Research Methodology*, **20**, 274. doi: 10.1186/s12874-020-01141-5.
Impact factor: 3.031

9) G. Rauch, K. Neumann, U. Grittner, **C. Herrmann**, J. Kruppa (2020). “Medizinische Statistik für Dummies.” 1st ed, Weinheim, Wiley-VCH Verlag GmbH & Co.
Impact factor: –

10) **C. Herrmann**, M. Pilz, M. Kieser, G. Rauch (2020). “A new conditional performance score for the evaluation of adaptive group sequential designs with sample size recalculation.” *Statistics in Medicine*, **39**:2067–2100. doi: 10.1002/sim.8534.
Impact factor: 1.783

11) M. Pilz, K. Kunzmann, **C. Herrmann**, G. Rauch, M. Kieser (2019). “A variational approach to optimal two-stage designs.” *Statistics in Medicine*, **38**:4159-4171. doi: 10.1002/sim.8291.
Impact factor: 1.783

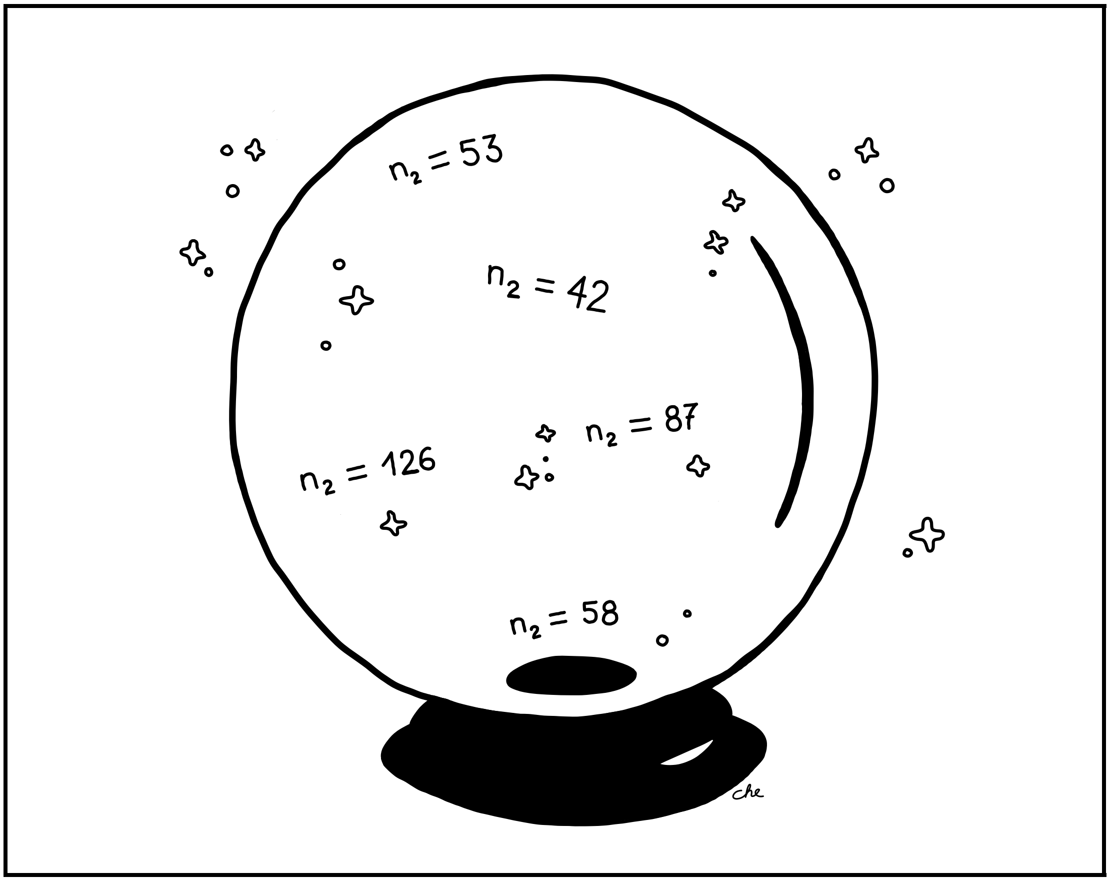
12) A. Müller, M. Olbert, A. Heymann, P. Zahn, K. Plaschke, V. von Dossow, D. Bitzinger, E. Barth, M. Meister, P. Kranke, **C. Herrmann**, K.-D. Wernecke, C. Spies (2019). “Relevance of peripheral cholinesterase activity on postoperative delirium in adult surgical patients (CESARO): A prospective observational cohort study.” *European Journal of Anaesthesiology*, **36**:114-122. doi: 10.1097/EJA.0000000000000888.
Impact factor: 4.140

Acknowledgements

I would particularly like to thank my excellent supervisor team, Prof. Dr. Geraldine Rauch, Prof. Dr. Meinhard Kieser und Dr. Jochen Kruppa. I am extremely thankful to Geraldine Rauch providing me with the very interesting topic of sample size recalculation in adaptive study designs as well as providing me with all necessary facilities for conducting the research. Geraldine Rauch and Meinhard Kieser were available anytime with their very profound knowledge on adaptive designs paired with considerable experience. Furthermore, my entire supervisor team provided very precious insights into the world of research. Geraldine Rauch and Jochen Kruppa shared their enthusiasm for teaching biostatistics, introduced me to research in the field of biostatistics education and supported me constantly with taking new steps.
Moreover, I would like to express my sincere gratitude to the German Research Foundation (grant RA 2347/4-1) for funding this work within the project ORACLE (Optimal Rules for Adaptive recalculation of sampLE size in clinical trials). Besides Geraldine Rauch and Meinhard Kieser, also Maximilian Pilz and Kevin Kunzmann were part of the ORACLE research. I am very grateful for extremely inspiring and productive discussions with the entire ORACLE team as well as to Corinna Kluge for a very stimulating collaboration.

I would like to extend my thanks to the entire PhD studies “Health Data Sciences”-team, especially to Prof. Dr. Dr. Tobias Kurth and Jessica Rohmann, as well as the other HDS-students for a great selection of offered courses and multiple inspiring discussions that expanded my interdisciplinary horizons. Moreover, I would like to thank all current and former colleagues at the Institute of Biometry and Clinical Epidemiology at the Charité – Universitätsmedizin Berlin for the very enjoyable work climate. I very much appreciated attending conferences, workshops and “Forschungscamps” with my colleagues, it was a great enrichment. I am very grateful to Annett Kaphahn and Dr. Dörte Huscher for their constant and very versatile support, and to Prof. Dr. Frank Konietschke and Dr. Uwe Schöneberg for their proficient advice at everyday work.

All this work would not be possible without the affectionate support of multiple people outside the world of work. Many happy and unforgettable moments with my friends in Germany and around the world, Rasmus as well as my family were a perfect source of motivation. Finally, I would like to express my deepest appreciation to my parents and my brother for their fond support during my studies.



ORACLE: “The second stage sample size is ...”

Drawn by Carolin Herrmann (2021).