**UNIVERSITÄTSKLINIKUM HAMBURG-EPPENDORF**

Institut für Medizinische Biometrie und Epidemiologie

Direktor: Prof. Dr. Heiko Becher

# Blinded sample size re-estimation in a confirmatory diagnostic accuracy study

**Dissertation**

zur Erlangung des Doktorgrades Dr. rer. biol. hum.
an der Medizinischen Fakultät der Universität Hamburg.

vorgelegt von:

Maria Stark
aus Augsburg

Hamburg 2022

**Angenommen von der**
**Medizinischen Fakultät der Universität Hamburg am:** 30.01.2023

**Veröffentlicht mit Genehmigung der**
**Medizinischen Fakultät der Universität Hamburg.**

**Prüfungsausschuss, der/die Vorsitzende:** Prof. Dr. Antonia Zapf

**Prüfungsausschuss, zweite/r Gutachter/in:** Prof. Dr. Levente Kriston

# Preface

I submitted my cumulative dissertation to the Medical Faculty at the University Medical Center Hamburg-Eppendorf within the Non-Medical PhD Programme.

I was supervised by Prof. Dr. Antonia Zapf (University Medical Center Hamburg-Eppendorf), Prof. Dr. Werner Brannath (University of Bremen) and Prof. Dr. Karl Wegscheider (University Medical Center Hamburg-Eppendorf).

My research was funded by the Deutsche Forschungsgemeinschaft (DFG) as part of the project *Flexible designs for diagnostic studies* (ZA 687/1-1).

My dissertation comprises three Thesis Articles which I summarize in the synopsis:

*Thesis Article 1:* Zapf, A., **Stark, M.**, Gerke, O., Ehret, C., Benda, N., Bossuyt, P., Deeks, J., Reitsma, J., Alonzo, T., & Friede, T. (2020). Adaptive trial designs in diagnostic accuracy research. *Statistics in Medicine, 39*(5), 591-601. https://doi.org/10.1002/sim.8430

*Thesis Article 2:* **Stark, M.**, & Zapf, A. (2020). Sample size calculation and re-estimation based on the prevalence in a single-arm confirmatory diagnostic accuracy study. *Statistical Methods in Medical Research*, *29*(10), 2958-2971. https://doi.org/10.1177/0962280220913588

*Thesis Article 3:* **Stark, M.**, Hesse, M., Brannath, W., & Zapf, A. (2022). Blinded sample size re-estimation in a comparative diagnostic accuracy study. *BMC Medical Research Methodology*, *22*, Article 115. https://doi.org/10.1186/s12874-022-01564-2

In the following synopsis, the personal pronoun ‘we’ refers to the group of researchers with whom I co-published these three Thesis Articles.

# Table of contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| CHMP | Committee for Medicinal Products for Human Use |
| CT | Computed Tomography |
| DFG | Deutsche Forschungsgemeinschaft |
| EMA | European Medicines Agency |
| EU | European Union |
| FDA | U.S. Food and Drug Administration |
| MCSE | Monte Carlo standard error |
| PET | Positron Emission Tomography |
| RMSE | Root mean squared error |

# List of Symbols

### General statistical symbols:

| | |
|---|---|
| $H_{0_{\text{global}}}$ | Global null hypothesis of the Intersection-Union test |
| $\mathrm{H}_{0_{\mathrm{Se}}}$ | Individual null hypothesis of the sensitivity |
| $\mathrm{H}_{0_{\mathrm{Sp}}}$ | Individual null hypothesis of the specificity |
| $\alpha_{\text{global}}$ | Global two-sided significance level of the Intersection-Union-test |
| $\alpha$ | Individual two-sided significance level per endpoint |
| $\mathrm{Power}_{\text{overall}}$ | Overall Power of the Intersection-Union test |
| $\mathrm{Power}_{\mathrm{Se}}$ | Individual power of the endpoint considering sensitivity |
| $\mathrm{Power}_{\mathrm{Sp}}$ | Individual power of the endpoint considering specificity |
| $\beta_{\text{overall}}$ | Overall type II error rate of the Intersection-Union test |
| $\beta_{\mathrm{Se}}$ | Individual type II error rate of the endpoint considering sensitivity |
| $\beta_{\mathrm{Sp}}$ | Individual type II error rate of the endpoint considering specificity |
| $Bin(k, p)$ | Binomial distribution with $k$ trials and success probability $p$ |
| $MVBin(k_{\mathrm{E}}, k_{\mathrm{C}}, p_{\mathrm{E}}, p_{\mathrm{C}}, \rho)$ | Multivariate binomial distribution with $k_{\mathrm{E}}$ and $k_{\mathrm{C}}$ trials in the experimental and comparator group, success probabilities $p_{\mathrm{E}}$ and $p_{\mathrm{C}}$ in the experimental and comparator group and the dependence between both tests $\rho$ |
| $z_{\cdot}$ | $(\cdot)$-quantile of standard normal distribution |
| $E(\cdot)$ | Expected value |
| $\Delta$ | Non-inferiority margin of the relevant endpoint |

### Diagnostic test:

| | |
|---|---|
| $\mathrm{Se}_{\mathrm{min}}$ | Minimum sensitivity |
| $\mathrm{Se}_{\mathrm{C}}$ | Sensitivity of the comparator test |
| $\mathrm{Se}_{\mathrm{E}}$ | Sensitivity of the experimental test |
| $\mathrm{Sp}_{\mathrm{min}}$ | Minimum specificity |
| $\mathrm{Sp}_{\mathrm{C}}$ | Specificity of the comparator test |
| $\mathrm{Sp}_{\mathrm{E}}$ | Specificity of the experimental test |
| $\delta_{\mathrm{Se}}$ | Difference between the experimental and comparator test regarding their sensitivities |

| | |
|---|---|
| $\delta_{\mathrm{Sp}}$ | Difference between the experimental and comparator test regarding their specificities |
| $\pi$ | Disease prevalence |
| $\pi_{\mathrm{ass.}}$ | Assumed disease prevalence |
| $\pi_{\mathrm{true}}$ | True disease prevalence |
| $\hat{\pi}$ | Estimated disease prevalence in the interim analysis |

**Single-test and unpaired study design:**

| | |
|---|---|
| $n_{\mathrm{D_E}}$ | Number of diseased individuals in the experimental group in the unpaired design |
| $n_{\mathrm{D_C}}$ | Number of diseased individuals in the comparator group in the unpaired design |
| $n_{\mathrm{ND_E}}$ | Number of non-diseased individuals in the experimental group in the unpaired design |
| $n_{\mathrm{ND_C}}$ | Number of non-diseased individuals in the comparator group in the unpaired design |
| $\mathrm{TP_C}$ | True positive results of the comparator test |
| $\mathrm{TN_C}$ | True negative results of the comparator test |
| $\mathrm{FP_C}$ | False positive results of the comparator test |
| $\mathrm{FN_C}$ | False negative results of the comparator test |
| $\mathrm{TP_E}$ | True positive results of the experimental test |
| $\mathrm{TN_E}$ | True negative results of the experimental test |
| $\mathrm{FP_E}$ | False positive results of the experimental test |
| $\mathrm{FN_E}$ | False negative results of the experimental test |

**Paired study design:**

| | |
|---|---|
| $n_{\mathrm{D}_{11}}$ | True-positive-positive results of the comparator and experimental test in the diseased population |
| $n_{\mathrm{D}_{00}}$ | False-negative-negative results of the comparator and experimental test in the diseased population |
| $n_{\mathrm{D}_{10}}$ | True-positive results of the experimental test and false-negative results of the comparator test in the diseased population |
| $n_{\mathrm{D}_{01}}$ | False-negative results of the experimental test and true-positive results of the comparator test in the diseased population |

| | |
|---|---|
| $n_{\mathrm{ND}_{11}}$ | False-positive-positive results of the comparator and experimental test in the non-diseased population |
| $n_{\mathrm{ND}_{00}}$ | True-negative-negative results of the comparator and experimental test in the non-diseased population |
| $n_{\mathrm{ND}_{10}}$ | False-positive results of the experimental test and true-negative results of the comparator test in the non-diseased population |
| $n_{\mathrm{ND}_{01}}$ | True-negative results of the experimental test and false-positive results of the comparator test in the non-diseased population |
| $\psi_{\mathrm{D}}$ | Proportion of discordant test results in the diseased population |
| $\psi_{\mathrm{ND}}$ | Proportion of discordant test results in the non-diseased population |
| $\psi_{\mathrm{D}_{\mathrm{min}}}$ | Minimal proportion of discordant rest results in the diseased population |
| $\psi_{\mathrm{ND}_{\mathrm{min}}}$ | Minimal proportion of discordant test results in the non-diseased population |
| $\hat{\psi}_{\mathrm{D}}$ | Estimated proportion of discordant test results in the diseased population in the interim analysis |
| $\hat{\psi}_{\mathrm{ND}}$ | Estimated proportion of discordant test results in the non-diseased population in the interim analysis |
| $\mathrm{TPPR}$ | True-positive-positive-rate |
| $\mathrm{TNNR}$ | True negative-negative-rate |
| $\mathrm{TPPR}_{\mathrm{max}}$ | Maximal true-positive-positive-rate |
| $\mathrm{TNNR}_{\mathrm{max}}$ | Maximal true-negative-negative-rate |
| $\widehat{TPPR}$ | Estimated true-positive-positive-rate in the interim analysis |
| $\widehat{TNNR}$ | Estimated true-negative-negative-rate in the interim analysis |

**Sample size (re-)estimation:**

| | |
|---|---|
| $n_{\mathrm{D}}$ | Number of diseased individuals |
| $n_{\mathrm{ND}}$ | Number of non-diseased individuals |
| $N_{\mathrm{Se}}$ | Sample size needed to show sufficient sensitivity |
| $N_{\mathrm{Sp}}$ | Sample size needed to show sufficient specificity |
| $N$ | Initial total sample size |
| $n$ | Sample size used for the interim analysis |
| $\lambda$ | Size of the internal pilot study relative to the initial sample size |
| $\widehat{N}$ | Re-estimated sample size |
| $N_{\mathrm{true}}$ | True sample size based on true nuisance parameters |
| $R$ | Quotient of the re-estimated sample size and the true sample size |

**Simulation study:**

| | |
|---|---|
| $n_{\mathrm{sim}}$ | Simulation runs per scenario |
| $\mathrm{MCSE}_{\text{single-test}}$ | Monte Carlo standard error in the single-test design |
| $\mathrm{MCSE}_{\mathrm{comparative}}$ | Monte Carlo standard error in a comparative design |
| $\hat{\pi}_{\mathrm{mean}}$ | Mean of estimated prevalence in interim analysis of all simulation runs per scenario |
| $\hat{\psi}_{\mathrm{D}_{\mathrm{mean}}}$ | Mean of estimated proportions of discordant test results in the diseased population in interim analysis of all simulation runs per scenario |
| $\hat{\psi}_{\mathrm{ND}_{\mathrm{mean}}}$ | Mean of estimated proportions of discordant test results in the non-diseased population in interim analysis of all simulation runs per scenario |

# Abstract

**Background:** A confirmatory phase III diagnostic study compares the sensitivity and specificity of an experimental test either against pre-defined minimum thresholds in a single-test design or against a comparator test in an unpaired or paired design. Hereby, it combines sensitivity and specificity as co-primary endpoints. In conventional sample size calculation, the problem of an overpowered study can arise because the final sample size represents the maximum of the individual sample sizes needed to show sufficient sensitivity or specificity. For this sample size calculation, one needs assumptions about nuisance parameters which are the disease prevalence and, in a paired design, proportions of discordant test results in the diseased and non-diseased population. Often these assumptions are uncertain leading to an incorrect specification of the sample size. This thesis aims to improve the initial sample size calculation and to develop blinded adaptive designs for the sample size re-estimation.

**Methods:** As there have been no adaptive designs to adjust the sample size during a diagnostic study so far, my co-authors and I elaborate on possibilities for developing adaptive designs in *Thesis Article 1*. In *Thesis Articles 2 and 3*, we develop the optimal sample size calculation and blinded adaptive design based on the optimal sample size calculation. We use the adaptive design to adjust the sample size by estimating nuisance parameters in each of the three study designs. We aim to show superiority or non-inferiority of the experimental test in both endpoints or combinations of superiority and non-inferiority, respectively. We conduct simulation studies to evaluate the performance of the blinded adaptive design and compare it to a fixed design without sample size re-estimation. Furthermore, we compare the blinded adaptive design to the existing approach of McCray et al. (2017) within a paired example study.

**Results:** The optimal sample size calculation and blinded adaptive design support reaching the target power. The proposed adaptive design controls the type I error rate due to blinded sample size re-estimation. Nuisance parameters are estimated without any relevant bias. Adjusted sample sizes are close to true sample sizes. Our blinded adaptive design leads to a smaller sample size than the already existing approach.

**Conclusions:** We suggest applying the optimal sample size calculation and the blinded adaptive design in confirmatory diagnostic accuracy studies as both support reaching the target power. Their application does not require much additional effort.

**Zusammenfassung**

**Hintergrund:** Eine konfirmatorische Diagnosestudie vergleicht die Sensitivität und Spezifität eines experimentellen Tests mit vordefinierten Grenzen in einem Ein-Test Design bzw. mit einem Komparatortest in einem ungepaarten oder gepaarten Design. Die Studie kombiniert Sensitivität und Spezifität zu co-primären Endpunkten. In der konventionellen Fallzahlplanung kann eine zu hohe Power auftreten, weil die finale Fallzahl das Maximum der individuellen Fallzahlen der beiden Endpunkte darstellt. Für die Fallzahlplanung braucht man Annahmen über Störparameter wie die Prävalenz, und im gepaarten Design, die Anteile der diskordanten Testergebnisse. Oftmals sind diese Annahmen unsicher, was zu einer ungenauen Berechnung der Fallzahl führen kann. Das Ziel meiner Doktorarbeit ist die Verbesserung der initialen Fallzahlplanung und die Entwicklung eines verblindeten adaptiven Designs zur Fallzahlneuschätzung.

**Methoden:** Da bisher keine adaptiven Designs zur Fallzahlanpassung in einer Diagnosestudie existiert haben, erörtern meine Koautoren und ich die Möglichkeiten zur Entwicklung von adaptiven Designs in *Publikation 1*. In *Publikation 2 und 3* entwickeln wir die optimale Fallzahlplanung und ein verblindetes adaptive Design, das auf die optimale Fallzahlplanung zurückgreift. Es dient zur Anpassung der Fallzahl mit Hilfe der geschätzten Störparameter in den drei Studiendesigns, die entweder darauf abzielen die Überlegenheit, Nicht-Unterlegenheit oder eine Kombination aus Überlegenheit und Nicht-Unterlegenheit des experimentellen Tests in beiden Endpunkten zu zeigen. Wir verwenden Simulationsstudien um das Verhalten des verblindeten adaptiven Designs zu evaluieren und es mit einem fixen Design ohne Fallzahlneuschätzung zu vergleichen. Außerdem vergleichen wir an einer gepaarten Beispielstudie das verblindete adaptive Design mit dem existierenden Ansatz von McCray et al. (2017).

**Ergebnisse:** Die optimale Fallzahlplanung und das adaptive Design tragen zum Erreichen der Zielpower bei. Durch die Verblindung kontrolliert das adaptive Design den Fehler 1. Art. Die Störparameter werden ohne relevante Verzerrung geschätzt. Die angepassten Fallzahlen kommen den wahren Fallzahlen nahe. Unser adaptives Design führt zu kleineren Fallzahlen als der bereits existierende Ansatz.

**Schlussfolgerungen:** Wir schlagen die Anwendung der optimalen Fallzahlplanung und des verblindeten adaptiven Designs vor, weil beide Methoden das Erreichen der Zielpower unterstützen. Ihre Anwendung erfordert keinen erheblichen Zusatzaufwand.

# 1 Synopsis

## 1.1 Introduction

On 26 May 2021, after a four-year transitional period, the European Union (EU) Regulation 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices came into force . The regulation aims to create uniform rules for developing and approving medical devices in the European Union (EU Regulation of 5 April 2017, 2017). According to the regulation, a medical device can be used, among other things, to correctly diagnose a disease or injury (EU Regulation of 5 April 2017, 2017). Hence, it serves as a diagnostic test by revealing a patient's disease status to support attending physicians in their choice of necessary treatments (Zhou et al., 2011). During development, a medical device should pass some consecutive phases including exploratory, confirmatory and post-market clinical follow-up studies (EU Regulation of 5 April 2017, 2017). In 1990, Köbberling et al. structured these studies into four phases: Phase I studies serve to verify the functionality and safety of the diagnostic test. Then, exploratory phase II studies give first insights into the diagnostic accuracy in patients with known disease status and help to identify the cut-off value to distinguish between diseased and non-diseased ones. Confirmatory phase III studies aim to prove the diagnostic accuracy with the pre-defined cut-off value in patients with a priori unknown disease status. Finally, phase IV studies evaluate patients' benefits of the diagnostic test after the market launch in combination with subsequent medical treatments.

In confirmatory phase III studies, several study designs evaluate the performance of the experimental diagnostic test. In each case, the reference standard determines the true disease status of each study participant. Figure 1 shows these study designs. The single-test design compares the experimental test to pre-defined minimum thresholds (Zhou et al., 2011). A comparative study design compares the experimental test to a comparator test in an unpaired or paired way (Zhou et al., 2011). The unpaired design randomizes each study participant to either the experimental or comparator test, in addition to the reference standard (Pepe, 2003). In the paired design, participants have all three diagnostic tests performed on them (Bossuyt et al., 2006). The *Guideline on clinical evaluation of diagnostic agents* issued by the European Medicines Agency (EMA) through its Committee for Medicinal Products for Human Use (CHMP) (2009) prefers the paired design to the unpaired design if it is practicable and ethically

justifiable. The reason is that the comparison of diagnostic tests within the same individual diminishes variability of measurements. Otherwise, in case of an invasive experimental and comparator test, the unpaired design might be more appropriate (Alonzo et al., 2002).

EMA and U.S. Food and Drug Administration (FDA) recommend to use sensitivity and specificity as co-primary endpoints in diagnostic accuracy phase III studies (CHMP, 2009; FDA, 2007). Sensitivity determines the probability to correctly diagnose diseased individuals (Pepe, 2003). Whereas specificity denotes the probability to correctly diagnose non-diseased individuals. The Intersection-Union Test combines sensitivity and specificity with a joint hypothesis (Hamasaki et al., 2018; Korevaar et al., 2019). In this context, the study goal is either to show superiority or non-inferiority of the experimental test in both endpoints. A combination of superiority and non-inferiority regarding both endpoints is also possible.

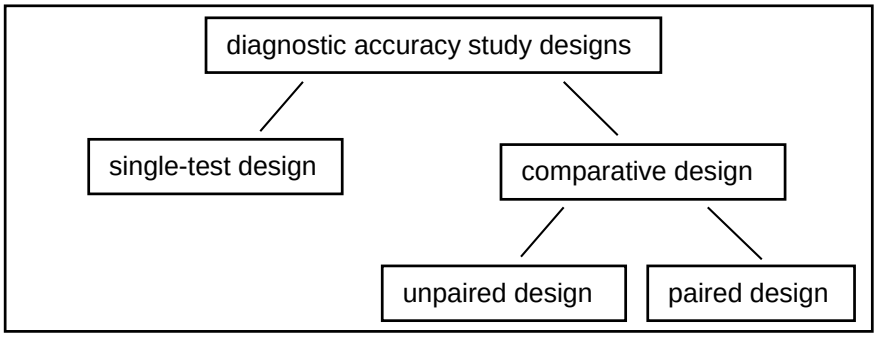



**Figure 1.** Study designs of a confirmatory diagnostic accuracy trial (*Thesis Article 3*: Stark et al., 2022).

EU regulation 2017/745 (EU Regulation of 5 April 2017, 2017) requires justification of the sample size needed to evaluate a medical device. Hence, the sponsor of a study used for approving the medical device has to provide a precise sample size calculation for such a study. Referring again to confirmatory diagnostic accuracy phase III studies, the *Guideline on clinical evaluation of diagnostic agents* (CHMP, 2009) points out that the sample size strongly depends on the disease prevalence. The conventional sample size calculation consists of three steps combining the co-primary endpoints in reference to the prevalence. The first step determines the needed sample sizes of diseased and non-diseased individuals by performing separate calculations for each endpoint. The second step relates these sample sizes to the prevalence to receive representative samples to show sufficient sensitivity or specificity, respectively (Flahault et al., 2005; Hajian-Tilaki, 2014). The two representative samples sizes are not necessarily equal, especially if there is a low or high disease prevalence. In the

third step, the maximum of both representative sample sizes leads to the final sample size (Buderer, 1996). In *Thesis Article 2*, we explain the conventional approach by planning needed sample sizes of diseased and non-diseased participants with an individual power of 90% each in the first step (Stark & Zapf, 2020). As both groups of diseased and non-diseased individuals are independent, the study reaches an overall power of at least 80%. However, we highlight in *Thesis Articles 2 and 3* that this conventional approach can lead to a too large final sample size and hence to an overpowered study (Stark et al., 2022; Stark & Zapf, 2020). This problem arises from choosing the maximum of both representative sample sizes in the third step. Such an overpowered study raises ethical concerns because as few patients as possible should be exposed to investigations in a trial to avoid unnecessary burdens. This thesis addresses this research question by providing an optimal sample size calculation approach depending on the prevalence to avoid overpowered confirmatory diagnostic accuracy studies. This approach covers the sample size calculation for single-test, unpaired and paired diagnostic studies aiming to show superiority, non-inferiority or a combination of both in the co-primary endpoints.

Often, assumptions needed for sample size calculation are uncertain during the planning phase of a confirmatory study (CHMP, 2007). Adaptive designs serve, among other aims, to counteract these uncertain assumptions by estimating predefined parameters included in the initial sample size calculation and, then, adjusting the sample size during the study (Bhatt & Mehta, 2016). The *Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design* states that, in general, adaptive study designs use “statistical methodology [which] allows the modification of a design element … at an interim analysis with full control of the type I error” (CHMP, 2007). A subgroup of adaptive designs are group-sequential designs which enable an early stop of the trial due to futility or efficacy of the experimental diagnostic test (FDA, 2018; Zapf et al., 2020). Adaptive designs can be based on either blinded or unblinded interim analyses, with the unblinded design revealing the sensitivity and specificity of the experimental test in diagnostic studies (Zapf et al., 2020). In contrast, the blinded design keeps them secret but estimates nuisance parameters, e.g. the disease prevalence and proportions of concordant or discordant test results in the paired design (McCray et al., 2017; Proschan, 2005). In *Thesis Article 1*, we worked out that adaptive designs are already well-known and much appreciated in the context of therapeutic trials, but hardly developed for

diagnostic studies (Zapf et al., 2020). Consequently, we confirm the need for adaptive designs in diagnostic studies and elaborate on their potential uses. However, sample size re-estimation is particularly important in diagnostic studies because some parameters needed for sample size calculation might be unknown in most cases, e.g. proportions of discordant test results in a paired design (Gerke et al., 2012). Therefore, the implementation of adaptive designs in diagnostic studies is necessary. This thesis contributes to this extensive area of research by developing blinded adaptive designs to re-estimate the sample size based on nuisance parameters in confirmatory diagnostic accuracy studies.

To summarize the state of the literature on sample size calculation in diagnostic studies, Jones et al. (2003), Bachmann et al. (2006) as well as Bochmann et al. (2007) ascertain that hardly any medical publication on diagnostic studies reports an appropriate sample size calculation. However, there is statistical methodology to calculate sample sizes in diagnostic accuracy studies. Hajian-Tilaki (2014) provides a sample size calculation approach applicable if the disease status is known. For prospective diagnostic studies with unknown disease status of study participants, Buderer (1996) and Flahault et al. (2005) include the disease prevalence into sample size calculation to obtain a representative sample. No publication addresses the problem of an overpowered study that can arise in the context of co-primary endpoints. This gives evidence for developing the optimal sample size calculation in this thesis.

The literature contains well-established methodology for adaptive designs in therapeutic studies, including group-sequential designs (Bauer et al., 2016; Chow & Chang, 2006; CHMP, 2007; Jennison & Turnbull, 2000; Moyé, 2006; Todd, 2007; Wald, 2014; Wassmer & Brannath, 2016; Whitehead, 1997). Referring to blinded adaptive designs in therapeutic studies, there are several approaches to re-estimate the sample size based on nuisance parameters with either a continuous, time-to-event or binary endpoint (Friede et al., 2019; Friede & Kieser, 2006, 2013; Friede & Miller, 2012; Proschan, 2009; Sander et al., 2017). Asakura et al. (2017) propose a group-sequential design to early stop for futility or efficacy in therapeutic trials with co-primary endpoints measured within the same individual. However, sensitivity and specificity are independent co-primary endpoints in confirmatory diagnostic studies, which is why the approach of Asakura et al. (2017) cannot be applied to them. In *Thesis Article 1*, we collect research on adaptive designs for several development phases of a diagnostic

test, especially for phase II studies (Zapf et al., 2020). McCray et al. (2017) publish a blinded adaptive design to re-estimate the sample size based on the prevalence and the proportion of concordant test results in a paired diagnostic phase III study. However, they do not address the problem of an overpowered sample size which occurs due to co-primary endpoints. In general, the small amount of existing literature and conclusions based on *Thesis Article 1* confirm the need for research regarding blinded adaptive designs in diagnostic studies.

In summary of elaborations described above, the overall goal of my thesis is to develop methods supporting to reach the target power in confirmatory diagnostic accuracy studies. In detail, there are three aims:

1. Verification of the need for adaptive designs in diagnostic accuracy studies and elaboration on their potential uses in *Thesis Article 1* (Zapf et al., 2020)
2. Development of the optimal sample size calculation to avoid overpowered diagnostic accuracy studies in the single-test, unpaired and paired design either testing for superiority or non-inferiority in both endpoints or combinations of both in *Thesis Articles 2 and 3* (Stark et al., 2022; Stark & Zapf, 2020)
3. Development of blinded adaptive designs for diagnostic accuracy studies to adjust the sample size based on the estimation of nuisance parameters during the study in the single-test, unpaired and paired design either testing for superiority or non-inferiority in both endpoints or combinations of both in *Thesis Articles 2 and 3* (Stark et al., 2022; Stark & Zapf, 2020)

The synopsis is structured the following way: After this Introduction, in the Methods Section, I explain the conventional and optimal sample size calculation in a confirmatory diagnostic accuracy study theoretically and by means of an example study at first. Second, I introduce the blinded sample size re-estimation procedure and third, I describe simulation studies performed for evaluating the adaptive design. In the Results Section, I show results of the blinded sample size re-estimation procedure applied to the example study and results of the simulation study. In the Discussions Section, I highlight strengths and weaknesses of proposed methods, give an outlook to further research. Finally, I provide a conclusion.

## 1.2 Methods

### 1.2.1 Statistical Methods

#### 1.2.1.1 Sample size calculation in a confirmatory diagnostic study

The first major task of this thesis represents the development of an optimal sample size calculation to avoid overpowered confirmatory diagnostic accuracy studies. To start, this section provides general information about the choice of endpoints in a confirmatory diagnostic study and gives first insights into parameters needed for sample size calculation. Subsection 1.2.1.1.1 describes the conventional sample size calculation for chosen endpoints and works out disadvantages related to this approach. Subsection 1.2.1.1.2 explains the optimal sample size calculation to address these disadvantages.

Both the *Guideline on clinical evaluation of diagnostic agents* (CHMP, 2009) and the *Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests* (FDA, 2007) require sensitivity and specificity as co-primary endpoints in confirmatory diagnostic studies. The Intersection-Union test combines these co-primary endpoints to test one joint hypothesis (Stark & Zapf, 2020). This joint global null hypothesis of the Intersection-Union test ($H_{0_{\text{global}}}$) represents the union of the individual null hypothesis of the sensitivity ($\mathrm{H}_{0_{\mathrm{Se}}}$) and the specificity ($\mathrm{H}_{0_{\mathrm{Sp}}}$) (Hamasaki et al., 2018). Referring to a comparative design which aims to show superiority of the experimental test regarding its sensitivity ($\mathrm{Se}_{\mathrm{E}}$) and specificity ($\mathrm{Sp}_{\mathrm{E}}$) compared to the sensitivity ($\mathrm{Se}_{\mathrm{C}}$) and specificity ($\mathrm{Sp}_{\mathrm{C}}$) of the comparator test, we define $H_{0_{\text{global}}}$ for equality as follows:

$$\mathrm{H}_{0_{\mathrm{Se}}}: \mathrm{Se}_{\mathrm{E}} = \mathrm{Se}_{\mathrm{C}} \text{ and } \mathrm{H}_{0_{\mathrm{Sp}}}: \mathrm{Sp}_{\mathrm{E}} = \mathrm{Sp}_{\mathrm{C}}$$
$$\mathrm{H}_{0_{\text{global}}} = \mathrm{H}_{0_{\mathrm{Se}}} \cup \mathrm{H}_{0_{\mathrm{Sp}}} \quad (1)$$

We will only reject $H_{0_{\text{global}}}$ if we can reject $\mathrm{H}_{0_{\mathrm{Se}}}$ and $\mathrm{H}_{0_{\mathrm{Sp}}}$ at the same time (Hamasaki et al., 2018). The overall power of the Intersection-Union test equals the product of powers resulting from each individual hypothesis because sensitivity and specificity refer to the independent diseased and non-diseased population (Stark & Zapf, 2020). The same applies to the global type I error rate. We can see superiority of the experimental test compared to the comparator test from point estimates with p-values or confidence intervals (Stark et al., 2022). Furthermore, a comparative study can aim to show non-inferiority in both endpoints and a combination of superiority and non-inferiority. *Thesis Article 3* defines $H_{0_{\text{global}}}$ in each of these settings (Stark et al., 2022).

In analogy to the comparative design, we can define $H_{0_{\text{global}}}$ for the comparison of the sensitivity and specificity of the experimental test to a predefined minimum sensitivity ($\text{Se}_{\text{min}}$) and minimum specificity ($\text{Sp}_{\text{min}}$) in the single-test design. In this design, the distinction between superiority and non-inferiority does not exist (Stark & Zapf, 2020).

**Table 1.** Parameters needed for sample size calculation in the respective confirmatory diagnostic study design.

| | Study design | | |
|---|---|---|---|
| **Parameters** | **Single-test** | **Unpaired** | **Paired** |
| **Parameters for diagnostic accuracy** | $\text{Se}_{\text{E}}, \text{Sp}_{\text{E}}$<br>$\text{Se}_{\text{min}}, \text{Sp}_{\text{min}}$ | $\text{Se}_{\text{E}}, \text{Sp}_{\text{E}}$<br>$\text{Se}_{\text{C}}, \text{Sp}_{\text{C}}$ | |
| **Nuisance parameters** | $\pi$ | $\pi$ | $\pi$<br>$\psi_{\text{D}}, \psi_{\text{ND}}$ |

Sample size calculation for the Intersection-Union test in the context of a confirmatory diagnostic study also requires assumptions about nuisance parameters, in addition to sensitivities and specificities already mentioned. Table 1 gives an overview of all needed parameters for sample size calculation. Nuisance parameters play an essential role in the blinded sample size re-estimation approach as they are included in the sample size calculation and can be estimated in a blinded way. Due to this importance, I will introduce them here in the context of the general sample size calculation to familiarize the reader with them. In a confirmatory diagnostic study, nuisance parameters are e.g. the disease prevalence ($\pi$) and, in the paired design, proportions of discordant test results in the diseased ($\psi_{\text{D}}$) and non-diseased population ($\psi_{\text{ND}}$) (Stark et al., 2022). The latter are those proportions in which the experimental and comparator test give different test results.

#### 1.2.1.1.1 Conventional sample size calculation

This section explains the conventional sample size calculation with the example of a paired study. The aim of this section is to highlight disadvantages of the conventional approach and to motivate the need for the optimal sample size calculation. In addition to this section, *Thesis Article 2* provides the conventional sample size calculation with a concrete example for the single-test design.

The example study for the paired design introduced by McCray et al. (2017) deals with the correct diagnosis of pancreatic cancer. It aims to show superiority regarding the sensitivity and specificity of the experimental combination of Positron Emission Tomography (PET) and Computed Tomography (CT) against the comparator test CT.

We perform the three steps of the conventional sample size calculation based on assumptions of the example study given in Table 2:

**Table 2.** Assumptions of the paired diagnostic accuracy trial for comparing the experimental Positron Emission Tomography (PET) combined with the computed tomography (CT) against the comparator test CT (McCray et al., 2017; *Thesis Article 3*: Stark et al., 2022).

| **General input parameters:**<br>**Significance level per endpoint: $\alpha = 0.05$ (two-sided)**,<br>**Overall Power: $\text{Power}_{\text{overall}} = 1 - \beta_{\text{overall}} = 0.8$**<br>**Power per endpoint: $\text{Power}_{\text{Se}} = \text{Power}_{\text{Sp}} = 1 - \beta_{\text{Se}} = 1 - \beta_{\text{Sp}} = 0.9$** | | | |
|---|---|---|---|
| Prevalence:<br>$\pi = 0.47$ | **Comparator test (CT)** | **Experimental Test (PET/CT)** | **Proportion of discordant test results** |
| **Diseased population** | $\text{Se}_\text{C} = 0.81$ | $\text{Se}_\text{E} = 0.90$ | $\psi_\text{D} = 0.09$ |
| **Non-diseased population** | $\text{Sp}_\text{C} = 0.66$ | $\text{Sp}_\text{E} = 0.80$ | $\psi_\text{ND} = 0.14$ |

1. Sample size of diseased individuals based on the formula of Miettinen et al. (1968):

$$n_\text{D} = \frac{\left(z_{1-\alpha/2} \cdot \psi_\text{D} + z_{1-\beta_\text{Se}}\sqrt{\psi_\text{D}^2 - \frac{1}{4}(\text{Se}_\text{C} - \text{Se}_\text{E})^2(3 + \psi_\text{D})}\right)^2}{\psi_D(\text{Se}_\text{C} - \text{Se}_\text{E})^2} = 74$$

   Sample size of non-diseased individuals:

$$n_\text{ND} = \frac{\left(z_{1-\alpha/2} \cdot \psi_\text{ND} + z_{1-\beta_\text{Sp}}\sqrt{\psi_\text{ND}^2 - \frac{1}{4}(\text{Sp}_\text{C} - \text{Sp}_\text{E})^2(3 + \psi_\text{ND})}\right)^2}{\psi_\text{ND}(\text{Sp}_\text{C} - \text{Sp}_\text{E})^2} = 47$$

2. Sample size needed to show the sensitivity including at least $n_\text{D}$ diseased individuals:

$$N_\text{Se} = \frac{n_\text{D}}{\pi} = \frac{74}{0.47} = 157$$

   Sample size needed to show the specificity including at least $n_\text{ND}$ non-diseased individuals:

$$N_\text{Sp} = \frac{n_\text{ND}}{1 - \pi} = \frac{47}{1 - 0.47} = 88$$

3. $N = \max\left(N_\text{Se}, N_\text{Sp}\right) = 157$

The disadvantage of the conventional sample size calculation is that it can lead to an overpowered diagnostic study because the final sample size $N$ could be higher than necessary. This problem arises if the sample size needed to show the sensitivity ($N_\text{Se}$) and the sample size needed to show the specificity ($N_\text{Sp}$) are unequal. Choosing the maximum of $N_\text{Se}$ and $N_\text{Sp}$ in the third step leads to recruiting more study participants

than necessary to show the endpoint with the smaller sample size. In this example, $N_{\mathrm{Se}}$ is higher than $N_{\mathrm{Sp}}$ leading to an overpowered endpoint of the specificity. Figure 2 compares simulation results of the conventional and optimal sample size calculation regarding the empirical power and sample size of the example study if we vary the prevalence, based on 10,000 simulation runs (Stark et al., 2022). In this subsection, the focus is on the grey dashed line representing the conventional sample size calculation.

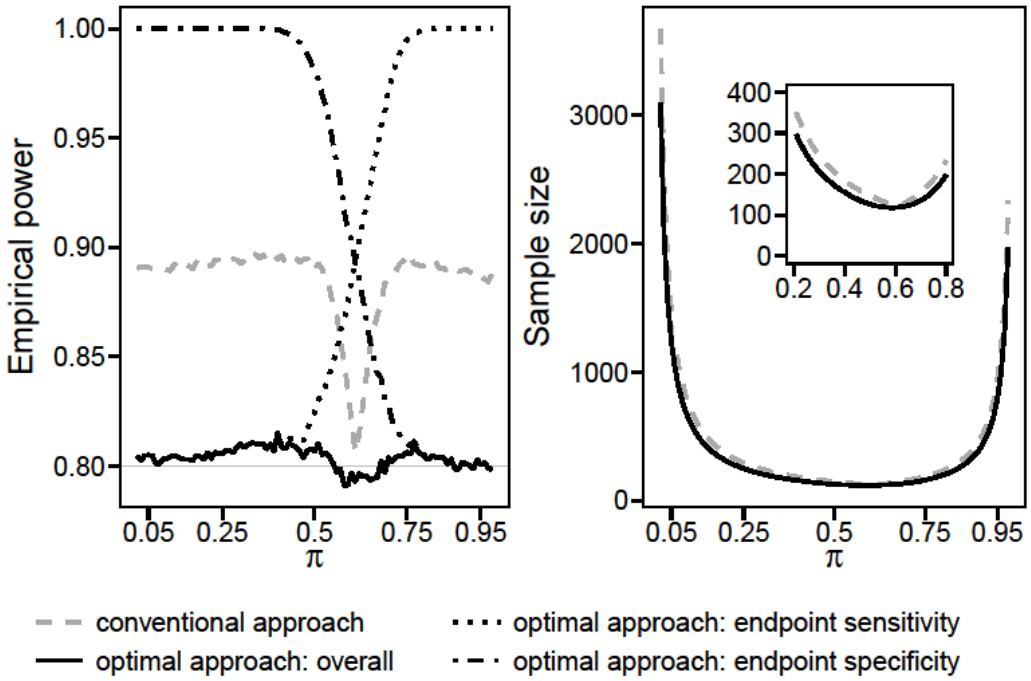



**Figure 2.** Comparison of the conventional and optimal sample size calculation for the paired example study with respect to a varying prevalence (π) regarding the resulting empirical overall power or sample size (*Thesis Article 3*: Stark et al., 2022).

By choosing the maximum of $N_{\mathrm{Se}}$ and $N_{\mathrm{Sp}}$ in the third step of the conventional approach, the final sample size is higher than needed to show the specificity. Differences between $N_{\mathrm{Se}}$ and $N_{\mathrm{Sp}}$ would be even larger if the disease prevalence was unbalanced leading to either an overpowered endpoint of the sensitivity or specificity depending on the direction of unbalance. These discrepancies between the sample sizes of both endpoints can result in an overpowered study. To address this disadvantage, we introduce the optimal sample size calculation in the following section.

#### 1.2.1.1.2 Optimal sample size calculation

We developed the optimal sample size calculation to address the problem of an overpowered confirmatory diagnostic study. This section presents the general idea of the optimal sample size calculation. *Thesis Articles 2 and 3* provide formulas for each study design aiming to show superiority or non-inferiority in both endpoints as well as the combination of both (Stark et al., 2022; Stark & Zapf, 2020).

The optimal sample size calculation splits the overall power ($\mathrm{Power}_{\mathrm{overall}}$) to both endpoints, so that both $N_{\mathrm{Se}}$ and $N_{\mathrm{Sp}}$ are equal and the product of the individual powers to show the sensitivity ($\mathrm{Power}_{\mathrm{Se}}$) and specificity ($\mathrm{Power}_{\mathrm{Sp}}$) results in the desired target overall power. Consequently, the final sample size is the smallest representative sample. The choice of the maximum of $N_{\mathrm{Se}}$ and $N_{\mathrm{Sp}}$ is not necessary if:

$$N_{\mathrm{Se}} \overset{!}{=} N_{\mathrm{Sp}} \tag{2}$$

$$\frac{n_{\mathrm{D}}}{\pi} \overset{!}{=} \frac{n_{\mathrm{ND}}}{1-\pi} \tag{3}$$

Under the condition:

$$\mathrm{Power}_{\mathrm{Se}} \cdot \mathrm{Power}_{\mathrm{Sp}} = \mathrm{Power}_{\mathrm{overall}} \tag{4}$$

$$(1-\beta_{\mathrm{Se}}) \cdot \left(1-\beta_{\mathrm{Sp}}\right) = \mathrm{Power}_{\mathrm{overall}} \tag{5}$$

$$\beta_{\mathrm{Sp}} = \frac{1-\beta_{\mathrm{Se}} - \mathrm{Power}_{\mathrm{overall}}}{1-\beta_{\mathrm{Se}}} = 1 - \frac{\mathrm{Power}_{\mathrm{overall}}}{1-\beta_{\mathrm{Se}}} \tag{6}$$

Figure 2 shows results of the optimal approach in black lines. The empirical overall power is the product of empirical powers of sensitivity and specificity. By individually splitting the overall power to both endpoints, the optimal sample size is lower than the conventional one especially for a low and high prevalence. As a result, empirical overall power of the optimal approach comes close to the target power.

### 1.2.1.2 Blinded sample size re-estimation

Sample size re-estimation verifies initial assumptions of sample size calculation and adjusts the sample size if necessary (CHMP, 2007). The second major task of this thesis is to implement a blinded sample size re-estimation procedure for confirmatory diagnostic studies based on the estimation of nuisance parameters.

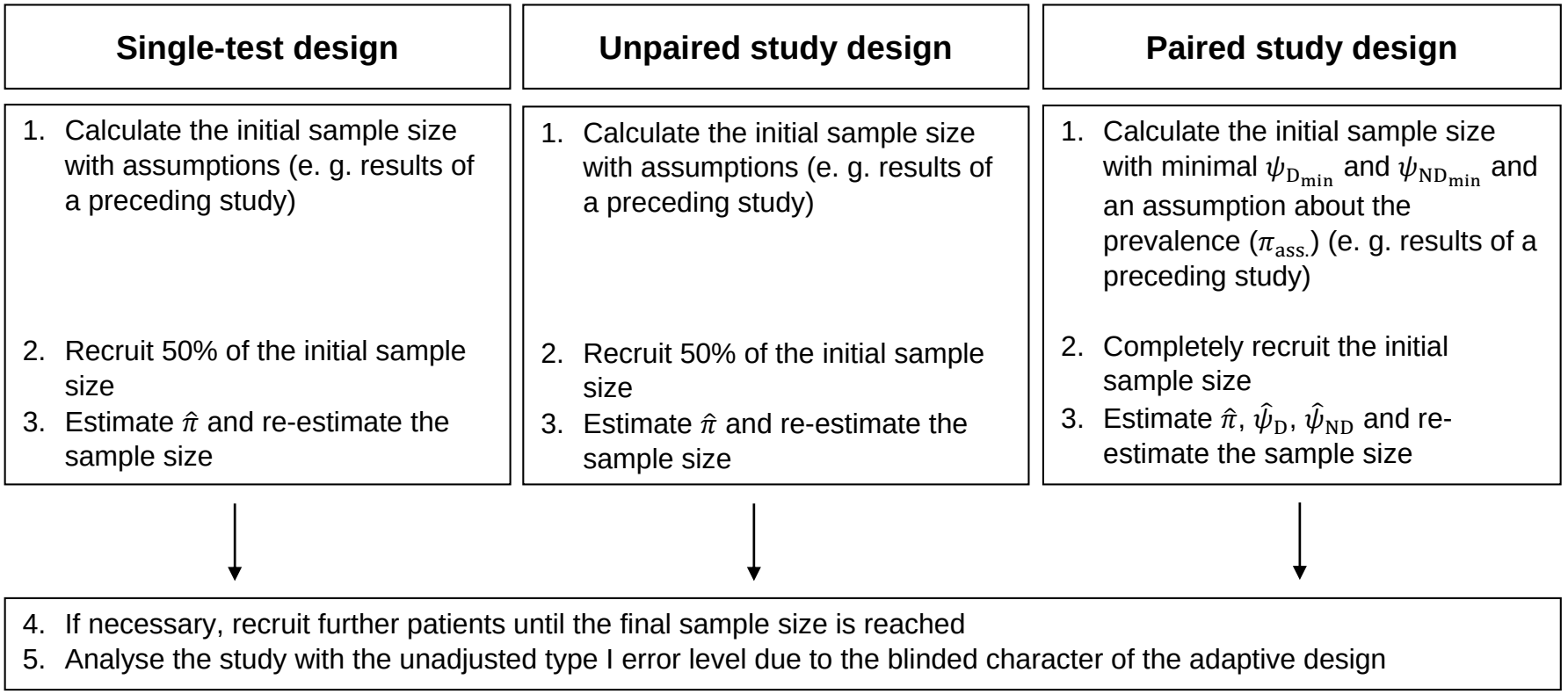



**Figure 3.** Procedure of the blinded adaptive design in the single-test, unpaired and paired diagnostic study (Adapted from *Thesis Article 3*: Stark et al., 2022).

The procedure for the blinded sample size re-estimation consists of five steps (Brinton et al., 2015). Figure 3 shows these steps for each study design. First, we calculate the initial sample size and second, we recruit the number of study participants to complete the sample size needed for the internal pilot study. Third, the internal pilot study serves to estimate nuisance parameters based on this interim data during the study (Friede & Kieser, 2006). Based on estimated nuisance parameters, we adjust the sample size. In this context, the re-estimated sample size may also be smaller than the initial one which Wittes et al. (1999) and Zucker et al. (1999) call the unrestricted sample size re-estimation.

Table 1 shows those nuisance parameters which we estimate depending on the study design.In the single-test and unpaired design, the only nuisance parameter is the disease prevalence ($\pi$), whereas proportions of discordant test results in the diseased ($\psi_{\mathrm{D}}$) and non-diseased ($\psi_{\mathrm{ND}}$) population additionally occur in the paired design. Discordant test results are those in which the experimental and comparator test yield different results.

In the fourth and fifth step, we recruit further study participants if necessary, to complete the adjusted sample size and finally analyse the study. This procedure does not reveal the sensitivity and specificity of the experimental test which is why we call it a blinded adaptive design (Zapf et al., 2020). Due to the blinded character, the sample size re-estimation controls the type I error rate. Hence, we may perform the final analysis without any adjustment for multiplicity (Friede & Kieser, 2013; Wu et al., 2008).

In *Thesis Article 2*, we explored that repeated estimation of the prevalence in the single-test design does not offer any advantages regarding the mean squared error of the estimated prevalence compared to the one-time estimation described above (Stark & Zapf, 2020). Hence, I do not include the repeated re-estimation procedure in this synopsis but we present results in *Thesis Article 2* (Stark & Zapf, 2020).

Figure 3 reveals that most prominent differences regarding the procedure of the adaptive design exist between study designs regarding both the estimated nuisance parameters and the sample size of the internal pilot study. Following subsections elaborate on these differences: Subsection 1.2.1.2.1 explains how to estimate nuisance parameters, and subsection 1.2.1.2.2 shows the sample size for interim analysis.

#### 1.2.1.2.1 Estimation of nuisance parameters

Figure 4 depicts how we can summarize the results of the internal pilot study. Given parameters denote results of interim data which we use to estimate nuisance parameters.

**(a) Single-test design**

| | | Reference Standard | |
|---|---|---|---|
| | | Diseased | Non-diseased |
| Experimental Test | Positive | True Positive $TP_E$ | False Positive $FP_E$ |
| | Negative | False Negative $FN_E$ | True Negative $TN_E$ |
| | | $n_D$ | $n_{ND}$ |

**(b) Unpaired design**

| | | Reference Standard | |
|---|---|---|---|
| | | Diseased | Non-diseased |
| Com-parator Test | Positive | $TP_C$ | $FP_C$ |
| | Negative | $FN_C$ | $TN_C$ |
| | | $n_{D_C}$ | $n_{ND_C}$ |

| | | Reference Standard | |
|---|---|---|---|
| | | Diseased | Non-diseased |
| Experi-mental Test | Positive | $TP_E$ | $FP_E$ |
| | Negative | $FN_E$ | $TN_E$ |
| | | $n_{D_E}$ | $n_{ND_E}$ |

**(c) Paired design**

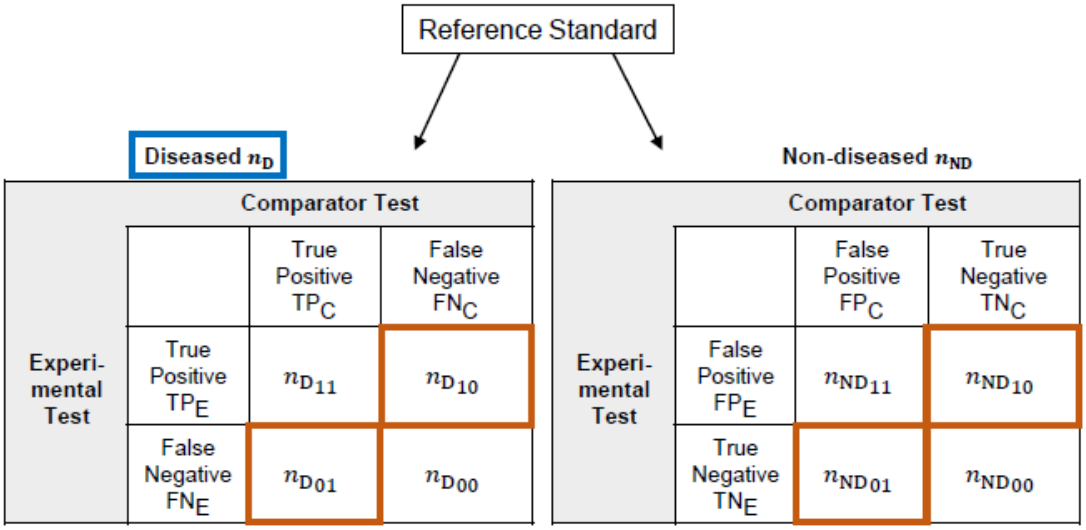



**Figure 4.** Results of the internal pilot study. The prevalence ($\pi$) is estimated based on the blue fields. Proportion of discordant test results are estimated based on the brown fields.

In the single-test and unpaired design, contingency tables tabulate positive and negative results of the experimental and comparator test, if necessary, in combination with the true disease status of the reference standard. In the paired design, the reference standard separates the study cohort into the diseased and non-diseased population whose results are each summarized in a contingency table. The split between both tables corresponds to the disease prevalence. Each table combines the results of the experimental and comparator test.

Figure 4 shows parameters needed to estimate the prevalence in blue colour and parameters needed to estimate proportions of discordant test results in brown colour. We estimate the prevalence ($\hat{\pi}$) with the maximum likelihood estimator of a binomial proportion (Brown et al., 2001). We consider the number of diseased individuals involved in the interim analysis ($n_{\mathrm{D}} = n_{\mathrm{D_E}} + n_{\mathrm{D_C}}$) and the total sample size for the interim analysis ($n$):

$$\hat{\pi} = \frac{n_{\mathrm{D}}}{n} = \frac{n_{\mathrm{D_E}} + n_{\mathrm{D_C}}}{n} \tag{7}$$

In the paired design, we estimate proportions of discordant test results in the diseased ($\hat{\psi}_{\mathrm{D}}$) and non-diseased ($\hat{\psi}_{\mathrm{ND}}$) population with the maximum likelihood estimator of a multinomial distribution (Held & Sabanés Bové, 2014):

$$\hat{\psi}_{\mathrm{D}} = \frac{n_{\mathrm{D}_{10}} + n_{\mathrm{D}_{01}}}{n_{\mathrm{D}}} \tag{8}$$

$$\hat{\psi}_{\mathrm{ND}} = \frac{n_{\mathrm{ND}_{10}} + n_{\mathrm{ND}_{01}}}{n_{\mathrm{ND}}} \tag{9}$$

#### 1.2.1.2.2 Sample size of the internal pilot study

The sample size of the internal pilot study describes the timing at which we perform the interim analysis to re-estimate the sample size. It differs between study designs as Table 3 shows.

**Table 3.** Size of the internal pilot study in each study design.

| | Study design | | |
|---|---|---|---|
| **Parameters** | **Single-test** | **Unpaired** | **Paired** |
| **Size of internal pilot study** | 50% of initial sample size based on an assumption about the prevalence | | sample size with minimal $\psi_{\mathrm{D}_{\mathrm{min}}}$ and $\psi_{\mathrm{ND}_{\mathrm{min}}}$ and an assumption about the prevalence |

In the single-test design, we perform sample size re-estimation after 50% of the initially calculated sample size. To determine whether this size of the internal pilot study is suitable, we evaluated in *Thesis Article 2* which proportion (λ) of the initially calculated sample size we need to recruit so that the re-estimated sample size ($\widehat{N}$) is close to the true sample size ($N_{\text{true}}$) (Stark & Zapf, 2020). Denne et al. (1999) denote this criterion as $R$:

$$R = \frac{\widehat{N}}{N_{\text{true}}} \tag{10}$$

We aim to re-estimate the true sample size as close as possible because it originates from true values of diagnostic accuracy and nuisance parameters. If this quotient $R$ is near 1, the re-estimated sample size is similar to the true sample size. Hence, the size of the internal pilot study is appropriate. Our investigations in *Thesis Article 2* reveal that the recruitment of 50% of the initial sample size is the minimum acceptable size. If we estimate nuisance parameters based on lower proportions of the initial sample size, re-estimated sample sizes can clearly differ from the true sample size. Otherwise, if we conduct the interim analysis later, we could not find smaller deviations of the re-estimated samples sizes from the true sample size. In the unpaired design, we adopted the timing of the interim analysis at 50% of the initial sample size without any further investigations because, in analogy to the single-test design, the prevalence is the only nuisance parameter to be estimated.

In the paired design, we calculate the initial sample size with minimal $\psi_{\text{D}}$ and $\psi_{\text{ND}}$ and an assumption about the prevalence. We completely recruit the initial sample size for interim analysis (McCray et al., 2017). Proportions of discordant test results vary between (Connor, 1987; Miettinen, 1968):

$$|\text{Se}_{\text{C}} - \text{Se}_{\text{E}}| \leq \psi_{\text{D}} \leq \text{Se}_{\text{C}} + \text{Se}_{\text{E}} - 2 \cdot \text{Se}_{\text{C}} \cdot \text{Se}_{\text{E}} \tag{11}$$

$$|\text{Sp}_{\text{C}} - \text{Sp}_{\text{E}}| \leq \psi_{\text{ND}} \leq \text{Sp}_{\text{C}} + \text{Sp}_{\text{E}} - 2 \cdot \text{Sp}_{\text{C}} \cdot \text{Sp}_{\text{E}} \tag{12}$$

Hence, minimal $\psi_{\text{D}}$ and $\psi_{\text{ND}}$ are:

$$\psi_{\text{D}_{\text{min}}} = |\text{Se}_{\text{C}} - \text{Se}_{\text{E}}| \tag{13}$$

$$\psi_{\text{ND}_{\text{min}}} = |\text{Sp}_{\text{C}} - \text{Sp}_{\text{E}}| \tag{14}$$

With $\psi_{\text{D}_{\text{min}}}$ and $\psi_{\text{ND}_{\text{min}}}$, the dependence between the experimental and comparator test is maximal which leads to the smallest possible sample size for the interim analysis. Only an incorrect assumption about the prevalence might enlarge the initial sample

size, and thus the sample size for the interim analysis, more than necessary. The results section 1.3.2 deals with possible consequences.

### 1.2.2 Simulation Studies

Through a simulation study, we compare the performance of the blinded adaptive design in each study design to a fixed design without sample size re-estimation. As performance measures, we use:

1. Type I error rate
2. Power
3. Mean of sample sizes of all simulation runs per scenario
4. Bias of the mean of estimated nuisance parameters:

$$\text{Bias of } \hat{\pi}_{\text{mean}} = \frac{\hat{\pi}_{\text{mean}} - \pi_{\text{true}}}{\pi_{\text{true}}} \quad (15)$$

$$\text{Bias of } \hat{\psi}_{\text{D}_{\text{mean}}} = \frac{\hat{\psi}_{\text{D}_{\text{mean}}} - \psi_{\text{D}_{\text{true}}}}{\psi_{\text{D}_{\text{true}}}} \quad (16)$$

$$\text{Bias of } \hat{\psi}_{\text{ND}_{\text{mean}}} = \frac{\hat{\psi}_{\text{ND}_{\text{mean}}} - \psi_{\text{ND}_{\text{true}}}}{\psi_{\text{ND}_{\text{true}}}} \quad (17)$$

5. Root mean squared error (RMSE) of the re-estimated sample size (Held & Sabanés Bové, 2014):

$$\text{RMSE} = \sqrt{E\left(\left(\hat{N} - N_{\text{true}}\right)^2\right)} \quad (18)$$

Table 4 shows simulated scenarios in each study design. In the single-test design, we evaluate a one-time re-estimation and a repeated re-estimation of the sample size. However, the repeated re-estimation procedure does not reveal any advantage compared to the one-time re-estimation which is why it is not depicted in Table 4. In the one-time re-estimation single-test design, we simulate scenarios with all possible parameter combinations which leads to 3888 scenarios. Whereas in comparative designs, there is an initial scenario based on the example study of McCray et al. (2017) testing for superiority in both endpoints. We extend the initial scenario to scenarios testing for a combination of superiority and non-inferiority or testing for non-inferiority in both endpoints. Starting from the initial scenario, we varied one parameter at a time, which leads to 58 scenarios in the unpaired design and 74 scenarios in the paired design.

**Table 4.** Simulated scenarios in the single-test, unpaired and paired design.Proportions of discordant test results are only relevant in the paired design. Section 1.3.2 shows results of scenarios highlighted in orange colour.

| | **Single-test design** | **Comparative design** | | | | |
|---|---|---|---|---|---|---|
| | One-time re-estimation | Initial scenario | Variation of initial scenario to show superiority in both endpoints | Variation of initial scenario to show non-inferiority in sensitivity and superiority in specificity | Variation of initial scenario to show superiority in sensitivity and non-inferiority in specificity | Variation of initial scenario to show non-inferiority in both endpoints |
| Simulation runs per scenario ($n_{sim}$) | 100,000 | 10,000 | | | | |
| Nominal significance level α | Per endpoint: 0.05 (two-sided)<br><br>Global: 0.05·0.05=0.0025 | Per endpoint: 0.05 (two-sided)<br><br>Global: 0.05·0.05=0.0025 | | Superiority endpoint: 0.05 (two-sided)<br>Non-inf. endpoint: 0.025 (one-sided)<br><br>Global: 0.05·0.025=0.00125 | | Per endpoint: 0.025 (one-sided)<br><br>Global: 0.025·0.025=0.000625 |
| Nominal overall target power | 0.8 | 0.8 | | | | |
| Minimum sensitivity $Se_{min}$ / sensitivity comparator test $Se_C$ | 0.6, 0.7, **0.8** | 0.8 | 0.6, 0.7 | 0.6, 0.7 | 0.6, 0.7 | 0.6, 0.7 |
| Minimum specificity $Sp_{min}$ / sensitivity comparator test $Sp_C$ | 0.6, **0.7**, 0.8 | 0.7 | 0.6, 0.8 | 0.6, 0.8 | 0.6, 0.8 | 0.6, 0.8 |
| True prevalence $\pi_{true}$ | 0.2, 0.4, 0.6, 0.8 | 0.2 | 0.4, 0.6, 0.8 | 0.4, 0.6, 0.8 | 0.4, 0.6, 0.8 | 0.4, 0.6, 0.8 |
| Assumed prevalence $\pi_{ass.}$ | $\pi_{true} - 0.1$, $\boldsymbol{\pi_{true} + 0.1}$ | $\pi_{true}$ + 0.1 | $\pi_{true}$ - 0.1<br>$\pi_{true}$ + 0.2<br>$\pi_{true}$ + 0.3 | $\pi_{true}$ - 0.1<br>$\pi_{true}$ + 0.2<br>$\pi_{true}$ + 0.3 | $\pi_{true}$ - 0.1<br>$\pi_{true}$ + 0.2<br>$\pi_{true}$ + 0.3 | $\pi_{true}$ - 0.1<br>$\pi_{true}$ + 0.2<br>$\pi_{true}$ + 0.3 |
| True discordant results diseased population $\psi_{D_{true}}$ | - | 0.11 (0.15, if: $Se_E - Se_C = 0.15$) | 0.16, 0.32 | 0.16, 0.32 | 0.18, 0.26 | 0.16, 0.32 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Assumed discordant results diseased population $\psi_{\mathrm{D}_{\mathrm{ass.}}}$ | - | 0.18 | 0.18 | 0.18 | 0.18 | 0.18 |
| True discordant results in the non-diseased population $\psi_{\mathrm{ND}_{\mathrm{true}}}$ | - | 0.14 (0.15, if: $\mathrm{Sp_E} - \mathrm{Sp_C} = 0.15$) | 0.24, 0.38 | 0.24, 0.38 | 0.21, 0.42 | 0.21, 0.42 |
| Assumed discordant results in the non-diseased population $\psi_{\mathrm{ND}_{\mathrm{ass.}}}$ | - | 0.24 | 0.24 | 0.24 | 0.24 | 0.24 |
| Non-inferiority margin Δ of relevant endpoint | - | - | - | 0.05, 0.1, 0.15 | 0.05, 0.1, 0.15 | 0.05, 0.1, 0.15 |
| Under the null hypothesis | $\mathrm{H}_{0_{\mathrm{Se}}}: \mathrm{Se_E} = \mathrm{Se_{min}} \cup \mathrm{H}_{0_{\mathrm{Sp}}}: \mathrm{Sp_E} = \mathrm{Sp_{min}}$ | $\mathrm{H}_{0_{\mathrm{Se}}}: \mathrm{Se_E} = \mathrm{Se_C} \cup \mathrm{H}_{0_{\mathrm{Sp}}}: \mathrm{Sp_E} = \mathrm{Sp_C}$ | | $\mathrm{H}_{0_{\mathrm{Se}}}: \mathrm{Se_E} \leq \mathrm{Se_C} - \Delta \cup \mathrm{H}_{0_{\mathrm{Sp}}}: \mathrm{Sp_E} = \mathrm{Sp_C}$ | $\mathrm{H}_{0_{\mathrm{Se}}}: \mathrm{Se_E} = \mathrm{Se_C} \cup \mathrm{H}_{0_{\mathrm{Sp}}}: \mathrm{Sp_E} \leq \mathrm{Sp_C} - \Delta$ | $\mathrm{H}_{0_{\mathrm{Se}}}: \mathrm{Se_E} \leq \mathrm{Se_C} - \Delta \cup \mathrm{H}_{0_{\mathrm{Sp}}}: \mathrm{Sp_E} \leq \mathrm{Sp_C} - \Delta$ |
| Fraction for re-estimation λ | 0.02, 0.1, 0.3, 0.5, 0.7 | Unpaired design 0.5<br>Paired design: sample size with minimal proportions of discordant test results | | | | |
| Sensitivity of the experimental test $\mathrm{Se_E}$ | $\mathrm{Se_{min}}$ | $\mathrm{Se_C}$ | $\mathrm{Se_C}$ | $\mathrm{Se_C} - \Delta$ | $\mathrm{Se_C}$ | $\mathrm{Se_C} - \Delta$ |
| Specificity of the experimental test $\mathrm{Sp_E}$ | $\mathrm{Sp_{min}}$ | $\mathrm{Sp_C}$ | $\mathrm{Sp_C}$ | $\mathrm{Sp_C}$ | $\mathrm{Sp_C} - \Delta$ | $\mathrm{Sp_C} - \Delta$ |
| Under the alternative hypothesis | $\mathrm{H}_{1_{\mathrm{Se}}}: \mathrm{Se_E} \neq \mathrm{Se_{min}} \cap \mathrm{H}_{1_{\mathrm{Sp}}}: \mathrm{Sp_E} \neq \mathrm{Sp_{min}}$ | $\mathrm{H}_{1_{\mathrm{Se}}}: \mathrm{Se_E} \neq \mathrm{Se_C} \cap \mathrm{H}_{1_{\mathrm{Sp}}}: \mathrm{Sp_E} \neq \mathrm{Sp_C}$ | | $\mathrm{H}_{0_{\mathrm{Se}}}: \mathrm{Se_E} > \mathrm{Se_C} - \Delta \cap \mathrm{H}_{0_{\mathrm{Sp}}}: \mathrm{Sp_E} \neq \mathrm{Sp_C}$ | $\mathrm{H}_{0_{\mathrm{Se}}}: \mathrm{Se_E} \neq \mathrm{Se_C} \cap \mathrm{H}_{0_{\mathrm{Sp}}}: \mathrm{Sp_E} > \mathrm{Sp_C} - \Delta$ | $\mathrm{H}_{0_{\mathrm{Se}}}: \mathrm{Se_E} > \mathrm{Se_C} - \Delta \cap \mathrm{H}_{0_{\mathrm{Sp}}}: \mathrm{Sp_E} > \mathrm{Sp_C} - \Delta$ |
| Fraction for re-estimation λ | 0.5 | Unpaired design 0.5<br>Paired design: sample size with minimal proportions of discordant test results | | | | |
| Sensitivity of the experimental test $\mathrm{Se_E}$ | $\mathrm{Se_{min}} + 0.05$, **$\mathrm{Se_{min}} + 0.1$,** $\mathrm{Se_{min}} + 0.15$ | $\mathrm{Se_C}$ + 0.1 | $\mathrm{Se_C}$ + 0.05<br>$\mathrm{Se_C}$ + 0.15 | $\mathrm{Se_C}$ | $\mathrm{Se_C}$ + 0.05<br>$\mathrm{Se_C}$ + 0.15 | $\mathrm{Se_C}$ |
| Specificity of the experimental test $\mathrm{Sp_E}$ | $\mathrm{Sp_{min}} + 0.05$, **$\mathrm{Sp_{min}} + 0.1$,** $\mathrm{Sp_{min}} + 0.15$ | $\mathrm{Sp_C}$ + 0.1 | $\mathrm{Sp_C}$ + 0.05<br>$\mathrm{Sp_C}$ + 0.15 | $\mathrm{Sp_C}$ + 0.05<br>$\mathrm{Sp_C}$ + 0.15 | $\mathrm{Sp_C}$ | $\mathrm{Sp_C}$ |

In the single-test design, we perform 100,000 simulation runs per scenario ($n_{\text{sim}}$). In those scenarios testing for superiority in both endpoints, this leads to a Monte Carlo standard error (MCSE) of (Morris et al., 2019):

$$\begin{aligned} \text{MCSE}_{\text{single-test}} &= \sqrt{\frac{\alpha_{\text{global}} \cdot \left(1 - \alpha_{\text{global}}\right)}{n_{\text{sim}}}} \\ &= \sqrt{\frac{0.0025 \cdot (1 - 0.0025)}{100\,000}} = 0.00016 \end{aligned} \tag{19}$$

To save computing capacity, we used 10 000 simulation runs per scenario in comparative designs. In those scenarios testing for superiority in both endpoints, this gives a MCSE of:

$$\text{MCSE}_{\text{comparative}} = \sqrt{\frac{0.0025 \cdot (1 - 0.0025)}{10\,000}} = 0.00050 \tag{20}$$

Table 5 shows distributions involved in the data generation mechanism. We use the statistical software R versions 3.5.0 and 4.0.2 to perform simulations with the default random number generator Mersenne-Twister, but with the own initialization method of *R* (Matsumoto & Nishimura, 1998; R Core Team, 2018, 2020).

**Table 5.** Data generation mechanism in the single-test, unpaired and paired design in the simulation study.($Bin$: binomial distribution, $MVBin$: multivariate binomial distribution, $k$: number of trials, $p$: success probability, $\rho$: dependence between both tests, $N$: total sample size, $n_{D_E}$: diseased individuals in experimental group, $n_{D_C}$: diseased individuals in comparator group)

| | Single-test design | Unpaired design | Paired design |
|---|---|---|---|
| **Diseased individuals ($n_D$) according to the reference standard** | $n_{\mathrm{D_E}} \sim Bin(k = N, p = \pi_{\mathrm{true}})$ | $n_{\mathrm{D_E}} \sim Bin(k = N, p = \pi_{\mathrm{true}})$<br>$n_{\mathrm{D_C}} \sim Bin(k = N, p = \pi_{\mathrm{true}})$ | $n_{\mathrm{D}} \sim Bin(k = N, p = \pi_{\mathrm{true}})$ |
| **True Positive Results (TP)** | $TP_{\mathrm{E}} \sim Bin(k = n_{\mathrm{D_E}}, p = \mathrm{Se_E})$ | $TP_{\mathrm{E}} \sim Bin(k = n_{\mathrm{D_E}}, p = \mathrm{Se_E})$<br>$TP_{\mathrm{C}} \sim Bin(k = n_{\mathrm{D_C}}, p = \mathrm{Se_C})$ | $(TP_{\mathrm{E}}, TP_{\mathrm{C}}) \sim MVBin(k_{\mathrm{E}} = n_{\mathrm{D_E}}, k_{\mathrm{C}} = n_{\mathrm{D_C}}, p_{\mathrm{E}} = Se_{\mathrm{E}}, p_{\mathrm{C}} = Se_{\mathrm{C}}, \rho = TPPR)$<br>with<br>$TPPR = \frac{Se_{\mathrm{C}} + Se_{\mathrm{E}} - \psi_{\mathrm{D_{true}}}}{2}$ |
| **True Negative Results (TN)** | $TN_{\mathrm{E}} \sim Bin(k = N - n_{\mathrm{D_E}}, p = \mathrm{Sp_E})$ | $TN_{\mathrm{E}} \sim Bin(k = N - n_{\mathrm{D_E}}, p = \mathrm{Sp_E})$<br>$TN_{\mathrm{C}} \sim Bin(k = N - n_{\mathrm{D_C}}, p = \mathrm{Sp_C})$ | $(TN_{\mathrm{E}}, TN_{\mathrm{C}}) \sim MVBin(k_{\mathrm{E}} = N - n_{\mathrm{D_E}}, k_{\mathrm{C}} = N - n_{\mathrm{D_C}}, p_{\mathrm{E}} = Sp_{\mathrm{E}}, p_{\mathrm{C}} = Sp_{\mathrm{C}}, \rho = TNNR)$<br>with<br>$TNNR = \frac{Sp_{\mathrm{C}} + Sp_{\mathrm{E}} - \psi_{\mathrm{ND_{true}}}}{2}$ |

## 1.3 Results

The following section applies the blinded sample size re-estimation procedure to the example study. Furthermore, this section provides results of simulation studies.

### 1.3.1 Example Study

McCray et al. already proposed in 2017 the blinded sample size re-estimation for a paired diagnostic study. However, their approach differs from ours in the definition of endpoints, hypothesis and sample size calculation. Table 6 summarizes differences and similarities between both approaches. The most important development of our approach is the implementation of the optimal sample size calculation in the adaptive design procedure. This section reveals the importance of the optimal sample size calculation by comparing the results of the sample size re-estimation between both approaches in the context of the example study introduced in section 1.2.1.1.1.

For sample size calculation, McCray et al. (2017) consider the quotient of sensitivities and specificities of both diagnostic tests as endpoints. Furthermore, they use the true-positive-positive-rate (TPPR) and the true-negative-negative-rate (TNNR) as a parameter of dependency between both tests. TPPR represents the proportion of test results in which the comparator and experimental test correctly diagnose a diseased individual. Vice versa, TNNR denotes the proportion of test results in which both tests correctly lead to a negative test results. McCray et al. (2017) use the maximal possible TPPR ($TPPR_{max}$) and maximal possible TNNR ($TNNR_{max}$) for the initial sample size calculation. They calculate the initial sample size with the conventional three steps with a power of 80% per endpoint which leads to a theoretical overall power of 64% resulting from the product of both individual powers.

In contrast to McCray et al. (2017), our approach applies the optimal sample size calculation using the difference in sensitivities and specificities of both diagnostic tests, as recommended by the *Guideline on clinical evaluation of diagnostic agents* (CHMP, 2009). Additionally, we use proportions of discordant test results as a parameter of dependency between both tests and plan the sample size with an overall power of 80%.

Table 6 compares initial sample sizes, sample sizes for interim analysis and re-estimated samples sizes of both adaptive designs. The optimal approach enables smaller samples sizes by avoiding to overpower one of both endpoints.

**Table 6.** Comparison of our blinded sample size re-estimation approach with McCray et al. (2017) (*Thesis Article 3*: Stark et al., 2022).

| | | McCray et al. (2017) | Our approach |
|---|---|---|---|
| **General information** | **Endpoint** | $\frac{\mathrm{Se_E}}{\mathrm{Se_C}}$ and $\frac{\mathrm{Sp_E}}{\mathrm{Sp_C}}$ | $\mathrm{Se_E} - \mathrm{Se_C}$ and $\mathrm{Sp_E} - \mathrm{Sp_C}$ |
| | **$\mathbf{H_{0_{global}}}$** | $\mathrm{H_{0_{Se}}}: \frac{\mathrm{Se_E}}{\mathrm{Se_C}} = 1 \cup$ $\mathrm{H_{0_{Sp}}}: \frac{\mathrm{Sp_E}}{\mathrm{Sp_C}} = 1$ | $\mathrm{H_{0_{Se}}}: \mathrm{Se_E} - \mathrm{Se_C} = 0 \cup$ $\mathrm{H_{0_{Sp}}}: \mathrm{Sp_E} - \mathrm{Sp_C} = 0$ |
| | **Sample size calculation** | Conventional approach<br>$\alpha$ per endpoint: 0.05 (two-sided)<br>Power per endpoint: 0.8 | Optimal approach<br>$\alpha$ per endpoint: 0.05 (two-sided)<br>Overall power: 0.8 |
| | **Parameter of dependency between both tests** | $\mathrm{TPPR} = \frac{n_{\mathrm{D}_{11}}}{n_{\mathrm{D}}}$ $\mathrm{TNNR} = \frac{n_{\mathrm{ND}_{00}}}{n_{\mathrm{ND}}}$ | $\psi_{\mathrm{D}} = \frac{n_{\mathrm{D}_{10}} + n_{\mathrm{D}_{01}}}{n_{\mathrm{D}}}$ $\psi_{\mathrm{ND}} = \frac{n_{\mathrm{ND}_{10}} + n_{\mathrm{ND}_{01}}}{n_{\mathrm{ND}}}$ |
| **Initial sample size calculation** | **Size of internal pilot study** | $\mathrm{TPPR_{max}}$ and $\mathrm{TNNR_{max}}$ correspond to $\psi_{\mathrm{D_{min}}}$ and $\psi_{\mathrm{ND_{min}}}$ | |
| | **Parameter of dependency between both tests for initial sample size calculation** | $\mathrm{TPPR_{max}} = \mathrm{Se_C} = 0.81$ $\mathrm{TNNR_{max}} = \mathrm{Sp_C} = 0.66$ | $\psi_{\mathrm{D_{min}}} = \lvert\mathrm{Se_C} - \mathrm{Se_E}\rvert = 0.09$ $\psi_{\mathrm{ND_{min}}} = \lvert\mathrm{Sp_C} - \mathrm{Sp_E}\rvert = 0.14$ |
| | **Initial sample size, size of internal pilot study** | 186 | 133 |
| **Sample size re-estimation** | **Estimation of nuisance parameters** | $\hat{\pi} = 0.44$ $\widehat{TPPR} = 0.80$ $\widehat{TNNR} = 0.66$ | $\hat{\pi} = 0.44$ $\hat{\psi}_{\mathrm{D}} = 0.11$ $\hat{\psi}_{\mathrm{ND}} = 0.14$ |
| | **Re-estimated sample size** | 242 | 200 |

### 1.3.2 Simulation Studies

This section summarizes simulations results in those scenarios highlighted in orange colour in Table 4. These scenarios include the initial scenario of the example study with a variation of the true prevalence ($\pi_{\mathrm{true}}$) in those settings testing for superiority or non-inferiority in both endpoints and both possible combinations of superiority and non-inferiority. This section neither evaluates the choice of the sample size for interim analysis nor performance metrics of the repeated estimation of nuisance parameters in the single-test design. Instead, *Thesis Article 2* shows these results (Stark & Zapf, 2020). *Thesis Article 3* provides further simulation results of scenarios testing for non-inferiority in both endpoints and combinations of superiority and non-inferiority (Stark et al., 2022).

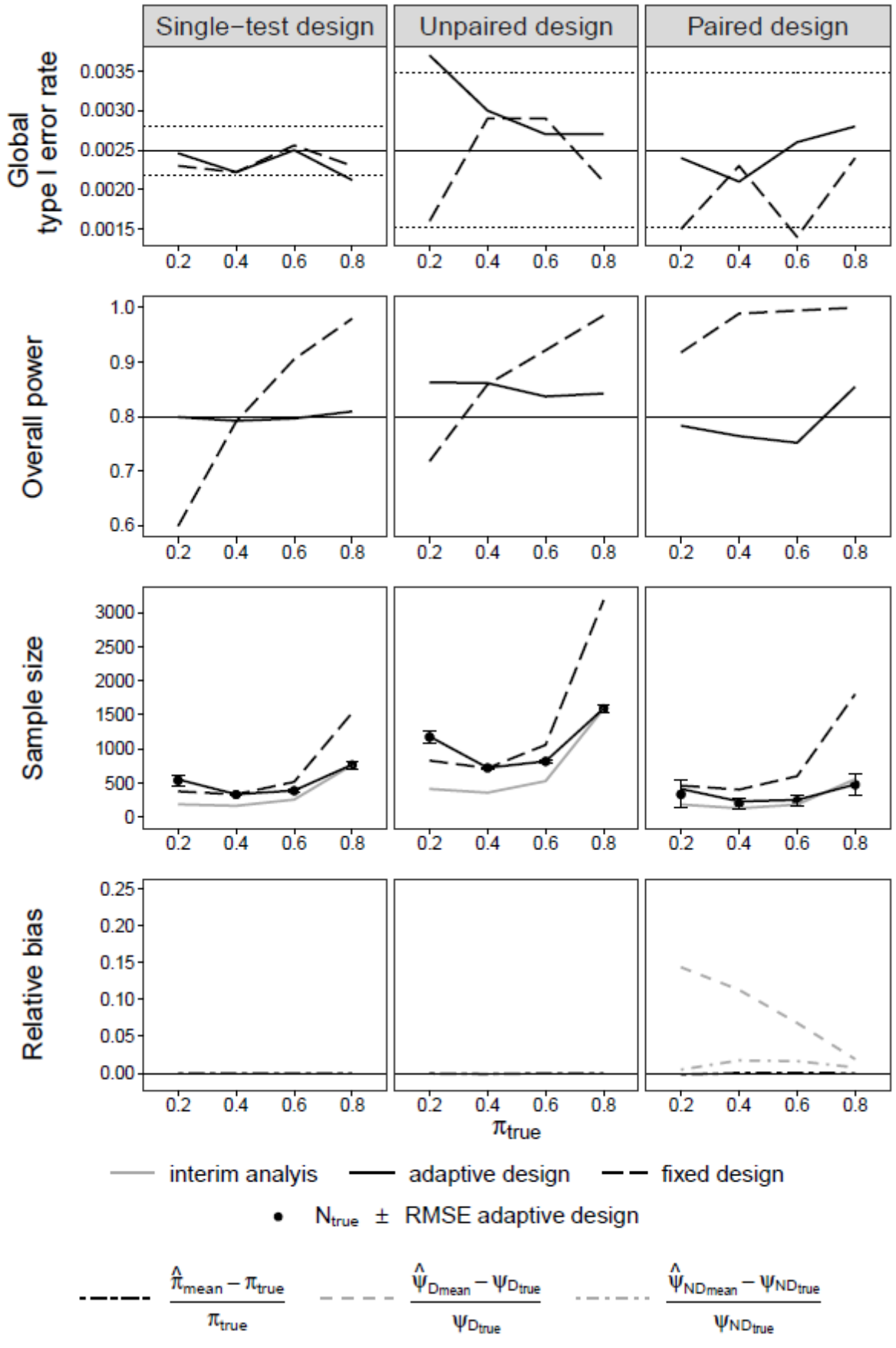



**Figure 5.** Type I error rate, power, sample sizes and relative bias in the single-test, unpaired and paired design. Black dotted lines represent intervals based on the Monte Carlo standard error (single-test design: ±1.96·0.00016=0.0003, comparative design: ±1.96·0.0005=0.00098)

Figure 5 opposes global type I error rates, overall powers, sample sizes and biases of the fixed and adaptive design in the single-test, unpaired and paired design testing for superiority in both endpoints. Generally, the adaptive design controls the global type I error rate. The same applies to scenarios testing for non-inferiority in both endpoints and combinations of superiority and non-inferiority (see Online Supplement of *Thesis Article 3*: Stark et al., 2022).

The overall power of the fixed design increases with an increasing true prevalence. With a low prevalence, the sample size needed to show the sensitivity determines the final sample size. In depicted scenarios, the assumed prevalence is higher than the true prevalence which leads to the problem of dividing by a too high prevalence during the sample size calculation. In the single-test and unpaired fixed design, this leads to a smaller sample size than the true sample size. Thus, the study is underpowered. Vice versa, with a high prevalence, the sample size needed to show the specificity is divided by a too low number of non-diseased individuals. This results in a higher sample size than the true sample size in the fixed design. Hence, the study in the single-test and unpaired design is overpowered. In the paired design, these mechanisms work similarly, but in addition, assumed proportions of discordant test results are too high. Overall, sample sizes are higher than truly necessary. Consequently, the fixed design in the depicted scenarios is overpowered. *Thesis Article 3* presents the influence of varying true proportions of discordant test results on the overall power and sample size in the paired design (Stark et al., 2022). In contrast to the fixed design, the adaptive design compensates for wrong assumptions about nuisance parameters and comes close to the overall target power.

Table 7 takes an even closer look at the power of the adaptive design. It picks up the power in the unpaired and paired design shown in Figure 5 and adds the power of those settings testing for non-inferiority in at least one endpoint. In the unpaired design, the empirical overall power is slightly higher than the overall target power. In those scenarios testing for non-inferiority in at least one endpoint, empirical overall power equals the overall target power or is even lower.

**Table 7.** Simulated empirical overall power in the unpaired and paired adaptive design depending on the hypothesis and true prevalence ($\pi_{true}$).

| Hypothesis | $\pi_{\text{true}}$ | Unpaired design | Paired design |
|---|---|---|---|
| **Superiority in both endpoints** | 0.2 | 0.863 | 0.783 |
| | 0.4 | 0.861 | 0.764 |
| | 0.6 | 0.837 | 0.752 |
| | 0.8 | 0.842 | 0.855 |
| **Superiority in sensitivity Non-inferiority in specificity** | 0.2 | 0.872 | 0.788 |
| | 0.4 | 0.840 | 0.810 |
| | 0.6 | 0.808 | 0.826 |
| | 0.8 | 0.815 | 0.829 |
| **Non-inferiority in sensitivity Superiority in specificity** | 0.2 | 0.804 | 0.822 |
| | 0.4 | 0.814 | 0.822 |
| | 0.6 | 0.831 | 0.756 |
| | 0.8 | 0.844 | 0.855 |
| **Non-inferiority in both endpoints** | 0.2 | 0.706 | 0.814 |
| | 0.4 | 0.698 | 0.829 |
| | 0.6 | 0.708 | 0.825 |
| | 0.8 | 0.722 | 0.828 |

Vice versa, referring to Figure 5, the empirical overall power of the paired adaptive design is lower than the overall target power unless $\pi_{\text{true}}$ equals 80%. In this scenario, the proportion of non-diseased individuals is initially assumed to be lower than in truth. This leads to the fact that the sample size for the interim analysis is already larger than the true sample size. Consequently, the paired adaptive design is overpowered in this scenario. Table 7 shows that empirical overall power rises in those scenarios testing for non-inferiority in at least one of both endpoints in the paired design.

Considering Figure 5, RMSE is highest with a low or high true prevalence, especially in the paired design. With a low or high true prevalence, there is either a low number of diseased or non-diseased individuals which can result in an uncertain estimation of the prevalence and of proportions of discordant test results in the diseased or non-diseased population. It is also noticeable that the sample size in the paired design is fundamentally lower than the one in the unpaired design.

The adaptive design estimates the prevalence and proportions of discordant test results in the non-diseased population without any relevant bias. The bias of estimated proportions of discordant test results in the diseased population is 14% if the prevalence is the lowest. With an increasing true prevalence, the bias decreases as there is a higher number of diseased individuals on the basis of which we estimate proportions of discordant test results in the diseased population.

## 1.4 Discussion

This thesis contributes to answer three research questions:

1. Verification of the need for adaptive designs in diagnostic accuracy studies in *Thesis Article 1* (Zapf et al., 2020)
2. Development of the optimal sample size calculation for two independent co-primary endpoints to avoid overpowered diagnostic accuracy studies in *Thesis Articles 2 and 3* (Stark et al., 2022; Stark & Zapf, 2020)
3. Development of blinded adaptive designs for diagnostic accuracy studies to adjust the sample size based on the estimation of nuisance parameters in *Thesis Articles 2 and 3* (Stark et al., 2022; Stark & Zapf, 2020)

On the one hand, there are several strengths of the proposed methods. First, the idea of the optimal sample size calculation is not limited to diagnostic studies considering sensitivity and specificity as co-primary endpoints. To remember, the idea of the optimal sample size calculation was to split the overall power to both independent co-primary endpoints so that the product of both individual powers results in the target overall power. In my view, it can be transferred to any studies with two independent co-primary endpoints.

Second, the comparison with the approach of McCray et al. (2017) reveals that our blinded adaptive design approach based on the optimal sample size calculation requires lower sample sizes and comes closer to the target power, especially if the dependence between both diagnostic procedures is the highest. Additionally, we offer the possibility to apply our approach in settings either testing for non-inferiority in both endpoints or a combination of superiority and non-inferiority which is not thought of with the approach of McCray et al. (2017).

Third, we evaluate the performance of proposed methods in realistic simulation scenarios and in scenarios with extreme parameter combinations.

On the other hand, I would like to point out these limitations: first, sample size calculations fit to evaluations with Wald confidence intervals. However, coverage probabilities of Wald confidence intervals are poor (Agresti & Caffo, 2000; Agresti & Coull, 1998; Agresti & Min, 2005). Therefore, we propose to use the logit transformed Wald confidence interval for analysis in the single-test design. Our simulation results

in *Thesis Article 2* show that the empirical overall power comes close to the target power (Stark & Zapf, 2020). For comparative designs, Fagerland et al. (2015) and Fagerland et al. (2014) evaluate that the Miettinen-Nurminen confidence interval and Tango's confidence interval have higher coverage probability than the simple Wald confidence interval for the difference of two independent or dependent proportions. The Miettinen-Nurminen interval is an asymptotic score confidence interval for the difference of two independent proportions (Fagerland et al., 2015). In contrast, Tango's interval represents an asymptotic score confidence interval for the difference of two paired proportions (Fagerland et al., 2014). However, it is noticeable in our simulation results that the empirical overall power in the paired design is lower than the target power if both sensitivity and specificity are tested for superiority. Whereas in those scenarios testing for non-inferiority in both endpoints, the target power is reached. Tango (1998) shows that his proposed confidence interval has a high coverage probability in non-inferiority settings. Vice versa, in the unpaired design, our simulated empirical overall power is higher than the target power if both endpoints are tested for superiority and lower than the target power testing for non-inferiority twice. Nevertheless, we propose these two score confidence intervals for consistency of analyses in both comparative designs.

Second, the blinded adaptive design procedure does not include the possibility to define a maximum achievable sample size that the re-estimated sample size must not exceed, as e. g. Friede & Kieser (2011) did instead. Therefore, the re-estimated sample size may become unrealistically high in our approach. However, even if there is a maximum achievable sample size, the question arises whether it is meaningful to continue the study until the maximum achievable sample size is recruited, if it is far below the re-estimated sample size. This would lead to a lower power than desired.

Third, we did not compare our adaptive design approach to the approach of McCray et al. (2017) in an extensive simulation study because we choose different endpoints. However, as mentioned above, we compare both approaches for the example study.

There are three further aspects which I would like to discuss: First, considering the adaptive design in a paired study, I would like to reflect whether sample size re-estimation is still blinded by revealing both proportions of discordant test results. Chang (2014) describes a semi-blinded sample size re-estimation procedure in which he derives the treatment effect in a therapeutic study from an estimated pooled

variance of normally distributed data. In a paired diagnostic study, differences between the experimental and comparator test are (Agresti & Min, 2005):

$$\delta_{\mathrm{Se}} = \mathrm{Se}_{\mathrm{E}} - \mathrm{Se}_{\mathrm{C}} = \frac{n_{\mathrm{D}_{10}} - n_{\mathrm{D}_{01}}}{n_{\mathrm{D}}} \tag{21}$$

$$\delta_{\mathrm{Sp}} = \mathrm{Sp}_{\mathrm{E}} - \mathrm{Sp}_{\mathrm{C}} = \frac{n_{\mathrm{ND}_{10}} - n_{\mathrm{ND}_{01}}}{n_{\mathrm{ND}}} \tag{22}$$

In my view, we cannot derive $\delta_{\mathrm{Se}}$ and $\delta_{\mathrm{Sp}}$ from estimated proportions of discordant test results given in equations (8) and (9). Together with controlling the type I error rate in simulations, this argues for a fully blinded adaptive design in the paired diagnostic study.

Second, Wittes et al. (1999) report type I error rate inflation in unblinded sample size re-estimation procedures with unrestricted design especially for small sample sizes. Proschan (2005) explains that potential issues regarding the type I error rate in the unblinded unrestricted design can arise if only a few additional individuals need to be recruited after an interim analysis. Kieser & Friede (2003) report that the type I error rate is not inflated in the blinded sample size re-estimation procedure with unrestricted design. Our investigations show that the type I error rate is controlled in blinded sample size re-estimation with the unrestricted design in diagnostic studies. Based on these results, we further recommend the unrestricted design as it offers the advantage of reducing the sample size in the interim analysis and avoiding an overpowered study. This may be useful in the single-test, unpaired and even paired design because an incorrect assumption about the disease prevalence may lead to a too high initial sample size in all three study designs.

Third, the question may arise whether the sample size for the interim analysis in the paired design is large enough to reliably estimate nuisance parameters as it is based on minimal proportions of discordant test results. Simulation results show a small bias in estimating proportions of discordant test results in the case of a small or high prevalence. Bias arises as there are only a few diseased or non-diseased individuals to estimate proportions of discordant test results in the diseased or non-diseased population, respectively. However, re-estimated sample sizes are close to true samples sizes based on true nuisance parameters. Hence, there is no indication that sample sizes for interim analyses are too small in the paired design.

In summary, I recommend applying blinded adaptive designs based on the optimal sample size calculation in diagnostic accuracy studies. They support to reach the target power without much additional effort.

In the DFG project ZA 687/1-1 which funded my thesis, my supervisor Prof. Dr. Antonia Zapf and colleagues additionally developed unblinded adaptive designs for early stopping due to futility or efficacy or to adjust the sample size based on estimated sensitivities and specificities of the experimental test. Furthermore, they evaluated adaptive designs for phase IV test-treatment studies as well as seamless designs to combine two phases of the development process of a diagnostic test.

In pandemic situations, blinded adaptive designs contribute to the pandemic response. In this context, diagnostic studies are related via a feedback-loop to national testing strategies and dynamic models to predict infection events. Governments specify which groups of people should be tested for the infectious disease through national testing strategies. Diagnostic studies evaluate the diagnostic accuracy of the test for this group of people and provide important insights for the modelling of the infection process. Based on the predicted infection process, governments can adjust the national testing strategy. Prof. Dr. Antonia Zapf and colleagues will establish the feedback-loop in further research in the DFG project ZA 687/3-1 on *Adaptive (seamless) designs for real-time evaluation of diagnostic tests and their usefulness for the parameterisation of dynamic spread models in epidemic and pandemic settings*.

In addition, Prof. Dr. Antonia Zapf and colleagues will explore the use of adaptive designs in diagnostic studies to estimate proportions of missing values in the DFG project ZA 687/6-1 on *Estimands and missing values in diagnostic studies*. Estimands in diagnostic studies become increasingly topical especially in view of the new Medical Device EU Regulatory 2017/745.

## 1.5 Conclusion

A confirmatory diagnostic accuracy study combines sensitivity and specificity as co-primary endpoints with the experimental diagnostic test either evaluated against pre-defined minimum thresholds in a single-test design or compared against a comparator diagnostic test in an unpaired or paired design.

In this thesis, I established two methods to support reaching the target power in a confirmatory diagnostic accuracy study. I developed both methods for the single-test, unpaired and paired design testing for superiority or non-inferiority in both endpoints or combinations of superiority and non-inferiority, respectively. The first method represents the optimal sample size calculation which helps to avoid over- or underpowered studies with two independent co-primary endpoints. The optimal sample size calculation individually splits the overall power to both endpoints so that the product of both individual powers gives the overall target power. The second method consists of a blinded adaptive design including the optimal sample size calculation to adjust the sample size based on the estimation of nuisance parameters. One of those nuisance parameters to be estimated in an interim analysis is the disease prevalence. In the paired design, we additionally estimate proportions of discordant test results in the diseased and non-diseased population.

I evaluated the blinded adaptive design based on the optimal sample size calculation using simulation studies and by applying it to a paired example study. Results show that the blinded sample size re-estimation procedure controls the type I error rate, comes close to the target power and estimates nuisance parameters without any relevant bias. Adjusted sample sizes are close to the true sample sizes which base on true nuisance parameters.

Following these results, I recommend applying blinded adaptive designs based on the optimal sample size calculation. Proposed methods support researchers planning and executing efficient confirmatory diagnostic accuracy studies reaching the target power without having to expend much additional effort. In light of the new Medical Device EU Regulatory 2017/745, the optimal sample size calculation approach and blinded adaptive designs contribute to the validity of diagnostic phase III studies for the approval of experimental tests.

# Bibliography

Agresti, A., & Caffo, B. (2000). Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures. *The American Statistician*, *54*(4), 280-288. https://doi.org/10.1080/00031305.2000.10474560

Agresti, A., & Coull, B. A. (1998). Approximate is better than “Exact” for Interval Estimation of Binomial Proportions. *The American Statistician*, *52*(2), 119-126. https://doi.org/10.1080/00031305.1998.10480550

Agresti, A., & Min, Y. (2005). Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine*, *24*(5), 729-740. https://doi.org/10.1002/sim.1781

Alonzo, T. A., Pepe, M. S., & Moskowitz, C. S. (2002). Sample size calculations for comparative studies of medical tests for detecting presence of disease. *Statistics in Medicine*, *21*(6), 835-852. https://doi.org/10.1002/sim.1058

Asakura, K., Hamasaki, T., & Evans, S. R. (2017). Interim evaluation of efficacy or futility in group-sequential trials with multiple co-primary endpoints. *Biometrical Journal*, *59*(4), 703-731. https://doi.org/10.1002/bimj.201600026

Bachmann, L. M., Puhan, M. A., ter Riet, G., & Bossuyt, P. M. (2006). Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ*, *332*, 1127-1129. https://doi.org/10.1136/bmj.38793.637789.2F

Bauer, P., Bretz, F., Dragalin, V., König, F., & Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine*, *35*(3), 325-347. https://doi.org/10.1002/sim.6472

Bhatt, D. L., & Mehta, C. (2016). Adaptive Designs for Clinical Trials. *The New England Journal of Medicine*, *375*, 65-74. https://doi.org/10.1056/NEJMra1510061

Bochmann, F., Johnson, Z., & Azuara-Blanco, A. (2007). Sample size in studies on diagnostic accuracy in ophthalmology: a literature survey. *British Journal of Ophthalmology*, *91*, 898-900. https://doi.org/10.1136/bjo.2006.113290

Bossuyt, P. M., Irwig, L., Craig, J., & Glasziou, P. (2006). Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ*, *332*, 1089-1092. https://doi.org/10.1136/bmj.332.7549.1089

Brinton, J. T., Ringham, B. M., & Glueck, D. H. (2015). An internal pilot design for prospective cancer screening trials with unknown disease prevalence. *Trials*, *16*, Article 458. https://doi.org/10.1186/s13063-015-0951-3

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, *16*(2), 101-133. https://doi.org/10.1214/ss/1009213286

Buderer, N. M. (1996). Statistical Methodology: I. Incorporating the Prevalence of Disease into the Sample Size Calculation for Sensitivity and Specificity. *Academic Emergency Medicine*, *3*(9), 895-900. https://doi.org/10.1111/j.1553-2712.1996.tb03538.x

Chang, M. (2014). *Adaptive Design Theory and Implementation Using SAS and R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b17761

Chow, S.-C., & Chang, M. (2006). *Adaptive Design Methods in Clinical Trials*. Chapman & Hall/CRC Press. https://doi.org/10.1201/9781584887775

Committee for Medicinal Products for Human Use (CHMP). (2007). *Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design*. European Medicines Agency. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf

Committee for Medicinal Products for Human Use (CHMP). (2009). *Guideline on clinical evaluation of diagnostic agents*. European Medicines Agency. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-evaluation-diagnostic-agents_en.pdf

Connor, R. J. (1987). Sample Size for Testing Differences in Proportions for the Paired-Sample Design. *Biometrics*, *43*(1), 207-211. https://doi.org/10.2307/2531961

Denne, J. S., & Jennison, C. (1999). Estimating the sample size for a t-test using an internal pilot. *Statistics in Medicine*, *18*(13), 1575-1585. https://doi.org/10.1002/(SICI)1097-0258(19990715)18:13<1575::AID-SIM153>3.0.CO;2-Z

European Union Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. (2017). *Official Journal of the European Union,* L 117.

Fagerland, M. W., Lydersen, S., & Laake, P. (2014). Recommended tests and confidence intervals for paired binomial proportions. *Statistics in Medicine*, *33*(16), 2850-2875. https://doi.org/10.1002/sim.6148

Fagerland, M. W., Lydersen, S., & Laake, P. (2015). Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research*, *24*(2), 224-254. https://doi.org/10.1177/0962280211415469

Flahault, A., Cadilhac, M., & Thomas, G. (2005). Sample size calculation should be performed for design accuracy in diagnostic test studies. *Journal of Clinical Epidemiology*, *58*(8), 859-862. https://doi.org/10.1016/j.jclinepi.2004.12.009

Friede, T., Häring, D. A., & Schmidli, H. (2019). Blinded continuous monitoring in clinical trials with recurrent event endpoints. *Pharmaceutical Statistics*, *18*(1), 54-64. https://doi.org/10.1002/pst.1907

Friede, T., & Kieser, M. (2006). Sample Size Recalculation in Internal Pilot Study Designs: A Review. *Biometrical Journal*, *48*(4), 537-555. https://doi.org/10.1002/bimj.200510238

Friede, T., & Kieser, M. (2011). Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis. *Pharmaceutical Statistics*, *10*(1), 8-13. https://doi.org/10.1002/pst.398

Friede, T., & Kieser, M. (2013). Blinded sample size re-estimation in superiority and noninferiority trials: bias versus variance in variance estimation. *Pharmaceutical Statistics*, *12*(3), 141-146. https://doi.org/10.1002/pst.1564

Friede, T., & Miller, F. (2012). Blinded continuous monitoring of nuisance parameters in clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *61*(4), 601-618. https://doi.org/10.1111/j.1467-9876.2011.01029.x

Gerke, O., Høilund-Carlsen, P. F., Poulsen, M. H., & Vach, W. (2012). Interim analyses in diagnostic versus treatment studies: differences and similarities. *American Journal of Nuclear Medicine and Molecular Imaging*, *2*(3), 344-352.

Hajian-Tilaki, K. (2014). Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of Biomedical Informatics*, *48*, 193-204. https://doi.org/10.1016/j.jbi.2014.02.013

Hamasaki, T., Evans, S. R., & Asakura, K. (2018). Design, data monitoring, and analysis of clinical trials with co-primary endpoints: A review. *Journal of Biopharmaceutical Statistics*, *28*(1), 28-51. https://doi.org/10.1080/10543406.2017.1378668

Held, L., & Sabanés Bové, D. (2014). *Applied Statistical Inference*. Springer. https://doi.org/10.1007/978-3-642-37887-4

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Chapman & Hall/CRC.

Jones, S. R., Carley, S., & Harrison, M. (2003). An introduction to power and sample size estimation. *Emergency Medicine Journal*, *20*, 453-458. https://doi.org/10.1136/emj.20.5.453

Kieser, M., & Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine*, *22*(23), 3571-3581. https://doi.org/10.1002/sim.1585

Köbberling, J., Trampisch, H.-J., & Windeler, J. (1990). Memorandum for the Evaluation of Diagnostic Measures. *Journal of Clinical Chemistry and Clinical Biochemistry*, *28*(12), 873-879.

Korevaar, D. A., Gopalakrishna, G., Cohen, J. F., & Bossuyt, P. M. (2019). Targeted test evaluation: a framework for designing diagnostic accuracy studies with clear study hypotheses. *Diagnostic and prognostic research*, *3*, Article 22. https://doi.org/10.1186/s41512-019-0069-2

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, *8*(1), 3-30. https://doi.org/10.1145/272991.272995

McCray, G. P., Titman, A. C., Ghaneh, P., & Lancaster, G. A. (2017). Sample size re-estimation in paired comparative diagnostic accuracy studies with a binary response. *BMC Medical Research Methodology*, *17*, Article 102. https://doi.org/10.1186/s12874-017-0386-5

Miettinen, O. S. (1968). The Matched Pairs Design in the Case of All-or-None Responses. *Biometrics*, *24*(2), 339-352. https://doi.org/10.2307/2528039

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074-2102. https://doi.org/10.1002/sim.8086

Moyé, L. A. (2006). *Statistical Monitoring of Clinical Trials: Fundamentals for Investigators*. Springer Science + Business Media.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Proschan, M. A. (2005). Two-Stage Sample Size Re-Estimation Based on a Nuisance Parameter: A Review. *Journal of Biopharmaceutical Statistics*, *15*(4), 559-574. https://doi.org/10.1081/bip-200062852

Proschan, M. A. (2009). Sample size re-estimation in clinical trials. *Biometrical Journal*, *51*(2), 348-357. https://doi.org/10.1002/bimj.200800266

R Core Team. (2018). *R: A language and environment for statistical computing* (Version 3.5.0) [Statistical software]. R Foundation for Statistical Computing. https://cran.r-project.org/bin/windows/base/old/3.5.0/

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.2) [Statistical software]. R Foundation for Statistical Computing. https://cran.r-project.org/bin/windows/base/old/4.0.2/

Sander, A., Rauch, G., & Kieser, M. (2017). Blinded sample size recalculation in clinical trials with binary composite endpoints. *Journal of Biopharmaceutical Statistics*, *27*(4), 705-715. https://doi.org/10.1080/10543406.2016.1198371

Stark, M., Hesse, M., Brannath, W., & Zapf, A. (2022). Blinded sample size re-estimation in a comparative diagnostic accuracy study. *BMC Medical Research Methodology*, *22*, Article 115. https://doi.org/10.1186/s12874-022-01564-2

Stark, M., & Zapf, A. (2020). Sample size calculation and re-estimation based on the prevalence in a single-arm confirmatory diagnostic accuracy study. *Statistical Methods in Medical Research*, *29*(10), 2958-2971. https://doi.org/10.1177/0962280220913588

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine*, *17*(8), 891-908. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<891::AID-SIM780>3.0.CO;2-B

Todd, S. (2007). A 25-year review of sequential methodology in clinical studies. *Statistics in Medicine*, *26*(2), 237-252. https://doi.org/10.1002/sim.2763

U.S. Food and Drug Administration (FDA). (2007). *Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests - Guidance for Industry and FDA Staf*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/statistical-guidance-reporting-results-studies-evaluating-diagnostic-tests-guidance-industry-and-fda

U.S. Food and Drug Administration (FDA). (2018). *Adaptive Designs for Clinical Trials of Drugs and Biologics*. https://www.fda.gov/media/78495/download

Wald, A. (2014). *Sequential analysis*. Dover Publications.

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

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials* (2nd ed.). John Wiley & Sons.

Wittes, J., Schabenberger, O., Zucker, D., Brittain, E., & Proschan, M. (1999). Internal pilot studies I: type I error rate of the naive t-test. *Statistics in Medicine*, *18*(24), 3481-3491. https://doi.org/10.1002/(sici)1097-0258(19991230)18:24<3481::aid-sim301>3.0.co;2-c

Wu, C., Liu, A., & Yu, K. F. (2008). An Adaptive Approach to Designing Comparative Diagnostic Accuracy Studies. *Journal of Biopharmaceutical Statistics*, *18*(1), 116-125. https://doi.org/10.1080/10543400701668282

Zapf, A., Stark, M., Gerke, O., Ehret, C., Benda, N., Bossuyt, P., Deeks, J., Reitsma, J., Alonzo, T., & Friede, T. (2020). Adaptive trial designs in diagnostic accuracy research. *Statistics in Medicine*, *39*(5), 591-601. https://doi.org/10.1002/sim.8430

Zhou, X.-H., McClish, D. K., & Obuchowski, N. A. (2011). *Statistical Methods in Diagnostic Medicine* (2nd ed.). John Wiley & Sons.

Zucker, D. M., Wittes, J. T., Schabenberger, O., & Brittain, E. (1999). Internal pilot studies II: comparison of various procedures. *Statistics in Medicine*, *18*(24), 3493-3509. https://doi.org/10.1002/(sici)1097-0258(19991230)18:24<3493::aid-sim302>3.0.co;2-2

# 2 Thesis Articles

## 2.1 Thesis Article 1

Zapf, A., **Stark, M.**, Gerke, O., Ehret, C., Benda, N., Bossuyt, P., Deeks, J., Reitsma, J., Alonzo, T., & Friede, T. (2020). Adaptive trial designs in diagnostic accuracy research. *Statistics in Medicine, 39*(5), 591-601. https://doi.org/10.1002/sim.8430

Received: 10 February 2019 | Revised: 23 October 2019 | Accepted: 26 October 2019

DOI: 10.1002/sim.8430

RESEARCH ARTICLE

WILEY Statistics in Medicine

# Adaptive trial designs in diagnostic accuracy research

**Antonia Zapf**[1] | **Maria Stark**[1] | **Oke Gerke**[2] | **Christoph Ehret**[3] | **Norbert Benda**[4,5] | **Patrick Bossuyt**[6] | **Jon Deeks**[7,8] | **Johannes Reitsma**[9] | **Todd Alonzo**[10] | **Tim Friede**[5]

[1]Department of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[2]Department of Nuclear Medicine, Odense University Hospital, Odense, Denmark

[3]Roche Diagnostics GmbH, Penzberg, Germany

[4]Federal Institute for Drugs and Medical Devices (BfArM), Bonn, Germany

[5]Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

[6]Department of Clinical Epidemiology and Biostatistics, University of Amsterdam, Amsterdam, The Netherlands

[7]Institute of Applied Health Research, University of Birmingham, Birmingham, UK

[8]NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Trust and the University of Birmingham, Birmingham, UK

[9]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht & University Utrecht, Utrecht, The Netherlands

[10]Keck School of Medicine, University of Southern California, Los Angeles, California

**Correspondence**
Antonia Zapf, Department of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, 20251 Hamburg, Germany.
Email: a.zapf@uke.de

**Funding information**
Deutsche Forschungsgemeinschaft, Grant/Award Number: ZA 687/1-1

The aim of diagnostic accuracy studies is to evaluate how accurately a diagnostic test can distinguish diseased from nondiseased individuals. Depending on the research question, different study designs and accuracy measures are appropriate. As the prior knowledge in the planning phase is often very limited, modifications of design aspects such as the sample size during the ongoing trial could increase the efficiency of diagnostic trials. In intervention studies, group sequential and adaptive designs are well established. Such designs are characterized by preplanned interim analyses, giving the opportunity to stop early for efficacy or futility or to modify elements of the study design. In contrast, in diagnostic accuracy studies, such flexible designs are less common, even if they are as important as for intervention studies. However, diagnostic accuracy studies have specific features, which may require adaptations of the statistical methods or may lead to specific advantages or limitations of sequential and adaptive designs. In this article, we summarize the current status of methodological research and applications of flexible designs in diagnostic accuracy research. Furthermore, we indicate and advocate future development of adaptive design methodology and their use in diagnostic accuracy trials from an interdisciplinary viewpoint. The term "interdisciplinary viewpoint" describes the collaboration of experts of the academic and nonacademic research.

**KEYWORDS**

adaptive designs, diagnostic accuracy, diagnostic studies, group sequential designs, sample size reestimation

## 1 | INTRODUCTION

Diagnostic tests undergo an development program including several phases.[1] Lijmer et al systematically reviewed 19 schemes for phased evaluations of medical tests and concluded that evaluations of technical efficacy, diagnostic accuracy, clinical performance, therapeutic efficacy, patient outcome, and societal aspects were common phases.[2,3] By diagnostic test, we mean any form of medical testing for diagnostic purposes, for example an entity derived from a sample (also sometimes referred to as biomarker) or an application of a diagnostic modality (eg, a maximum standard uptake value in positron emission tomography/computed tomography). Technical efficacy covers the technical aspects of a diagnostic test that are evaluated in the first phase.[2] These aspects comprise the applicability and the equipment, and the term clinical performance describes how useful the diagnostic test is to deduce the its desired diagnosis.[3] Hence, with a supportive result, the clinician is able to make a more informed diagnosis than without the diagnostic test. In this article, we focus on diagnostic accuracy studies, which aim at assessing how reliable a diagnostic test identifies specific subgroups, eg, diseased and nondiseased. Depending on the research question, different study designs, for instance, single-arm or parallel-arm designs, and accuracy measures are appropriate.

Both sequential trial methodology and adaptive designs have been used for several decades in intervention studies.[4-12] The "reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design"[13] defines the terms "group sequential design" and "adaptive design" and therefore makes apparent that group sequential designs fall within the class of adaptive designs: Sequential trials are hallmarked by preplanned interim analyses at which cumulating data are assessed with respect to early stopping for efficacy or futility with control of the overall type I error probability at a specified level. Adaptive clinical trial designs are characterized by preplanned interim analyses, at which planned modifications of the study design based on accumulating study data (or any other information available at the time of any interim analysis) are possible without undermining the trial's integrity and validity.[14] In the remainder, we refer to these flexible designs as adaptive designs with the understanding that these include (group) sequential designs. It should be noted that monitoring of safety data by a data monitoring committee or Data safety monitoring board is usually not part of this process but a separate issue, although some authors have advocated the use of similar stopping boundaries.[15]

On the one hand, adaptive designs are less common in diagnostic research. On the other hand, such flexible designs are just as important for diagnostic accuracy studies as they are for intervention studies to increase efficiency. However, diagnostic accuracy studies have specific features, which may require modifications of the statistical methods or lead to advantages or limitations for the application of adaptive designs.[16] For example, in general, the time between inclusion and completion of the study is very short for the individual participant. Therefore, the aim of this position paper is twofold: (1) to summarize the current status of methodology research and the use of adaptive designs in diagnostic accuracy research and (2) to advocate future development and use of adaptive designs in diagnostic accuracy trials by highlighting the characteristics of diagnostic research.

This work evolved from a workshop on flexible designs for diagnostic studies held in Göttingen, Germany, November 6–7, 2017.[17] The paper is structured as follows: First, the design aspects and the measures of diagnostic accuracy studies are described (see Section 2). Thereafter, in Section 3, the two main adaptive design types, namely combination tests and the conditional error function approach, are briefly described. In the main part (Section 4), methodological research and practical perspectives of adaptive designs for diagnostic accuracy studies are outlined. In Section 5, trial steering and data monitoring committees and their respective roles in trial conduct are summarized. A discussion in Section 6 closes the paper.

## 2 | DIAGNOSTIC ACCURACY STUDIES – DESIGN ASPECTS AND MEASURES

In early diagnostic trials, the disease status is often known in advance, determined by the reference standard, leading to a case-control study design. Diagnostic case-control designs may be applied with fairly balanced sample sizes of diseased and nondiseased in order to gather as much information on sensitivity as on the specificity, even though a prevalence of about 50% might not mirror the true prevalence in the target population. The aim in these studies is mainly to obtain a rough estimate of the overall diagnostic accuracy and to define a positivity threshold. In contrast, in confirmatory diagnostic trials, the disease status is often determined simultaneously to the diagnostic test(s) investigated, leading to a cohort study design. In these studies, a consecutive recruitment within a given time frame is recommended to obtain a representative sample regarding the prevalence; then, the ratio of diseased to diseased and nondiseased reflects the prevalence in the study population. The aim of confirmatory accuracy studies is to obtain a reliable estimate of the diagnostic accuracy at a specific threshold.

Another important design aspect is whether an experimental diagnostic test is compared with the reference standard only, or whether two or more diagnostic tests under evaluation are compared with each other (based on their comparison with the reference standard). In both scenarios estimation of a test's diagnostic accuracy requires knowledge, for each patient, of the true disease state (defined by the reference standard) and of the results of the diagnostic tests.

The third important design aspect is only valid for studies comparing two or more tests. The standard design, which is also recommended by the EMA guideline, is the within-subject design (all tests under evaluation in all patients, also called paired design).[1] However, if it is not feasible or ethically justifiable, the diagnostic tests under evaluation will be applied in independent groups, preferentially using a randomized allocation procedure. Furthermore, it can be appropriate to include two or more readers, which leads to a two-factorial design and entails observer agreement assessment.[18]

The choice of the accuracy measure depends on whether it is an early or a confirmatory diagnostic accuracy study. Early diagnostic accuracy trials may focus on overall estimates of diagnostic accuracy of tests on a continuous or ordinal scale, without defining a positivity threshold and considering sensitivity (true positive rate) and specificity (true negative rate) jointly. The standard approach for this scenario is to estimate a receiver operating characteristic (ROC) curve, which displays sensitivity versus 1 minus specificity for every possible cutoff value.[19-21] The area under the curve (AUC) is a measure for the overall diagnostic accuracy. More precisely, the AUC is "the probability that, when presented with a randomly chosen patient with disease and a randomly chosen patient without disease, the results of the diagnostic test will rank the patient with disease as having higher suspicion for disease than the patient without disease."[19] In general, the AUC is equal to 0.5 for a test as useful as flipping a coin and equal to 1 for a perfect test. Sometimes, not the whole AUC is used but a partial area (pAUC) for a specific minimum sensitivity or specificity. However, as the methodology is the same as for the whole AUC, we focus here on assessing the AUC rather than pAUC.

If the optimal cut point has already been determined or if the result of a diagnostic test is actually dichotomous, sensitivity and specificity will be both considered primary endpoints in confirmatory accuracy trials. Since both measures represent important characteristics of a diagnostic test, they are recommended to be used on equal footing as coprimary endpoints by the European Medicines Agency (EMA), evaluated separately.[1]

The positive predictive value (PPV) and the negative predictive value (NPV) as probabilities for a correct test result among the positive or negative test results, can be included as key-secondary endpoints. A reliable estimation of the predictive values requires either a representative sample (of the population in which the diagnostic test is intended to be applied) or the imputation of a known prevalence (for a given population) and the estimated sensitivity and specificity into the Bayes' formula.

Regarding the statistical hypotheses, it is necessary to distinguish between single test studies and studies for the comparison of two or more tests. If in single test studies the aim is not only to estimate the accuracy but to assess whether the accuracy meets some predefined required values, hypotheses have to be formulated accordingly. For the comparison of two or more diagnostic tests, hypothesis tests may be of interest to assess whether the performance of one test exceeds that of the other(s). The hypotheses can be formulated for each of the accuracy measures. However, for sensitivity and specificity as coprimary endpoints, the global null hypothesis can only be rejected if both hypotheses (regarding sensitivity and specificity) are rejected. If two tests are compared, ideally superiority in both sensitivity and specificity is achieved. However, this is often unrealistic. Hence, noninferiority is usually required in one coprimary endpoint and superiority in the other. In general, the hypotheses are tested using confidence intervals, and p-values are rarely used. For the comparison of two tests, confidence intervals for the differences (or ratios) of the accuracy measures are important for a meaningful interpretation.[22,23]

All different study designs described above and all mentioned accuracy measures are considered in this article. In some special cases, other endpoints could be appropriate, eg, positive and negative percent agreement (when no reference standard is available) or diagnostic odds ratios. However, this will not be covered in this article.

## 3 | STATISTICAL METHODS FOR ADAPTIVE DESIGNS IN INTERVENTION STUDIES

As mentioned in Section 1, adaptive designs are well established in intervention studies. This approach uses information from preplanned interim analyses to either decide to stop the trial early for efficacy or futility or, more generally, to modify design aspects. Interim analyses can be performed in a fully blinded or in an unblinded manner. Blinded interim analyses are based on data pooled across treatments. As this could also be done in open trials, the latest FDA guidance on adaptive designs refers to these as adaptations based on noncomparative data.[14] For instance, there is a wide range of sample size

reestimation procedures based on noncomparative data for various types of endpoints available.[24] More recently, there has been some interest in blinded continuous monitoring procedures which result in smaller variability of the final sample size compared to designs with only a single reestimation.[25,26] Interim analyses based on unblinded data may include formal statistical hypothesis testing. Commonly applied adaptations include sample size adjustments, treatment (or dose) selection, and subgroup selection or enrichment (study eligibility criteria). Thereby, adaptive trial designs can result in more efficient clinical studies and the chance of success may be increased. For group sequential designs, we refer here to the literature.[4,27] In the following, we briefly introduce combination tests and conditional error functions as these can be used to construct very flexible designs. We also outline how to deal with multiple hypotheses in so-called adaptive seamless designs.

Combination tests combine the p-values based on data from different stages of a trial. To control the type I error rate, the so-called p-clud condition must be fulfilled.[28] This is, for instance, the case when the data of the stages come from independent samples, and hypothesis tests are used that result under the null hypothesis in p-values uniformly distributed on the interval $[0, 1]$.[12] A combination test is specified by its combination function and its boundaries for early termination of the study. Early stopping is recommended if the p-value of the interim stage is smaller than the lower boundary or larger than the upper boundary. In the first case, the null hypothesis can be rejected. In the second case, the null hypothesis is not rejected, and the study is stopped due to futility. When the study is supposed to continue until the final stage, the null hypothesis will be rejected in the final analysis if the value returned by the combination function is smaller than or equal to the critical value.

An early proposal of a combination test is the Fisher's product test in which the p-values are multiplied with each other. The weighted Fisher's product test performs a weighted multiplicative combination of the p-values of each trial stage. Its usage is recommended if the sample sizes of the different stages are unequal. Hereby, stages with larger sample sizes obtain a higher weight than those with a smaller sample size. The inverse normal combination test is based on a weighted inverse normal combination function whereby the weights are again chosen according to the planned sample sizes of the different stages.[29] The inverse normal method is equivalent to an extension of group sequential tests by decomposing the test statistic as a weighted sum of the stagewise statistics with preplanned weights.[30]

The conditional error function approach represents a further approach to define the rejection area in an adaptive design.[12] One early form is the proposal by Proschan and Hunsberger for effect-based sample size reestimation. Analogous to the combination tests, the conditional error function approach is defined by the lower and upper boundaries of the rejection region and the conditional error function. The conditional error function returns the conditional type I error rate given the data of the first stage. Hence, the overall type I error rate is the probability to reject the null hypothesis at the first stage plus the expected value of the conditional error function in the interval between the lower and the upper boundaries of the rejection region.[31]

So far, we have only considered the situation of a single hypothesis. In adaptive seamless designs combining aspects of different development phases such as learning about the optimal dose or population with confirmatory testing, multiple hypotheses are considered. Control of the familywise type I error probability in the strong sense can be achieved by, eg, using combination tests on intersection hypotheses in a closed test procedure.[32-34] Considering adaptive designs for treatment or subgroup selection, the methods and aspects of their implementation have been comprehensively described (eg, simulation models and software) in a forthcoming manuscript.[34]

The combination test principle as well as the conditional error function approach can also be transferred from p-values to confidence intervals (see for example the work of Magirr et al[35] and Brannath et al[36]).

## 4 | ADAPTIVE DESIGNS FOR DIAGNOSTIC ACCURACY STUDIES

In Section 3, we mentioned that interim analyses could be performed in a blinded or in an unblinded manner. In the context of intervention studies for the comparison of two drugs, blinded interim analyses, in which treatment groups are not identified, ensure full integrity of the trial. In diagnostic studies, the connection of the results of the diagnostic test(s) with the outcomes of the reference standard may be blinded. For example, the prevalence can be estimated in a blinded manner by only using the results from the reference standard. In a diagnostic trial comparing two diagnostic tests, a blinded interim analysis could be achieved by summarizing the test results for a given reference standard (diseased or nondiseased) pooling the results of both diagnostic tests.

In sequential intervention trials, stopping for futility or efficacy can lead to reduced costs and trial duration/development time and save further study participants from harm or provide the benefit of the new therapy earlier to

patients outside the trial. In contrast, the results of experimental tests are typically not used to inform the care of participants in the study, so there is no additional risk of harm. Accordingly, the ethical imperative to halt a study at the earliest time to avoid harming patients is weak unless the experimental tests have direct negative consequences themselves. However, the advantages of completing a successful trial early remain.

The need for adaptive designs in AUC studies results from the fact that prior knowledge in the planning phase is often very limited. Pepe et al[37] described standards of study designs in pivotal diagnostic accuracy studies and mentioned planning for early termination, if appropriate. Modifications of design aspects can be motivated by the interim results or by external reasons; examples are adaptation of the reference standard or the eligibility criteria due to slow recruitment. If sample size reestimation is performed based on the results of the interim analysis, it can, for example, impact the point estimate of the diagnostic accuracy, the variability of the test results, the correlation between the results of the individual diagnostic tests (in the paired design), or the proportion of missing values.

The aim of adaptive designs for diagnostic trials with sensitivity and specificity as coprimary endpoints can be the reestimation of different parameters. Probably, the simplest case is the blinded reestimation of the prevalence, which requires adaptation of the overall sample size (in general in case of an overestimated prevalence), which will not affect significance testing for sensitivity, specificity, and AUC, but may do for predictive values. In contrast, for the reestimation of sensitivity and specificity, an unblinded interim analysis is needed. Furthermore, the reestimation of the proportion of discordant results between several diagnostic tests can be of interest; to this end, the correlation between these two diagnostic tests can be reevaluated. With the reestimation of these parameters, the sample size of the individual status groups can be adapted during the study. If deemed necessary, even adjustments to the reference standard are possible, for example, by changing individual components of a multicomponent reference standard. Another important issue is a possible modification of the positivity threshold (of the experimental test and/or of the reference standard) during the trial, which might be possible within adaptive seamless designs.

## 4.1 | Methodological research

### 4.1.1 | Adaptive designs for the AUC

Regarding group sequential designs in AUC studies without sample size reestimation or other modifications, there are several articles (for an overview see for example[38-40]). To our knowledge, the first article about group sequential designs in diagnostic research was written by Mazumdar and Liu, in which the authors propose an approach based on a binormal distribution (transferable to other distributions or nonparametric models) for the comparison of two AUCs.[41]

The implications of group sequential designs for comparative diagnostic accuracy trials and resulting guidelines for practitioners were presented by Mazumdar, here with the O'Brien-Fleming stopping boundaries.[42] Zhou et al presented a nonparametric group sequential design for the comparison of two AUCs in the paired design, based on the Brownian motion.[43] Tang et al proposed two group sequential designs for paired data: a nonparametric approach using a nonparametric family of weighted AUC statistics, and a semiparametric approach based on a proportional hazards model.[44] Liu et al also used a nonparametric approach, but in a more general sense for a single AUC and the comparison of two or more AUCs in the paired design, but also for independent groups.[45] Another nonparametric approach is the sequential conditional probability ratio test procedure for the comparison of two AUCs.[46] Koopmeiners and Feng derived the asymptotic properties of the sequential empirical ROC curve for case-control studies.[47] To identify the optimal design (stopping for efficacy only, for futility only, or for both) Kaizer et al suggested a loss function as decision criterion for two-stage biomarker validation studies.[48]

Regarding adaptive designs with sample size reestimation, the reestimation can be performed based on nuisance parameters without the need to adjust for the type I error.[49,50] In contrast, Tang and Liu proposed a nonparametric approach for sample size reestimation based on the estimated difference between two paired AUCs in a group sequential design with an error-spending function.[51] Brinton et al also used the idea of an internal pilot study to correct the sample size for the true disease prevalence and variance with a control of the type I error rate.[52]

### 4.1.2 | Adaptive designs for other accuracy measures

For the comparison of ROC curves, instead of AUCs, Ye and Tang derived asymptotic properties of the sequential differences of two empirical ROC curves at the process level.[40] Dong et al addressed the optimal sampling ratio including adaptations.[53]

**TABLE 1** Overview of flexible designs for the area under the curve (AUC) and other accuracy measures

| Method | Type of analysis | Design | Approach |
|---|---|---|---|
| **Group sequential** | **ROC curve** | Paired difference[36] | Parametric[36] |
| | **AUC** | Single group[41] | Parametric[37,38] |
| | | Paired difference[37-42,44] | Semiparametric[40] |
| | | Unpaired difference[41] | Nonparametric[37-42,44] |
| | **Sensitivity, specificity** | Single group[49,50] | Parametric[49,50] |
| | | | Nonparametric[50] |
| | **PPV, NPV** | Single group[52-54] | Parametric[52,53] |
| | | | Nonparametric[54] |
| **Adaptation of sample size** | **AUC** | Paired difference[41-43,45-48] | Parametric[45,47] |
| | | | Nonparametric[46,48] |
| | **Sensitivity, specificity** | Paired difference[51] | Parametric[51] |

Abbreviations: NPV, negative predictive value; PPV, positive predictive value; ROC, receiver operating characteristic.

Only a few studies were identified that dealt with adaptive designs in diagnostic trials, considering sensitivity and specificity as coprimary endpoints. Shu et al[54] proposed different group sequential designs to early terminate a diagnostic phase 2 trial if both the sensitivity and specificity are either good enough or below a minimally acceptable margin. Pepe et al[55] proposed a group sequential design for a diagnostic phase 2 or phase 3 biomarker study with the possibility to adjust for bias that is caused by early stopping. One method for sample size recalculation in a paired diagnostic study with sensitivity and specificity as coprimary endpoints was presented by McCray et al.[56] They reestimated the proportion of concordant test results via maximum likelihood estimation.

There exists some literature using group sequential designs to reevaluate the PPV and the NPV of a diagnostic test. Koopmeiners and Feng introduced a group sequential design in a diagnostic biomarker study by deriving the asymptotic results of the PPV and NPV curves.[47] Koopmeiners et al[57] used this group sequential design to decide about an early termination of a continuous diagnostic biomarker trial due to futility. In the case of an unknown prevalence, Koopmeiners and Feng[58] as well as Tayob et al[59] developed group sequential designs which can be used for the unbiased estimation of the PPV and NPV. See Table 1 for an overview of abovementioned adaptive designs for the AUC and further accuracy measures.

## 4.2 | Practical perspectives

Adaptive designs are rarely utilized for clinical trials in diagnostic research. Short enrollment periods, moderate savings in time or costs due to early stopping for success, and increased logistical complexity for executing interim analyses could be reasons why conventional fixed designs are traditionally implemented in diagnostic clinical trials instead.

Some examples of diagnostic accuracy studies using adaptive designs were identified. Shivakumar et al[60] reported the results of an interim analysis for the diagnosis of psychological distress in elderly seeking health care, without discussion of possible biases and type I error rate inflation. Snijder et al[61] presented the results of an interim analysis of a study about image-based ex vivo drug screening for patients with aggressive hematological malignancies. The authors did not mention a group sequential design or adjustment of the type I error. Ghaneh et al[62] applied an adaptive design for sample size reestimation based on the correlation between the test errors (false positives and false negatives) in a multicenter, prospective diagnostic accuracy study for the diagnosis of pancreatic cancer.

Nevertheless, as already discussed, adaptive designs in diagnostic accuracy trials may be beneficial. In the following, an early stop for futility is presented as one potential application of adaptive trial methodology.

To illustrate, the following scenario is considered. For the approval of an assay, a confirmatory study is needed to demonstrate that sensitivity fulfills a predefined acceptance criterion, ie, the lower limit of a two-sided 95% confidence interval (LLCI) is at least 90%, for a point estimate of 96%. Budget constraints prohibit the conduct of a pilot study in a clinical setting, leading to uncertainty around the assay's clinical performance. The disease prevalence is low (eg 10%), consequently subject recruitment can extend over the course of several years. Sample acquisition costs are high; therefore, an economical approach to meeting the study objectives is essential.

For this scenario, it is unlikely to reach a stop for success as the binomial distribution does not allow the effect size to be much larger than the expected 96%. Additionally, the experimentwise type I error probability needs to be controlled at level $\alpha$. Therefore, the required sample size to attain a LLCI above 90% is close to the final sample size if the significance

level α is evenly distributed between an interim and final analysis (97.5% confidence intervals each, Bonferroni correction). An α-spending function could be used to distribute the type I error more efficiently for this scenario (for example implemented in the R package gsdesign by Anderson[63]), but for reasons of simplification, we use here the Bonferroni correction. Furthermore, the optimal timing of the interim analysis is not necessarily at half-time. However, the simplified numerical example, with the specifications above and ignoring the random nature of the point estimate looks as follows for three design considerations:

- *Conventional fixed design*. For a single cohort fixed design with an α of 5% (two-sided), a minimum of 100 true positive cases is required to allow for a maximum of four false negatives so that the LLCI is at least 90% with a point estimate of 96%. Assuming a disease prevalence of 10%, a total sample size (diseased and nondiseased) of $N = 1000$ is necessary.
- *Adaptive design stopping early for success*. Using the Bonferroni correction, ie, α is evenly distributed between an interim and final analysis, a sample size of 79 true positive cases is needed to allow for two false negatives with the minimum LLCI of 90% (point estimate: 97.5%). However, if the performance of the assay is as expected at the interim analysis, the study cannot be stopped for success as the sample size is not large enough to attain a LLCI of at least 90%. As a result, a sample size of 110 for true positive cases is necessary for the final analysis to ensure a minimum LLCI of 90%. The total sample size (diseased and nondiseased) for the study including an interim analysis would be $N = 1100$, meaning 10% larger than for a conventional trial without an interim analysis.
- *Adaptive design stopping early for futility*. If only an early stop for futility is planned, it is not necessary to adjust the type I error for the interim analysis at 50% of the recruitment. The full $\alpha = 5\%$ could be used for the final analysis. If the number of false negatives is already 5 or more at 50% of the recruitment, the study could be terminated for futility, and costs for the recruitment of the remaining 50% of patients can be saved.

This example illustrates possible applications and corresponding implications of adaptive designs in diagnostic accuracy studies.

## 5 | DATA MONITORING COMMITTEE

Clinical trials can have a trial steering committee (TSC) and a data monitoring committee (DMC) or data monitoring and safety board. For more information about DMC, we refer to the relevant guidelines.[64-66] This does not only apply to intervention studies but also to diagnostic trials with patient-relevant outcomes, where the new diagnostic test may lead to an altered therapy or has any other consequences for the participants. In contrast, in diagnostic accuracy studies, such committees are not standard.

For fixed study designs and adaptive designs with blinded interim analysis, it may be appropriate and more efficient to combine the TSC and the DMC into an oversight committee (OC). An OC should be established if at least one of the following issues is present: (1) reasonable safety concerns, (2) considerable uncertainty about the assumptions for the sample size calculation, (3) the chance of external findings influencing the current study, or (4) resource intensive (in terms of budget and/or time). An OC should involve responsible members of the study group as well as independent members. The task of all OC members regarding adaptations should be to monitor, for example, if the new diagnostic tests lead to an obvious and unreasonable harm for the patients. Furthermore, all OC members would be involved in blinded interim analyses (in conducting the analyses or in discussing the results) and provide recommendations about next steps. The next steps could be stopping for futility, sample size reestimation, or other adaptations.

All OC members can also be involved in monitoring the recruitment rate. With regard to adaptive designs with unblinded interim analysis, the DMC should be established as an independent committee, because unblinded interim analyses leading to sample size reestimation or other adaptations have to be performed independently of the TSC. As a result, recommendations to the sponsor about continuing or stopping the trial (for efficacy or futility) should be made by the DMC.

## 6 | DISCUSSION

The evaluation of diagnostic accuracy with its inherent need for a reference standard is a characteristic phase for any diagnostic test. As such, dedicated research into how to apply adaptive trial methodology to diagnostic trials is necessary. Currently, group sequential techniques with the main purpose of sample size reassessment or possibly early termination

of the trial are applied, but not routinely. Furthermore, only few reports on interim analyses without discussion of type I error rate inflation were found. We strongly believe that the field of diagnostic research could be significantly advanced by the more frequent implementation of adaptive designs, in particular, with options for early stopping (due to efficacy or futility) or design modifications such as sample size reassessment. In our view, these areas are two promising fields for future methodological research. For this purpose, the development of further techniques for adaptive designs in diagnostic accuracy trials is necessary.

This paper is a timely status of the research in adaptive trial methodology based on an ad hoc literature search in PubMed/Medline and Google Scholar performed by three authors (AZ, MS, and OG) in July 2018. The literature search was not performed systematically. A strength of the project is that coauthors from The Netherlands, UK, Germany, Denmark, and US, working in diagnostic research in academia, a government agency, and industry contributed.

This study is, to the best of our knowledge, the first to summarize the current status of the topic from a diagnostic research point of view and to indicate potential future research subjects from an interdisciplinary standpoint. An interdisciplinary viewpoint describes the collaboration of experts of academic and nonacademic research areas, which helps to reveal different requirements adaptive designs in diagnostic trials must maintain.

One may think that adaptive trial methodology can be transferred to diagnostic research because it has been established and used for decades now.[7,11] To some extent, this may be true, especially when thinking of late phase diagnostic trials establishing patient benefit, thereby requiring randomized designs. However, diagnostic accuracy research is peculiar and prevents a simple application of preexisting techniques.

- A reference standard is a prerequisite for any diagnostic accuracy study.
- The primary outcome is twofold (sensitivity and specificity), implying an important role on the prevalence with respect to the achievable accuracy in parameter estimation.
- Diagnostic accuracy trials are often planned and conducted in a within-subject (or paired) design, thereby shifting the focus on discordant pairs of results and their (dis)agreement.
- Keeping the blind regards the interim analysis, meaning the results of diagnostic test(s) and reference standard, not randomization information with respect to study arms (as is the case in interventional research).

Tang et al argued that the use of adaptive designs in diagnostic accuracy studies is an obvious option since they are conducted so fast.[44] Hence, the speed of recruitment determines the applicability of a group sequential design (or in fact any other adaptive design): the longer the length of trial recruitment, the more realistic the application of a group-sequential design becomes.

In case of early termination of the study, risks and consequences of interim analyses with respect to possible bias need to be taken into consideration.[67-71] Our study stresses the need for continuing research into possible applications of adaptive designs in diagnostic accuracy research. Recent endeavors concerning late phase diagnostic trials on patient benefit, which are beyond the focus of this study, dealt with multiplicity issues in exploratory subgroup analysis, including adaptive biomarker-driven designs[72] and specified the application of an enrichment design comparing a new endovascular treatment with standard of care for ischemic stroke patients.[73] The following questions might be subject to future research:

- How can adaptive designs be applied with possible early stopping due to efficacy or futility as well as seamless designs?
- Can adaptive designs for the reestimation of PPV and NPV be transferred to the reestimation of the sensitivity and specificity?
- What is the optimal time point for an interim analysis—as early as possible, or as late as necessary? First interim analysis with 40%, 50%, or 60% of patients? This issue depends on the duration of evaluation , eg, histopathological examination of tissue following a diagnostic test or follow-up of at least 6 months as part of a composite reference standard.

Furthermore, this paper is limited to adaptive designs in diagnostic accuracy research, as these areas concern the very characterization of a diagnostic test; diagnostic thinking efficacy and therapeutic efficacy focus, in opposition, on clinical endpoints or surrogates of those for patient benefit, which, in turn, are investigated in patient outcome research later in the process. Adaptive designs for such studies are also subject to future research.

## ACKNOWLEDGEMENT

This work was supported by the Deutsche Forschungsgemeinschaft under grant ZA 687/1-1.

## CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## ORCID

*Antonia Zapf* https://orcid.org/0000-0001-5339-2472
*Tim Friede* https://orcid.org/0000-0001-5347-7441

## REFERENCES

1. European Medicines Agency. Guideline on clinical evaluation of diagnostic agents. 2009. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003580.pdf. Accessed July 24, 2018.
2. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. *Med Decis Making*. 2009;29(5):E13-E21.
3. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0746&from=DE. Accessed September 27, 2019.
4. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. 2nd ed. Chichester, UK: Wiley; 1997.
5. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC; 2000.
6. Moyé LA. *Statistical Monitoring of Clinical Trials: Fundamentals for Investigators*. New York, NY: Springer; 2006.
7. Todd S. A 25-year review of sequential methodology in clinical studies. *Statist Med*. 2007;26(2):237-252.
8. European Medicines Agency. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. 2007. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003616.pdf. Accessed July 24, 2018.
9. Chow SC, Chang M. *Adaptive Design Methods in Clinical Trials*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC; 2011.
10. Wald A. *Sequential Analysis*. New Impression ed. Mineola, NY: Dover Publications; 2014.
11. Bauer P, Bretz F, Dragalin V, König F, Wassmer G. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statist Med*. 2016;35(3):325-347.
12. Wassmer G, Brannath W. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. New York, NY: Springer; 2016.
13. Committee for Medicinal Products for Human Use. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. 2007. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf. Accessed September 27, 2019.
14. US Food and Drug Administration. Adaptive Design Clinical Trials for Drugs and Biologics (Draft Guidance). 2018. https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf. Accessed January 20, 2019.
15. Zhu L, Yao B, Xia HA, Jiang Q. Statistical monitoring of safety in clinical trials. *Stat Biopharm Res*. 2016;8(1):88-105.
16. Gerke O, Høilund-Carlsen PF, Poulsen MH, Vach W. Interim analysis in diagnostic versus treatment studies: differences and similarities. *Am J Nucl Med Mol Imaging*. 2012;2(3):344-352.
17. Workshop on Flexible Designs for Diagnostic Studies – From Diagnostic Accuracy to Personalized Medicine. 2017. http://www.ams.med.uni-goettingen.de/p-mgmt/Flexpn.html. Accessed July 26, 2018.
18. Gerke O, Vilstrup MH, Segtnan EA, Halekoh U, Høilund-Carlsen PF. How to assess intra- and inter-observer agreement with quantitative PET using variance component analysis: a proposal for standardisation. *BMC Med Imaging*. 2016;16(1):54.
19. Obuchowski NA. ROC analysis. *AJR Am J Roentgenol*. 2005;184(2):364-372.
20. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. 2nd ed. Hoboken, NJ: Wiley; 2011.
21. Obuchowski NA, Bullen JA. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys Med Biol*. 2018;63(7):07TR01.
22. Wenzel D, Zapf A. Difference of two dependent sensitivities and specificities: comparison of various approaches. *Biom J*. 2013;55(5):705-718.
23. Gerke O, Vach W, Høilund-Carlsen PF. PET/CT in cancer: methodological considerations for comparative diagnostic phase II studies with paired binary data. *Methods Inf Med*. 2008;47(6):470-479.
24. Friede T, Kieser M. Sample size recalculation in internal pilot study designs: a review. *Biom J*. 2006;48:537-555.
25. Friede T, Miller F. Blinded continuous monitoring of nuisance parameters in clinical trials. *J Royal Stat Soc Ser C*. 2012;61:601-618.
26. Friede T, Häring DA, Schmidli H. Blinded continuous monitoring in clinical trials with recurrent event endpoints. *Pharmaceutical Statistics*. 2019. In press.
27. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC; 2000.
28. Brannath W, Posch M, Bauer P. Recursive combination tests. *JASA*. 2002;97(457):236-244.
29. Chang M. *Adaptive Design Theory and Implementation Using SAS and R*. Boca Raton, FL: Chapman & Hall/CRC; 2014.

30. Cui L, Hung HM, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics*. 1999;55(3):853-857.
31. Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*. 2001;57(3):886-891.
32. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statist Med*. 1999;18(14):1833-1848.
33. Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. *Biom J*. 2006;48(4):623-634.
34. Friede T, Stallard N, Parsons N Seamless phase II/III clinical trials using early outcomes for treatment or subgroup selection: methods and aspects of their implementation. arXiv preprint arXiv:1901.08365. 2019.
35. Magirr D, Jaki T, Posch M, Klinglmueller F. Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. *Biometrika*. 2013;100(4):985-996.
36. Brannath W, Mehta CR, Posch M. Exact confidence bounds following adaptive group sequential tests. *Biometrics*. 2009;65(2):539-546.
37. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J Natl Cancer Inst*. 2008;100(20):1432-1438.
38. Dong T, Tang L. Sequential diagnostic trial designs. *Wiley Interdiscip Rev Comput Stat*. 2011;3:79-83.
39. Zou KH, Liu A, Bandos AI, Ohno-Machado L, Rockette HE. *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*. Boca Raton, FL: Chapman & Hall/CRC; 2012.
40. Ye X, Tang LL. Group sequential methods for comparing correlated receiver operating characteristic curves. In: Chen Z, Liu A, Qu Y, Tang L, Ting N, Tsong Y, eds. *Applied Statistics in Biomedicine and Clinical Trials Design*. New York, NY: Springer; 2015.
41. Mazumdar M, Liu A. Group sequential design for comparative diagnostic accuracy studies. *Statist Med*. 2003;22(5):727-739.
42. Mazumdar M. Group sequential design for comparative diagnostic accuracy studies: implications and guidelines for practitioners. *Med Decis Making*. 2004;24(5):525-533.
43. Zhou XH, Li SM, Gatsonis CA. Wilcoxon-based group sequential designs for comparison of areas under two correlated ROC curves. *Statist Med*. 2008;27(2):213-223.
44. Tang L, Emerson SS, Zhou XH. Nonparametric and semiparametric group sequential methods for comparing accuracy of diagnostic tests. *Biometrics*. 2008;64(4):1137-1145.
45. Liu A, Wu C, Schisterman EF. Nonparametric sequential evaluation of diagnostic biomarkers. *Statist Med*. 2008;27(10):1667-1678.
46. Tang L, Tan M, Zhou XH. A sequential conditional probability ratio test procedure for comparing diagnostic tests. *J Appl Stat*. 2011;38(8):1623-1632.
47. Koopmeiners JS, Feng Z. Asymptotic properties of the sequential empirical ROC, PPV and NPV curves under case-control sampling. *Ann Stat*. 2011;39(6):3234-3261.
48. Kaizer AM, Koopmeiners JS. Identifying optimal approaches to early termination in two-stage biomarker validation studies. *Appl Stat*. 2017;66:187-199.
49. Wu C, Liu A, Yu KF. An adaptive approach to designing comparative diagnostic accuracy studies. *J Biopharm Stat*. 2008;18(1):116-125.
50. Friede T, Kieser M. Blinded sample size re-estimation in superiority and noninferiority trials: bias versus variance in variance estimation. *Pharm Stat*. 2013;12(3):141-146.
51. Tang LL, Liu A. Sample size recalculation in sequential diagnostic trials. *Biostatistics*. 2010;11(1):151-163.
52. Brinton JT, Ringham BM, Glueck DH. An internal pilot design for prospective cancer screening trials with unknown disease prevalence. *Trials*. 2015;16:458.
53. Dong T, Tang LL, Rosenberger WF. Optimal sampling ratios in comparative diagnostic trials. *J Royal Stat Soc Ser C Appl Stat*. 2014;63(3):499-514.
54. Shu Y, Liu A, Li Z. Sequential evaluation of a medical diagnostic test with binary outcomes. *Statist Med*. 2007;26(24):4416-4427.
55. Pepe MS, Feng Z, Longton G, Koopmeiners J. Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility. *Statist Med*. 2009;28(5):762-779.
56. McCray GPJ, Titman AC, Ghaneh P, Lancaster GA. Sample size re-estimation in paired comparative diagnostic accuracy studies with a binary response. *BMC Med Res Methodol*. 2017;17(1):102.
57. Koopmeiners JS, Feng Z, Pepe MS. Conditional estimation after a two-stage diagnostic biomarker study that allows early termination for futility. *Statist Med*. 2012;31(5):420-435.
58. Koopmeiners JS, Feng Z. Group sequential testing of the predictive accuracy of a continuous biomarker with unknown prevalence. *Statist Med*. 2016;35(8):1267-1280.
59. Tayob N, Do KA, Feng Z. Unbiased estimation of biomarker panel performance when combining training and testing data in a group sequential design. *Biometrics*. 2016;72(3):888-896.
60. Shivakumar P, Sadanand S, Bharath S, Girish N, Philip M, Varghese M. Identifying psychological distress in elderly seeking health care. *Indian J Public Health*. 2015;59(1):18-23.
61. Snijder B, Vladimer GI, Krall N, et al. Image-based ex-vivo drug screening for patients with aggressive haematological malignancies: interim results from a single-arm, open-label, pilot study. *Lancet Haematol*. 2017;4(12):e595-e606.
62. Ghaneh P, Hanson R, Titman A, et al. PET-PANC: multicentre prospective diagnostic accuracy and health economic analysis study of the impact of combined modality 18fluorine-2-fluoro-2-deoxy-d-glucose positron emission tomography with computed tomography scanning in the diagnosis and management of pancreatic cancer. *Health Technol Assess*. 2018;22(7):1-114.
63. Anderson K. gsDesign: group sequential design. R Package Version. https://CRAN.R-project.org/package=gsDesign. 2016.

64. European Medicines Agency (EMA). Guideline on data monitoring committees. 2005. https://www.ema.europa.eu/documents/scientific-guideline/guideline-data-monitoring-committees_en.pdf. Accessed November 5, 2018.
65. US Food and Drug Administration (FDA). Guidance for Clinical Trial Sponsors - Establishment and Operation of Clinical Trial Data Monitoring Committees. 2006. https://www.fda.gov/downloads/regulatoryinformation/guidances/ucm127073.pdf. Accessed December 22, 2018.
66. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). E9 Statistical Principles for Clinical Trials. 1998. https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/
67. Wittes J. Stopping a trial early - and then what. *Clin Trials*. 2012;9(6):714-720.
68. Shu Y, Liu A, Li Z. Point and interval estimation of accuracies of a binary medical diagnostic test following group sequential testing. *Philos Trans Royal Soc A Math Phys Eng Sci*. 2008;366(1874):2335-2345.
69. Lee JW, DeMets DL. Estimation following group sequential tests with repeated measurements data. *Comput Stat Data Anal*. 1999;32:69-77.
70. Li Z, DeMets DL. On the bias of estimation of a Brownian motion drift following group sequential tests. *Statistica Sinica*. 1999;9:923-937.
71. Liu A, Hall WJ. Unbiased estimation following a group sequential test. *Biometrika*. 1999;86(1):71-78.
72. Dmitrienko A, Millen B, Lipkovich I. Multiplicity considerations in subgroup analysis. *Statist Med*. 2017;36(28):4446-4454.
73. Lai TL, Lavori PW, Tsang KW. Adaptive enrichment designs for confirmatory trials. *Statist Med*. 2018;38(4):613-624.

**How to cite this article:** Zapf A, Stark M, Gerke O, et al. Adaptive trial designs in diagnostic accuracy research. *Statistics in Medicine*. 2020;39:591–601. https://doi.org/10.1002/sim.8430

## 2.2 Thesis Article 2

### 2.2.1 Main Document

**Stark, M.**, & Zapf, A. (2020). Sample size calculation and re-estimation based on the prevalence in a single-arm confirmatory diagnostic accuracy study. *Statistical Methods in Medical Research*, *29*(10), 2958-2971. https://doi.org/10.1177/0962280220913588

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

# Sample size calculation and re-estimation based on the prevalence in a single-arm confirmatory diagnostic accuracy study

Statistical Methods in Medical Research
2020, Vol. 29(10) 2958–2971


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280220913588
journals.sagepub.com/home/smm
SAGE

**Maria Stark and Antonia Zapf**

## Abstract

**Introduction:** In a confirmatory diagnostic accuracy study, sensitivity and specificity are considered as co-primary endpoints. For the sample size calculation, the prevalence of the target population must be taken into account to obtain a representative sample. In this context, a general problem arises. With a low or high prevalence, the study may be overpowered in one subpopulation. One further issue is the correct pre-specification of the true prevalence. With an incorrect assumption about the prevalence, an over- or underestimated sample size will result.

**Methods:** To obtain the desired power independent of the prevalence, a method for an optimal sample size calculation for the comparison of a diagnostic experimental test with a prespecified minimum sensitivity and specificity is proposed. To face the problem of an incorrectly pre-specified prevalence, a blinded one-time re-estimation design of the sample size based on the prevalence and a blinded repeated re-estimation design of the sample size based on the prevalence are evaluated by a simulation study. Both designs are compared to a fixed design and additionally among each other.

**Results:** The type I error rates of both blinded re-estimation designs are not inflated. Their empirical overall power equals the desired theoretical power and both designs offer unbiased estimates of the prevalence. The repeated re-estimation design reveals no advantages concerning the mean squared error of the re-estimated prevalence or sample size compared to the one-time re-estimation design. The appropriate size of the internal pilot study in the one-time re-estimation design is 50% of the initially calculated sample size.

**Conclusions:** A one-time re-estimation design of the prevalence based on the optimal sample size calculation is recommended in single-arm diagnostic accuracy studies.

## Keywords

Adaptive design, co-primary endpoints, blinded sample size re-estimation, sensitivity, specificity

## 1 Introduction

The determination of the correct sample size is an essential component of a confirmatory study in general. If the sample size is too large, more patients than necessary will be exposed to a treatment or diagnostic test under investigation. Otherwise, if the sample size is too small, it will not be ensured to find a relevant effect on the basis of those patients who are involved. In each case, ethical and financial issues will arise. The special feature of a confirmatory diagnostic accuracy study is the combination of the sensitivity (as the true positive rate) and the specificity (as the true negative rate) to co-primary endpoints, measured in two independent subpopulations. This means that for both endpoints a separate sample size calculation is performed, giving the needed number of

University Medical Center Hamburg-Eppendorf, Institute of Medical Biometry and Epidemiology, Hamburg, Germany

**Corresponding author:**
Maria Stark, University Medical Center Hamburg-Eppendorf, Institute of Medical Biometry and Epidemiology, Martinistr. 52, Hamburg 20246, Germany.
Email: m.stark@uke.de

diseased and non-diseased individuals. Based on the prevalence the total sample size is calculated, which can be different for both endpoints. In this case, the maximum of the total sample sizes of both endpoints is the final sample size.

The guideline on clinical evaluation of diagnostic agents of the European Medicine Agency[1] demands the specification of the sample size in a confirmatory diagnostic accuracy study in the study protocol. The guideline highlights the dependency of the sample size on the prevalence. This means that the total sample size for the sensitivity and for the specificity may differ in the case of a low or high prevalence. This aspect leads to an unbalanced design. In a confirmatory diagnostic accuracy study, the sample size of each endpoint is often calculated with an individual power of 90% to reach an overall power of at least 80%. This is possible because sensitivity and specificity are estimated in independent subgroups. However, in the case of a low or high prevalence, the empirical overall power is noticeably larger than 80% due to the unbalanced design. This paper solves this problem of an overpowered sample size determination by providing an approach to calculate the optimal sample size depending on the prevalence. This approach is illustrated through the example of a study design containing the comparison of one experimental test to a prespecified minimum sensitivity and specificity in which the reference standard defines the true disease status. Furthermore, if the true prevalence is not known for the initial sample size calculation in this study design, a procedure for the blinded re-estimation of the sample size based on the prevalence is presented. This enables the adaptation of the sample size during the study.

In the literature, sample size calculations are hardly published in diagnostic studies.[2–4] In methodological research, there are several approaches which address the usage of binomial confidence intervals based on the normal approximation as the basis of the sample size calculation. Agresti and Coull[5] describe an own confidence interval which provides a better coverage probability than the standard Wald confidence interval. Beyond that, Piegorsch[6] gives a survey about binomial confidence intervals which are superior to the Wald confidence interval. Wei and Hutson[7] give a new sample size calculation method which is based on the expected width of the confidence interval under the assumption of an hypothesized proportion. Research for blinded sample size re-estimation in the context of clinical trials does already exist.[8–10] Asakura et al.[11] published an interim evaluation with co-primary endpoints in clinical trials. However, this approach is applicable only for co-primary endpoints measured on the same individuals. Flahault et al.[12] developed an approach for sample size calculation allowing for uncertainty in the prevalence. They determine the sample size so that the sample contains, with a predetermined probability, enough diseased and non-diseased people. No research is found addressing the problem of overpowering as a consequence of the sample size calculation for co-primary endpoints measured in independent subpopulations. Furthermore, the implementation of a blinded sample size re-estimation procedure based on the prevalence for the comparison of one experimental test to a prespecified minimum sensitivity and specificity neither could be found. This lack of research gives evidence to the present paper.

This publication is structured the following way: at first, the problem of overpowering with the conventional way of sample size calculation in diagnostic accuracy studies with co-primary endpoints is explained. The next two subsections present the theoretical basis and practical application of an approach to negotiate this problem by calculating the optimal sample size. After this, the procedure of a one-time and a repeated blinded sample size re-estimation based on the prevalence is presented. In Section 4, the results of the simulation study concerning the one-time and repeated sample size re-estimation design are compared to those of the fixed design and among each other. Finally, the results of the simulation study are discussed and a conclusion is given.

## 2 Sample size calculation in a confirmatory diagnostic accuracy study

### 2.1 Conventional sample size calculation

As already mentioned in Section 1, sensitivity and specificity are combined as co-primary endpoints which is done through the Intersection-Union Test. The global null hypothesis $H_{0_{global}}$ is defined as the union of the null hypothesis of the sensitivity $H_{0_{se}}$ and the null hypothesis of the specificity $H_{0_{sp}}$

$$H_{0_{global}}: \quad H_{0_{se}}: \theta_{se_0} = \theta_{se_1} \cup H_{0_{sp}}: \theta_{sp_0} = \theta_{sp_1} \tag{1}$$

$\theta_{se_1}$ and $\theta_{sp_1}$ represent the sensitivity and specificity of the experimental test. $\theta_{se_0}$ and $\theta_{sp_0}$ denote the minimum sensitivity and minimum specificity to which the experimental test is compared. $H_{0_{global}}$ can only be rejected if $H_{0_{se}}$ and $H_{0_{sp}}$ can be rejected. The overall power results as the product of the individual power of each endpoint, as the

endpoints are measured in independent subpopulations. In analogy, the global type I error rate is the product of the type I error rates of both endpoints. The global type I error rate is not inflated through the combination of both endpoints via the Intersection-Union Test.

For the sample size calculation in a confirmatory diagnostic accuracy trial, both endpoints must be considered. The true disease status of the patients is unknown at the time of enrolment into the study. The sample size is determined in three steps: first, the individual sample size for the sensitivity $n_{se}$ (the number of diseased individuals) and for the specificity $n_{sp}$ (the number of non-diseased individuals) is calculated by using in this paper the sample size formula for the Wald confidence interval for a single proportion.[13,14] The sample size formula is given in the example below. Second, the total sample size of both endpoints is calculated by dividing the individual sample size for the sensitivity by the prevalence, leading to the total sample size $N_{se}$, and by dividing the individual sample size for the specificity by one minus the prevalence, leading to the total sample size $N_{sp}$.[15] This must be done to obtain a representative sample with the correct ratio between cases and controls.[12] Hereby, the prevalence in a confirmatory diagnostic accuracy study means the proportion of diseased people in the target population for which the diagnostic test is developed. In the third step, the maximum of these both sample sizes represents the final sample size N of the study.[16]

This procedure is exemplified with a confirmatory single-arm diagnostic accuracy study for the diagnosis of pancreatic cancer. The example is based on a two-arm study used by McCray et al.[17] The experimental test to be examined is the computed tomography (CT).The biopsy is the reference standard. The positron emission tomography which serves as the experimental test in the publication of McCray et al.[17] is not considered here. The conventional sample size calculation is done so that an overall power ($= 1 - \beta$) of at least 80% should be reached by assigning a power of 90% to each individual endpoint. The sensitivity of the CT is expected to be $\theta_{se_1} = 0.81$ and it should be shown that it is larger than $\theta_{se_0} = 0.75$. The specificity of the CT is expected to be $\theta_{sp_1} = 0.66$ and the study aims to show that it is larger than $\theta_{sp_0} = 0.6$. The type I error rate is set to $\alpha = 0.05$ (two-sided) and the individual type II error rate of each endpoint is $\beta_{se} = \beta_{sp} = 0.1$. The prevalence $\pi$ is assumed to be 0.3. The variance of the parameter $\theta_.$ is defined as $V(\theta_.) = \theta \cdot (1 - \theta_.)$.[14] The upper $\alpha/2$ and $\beta$ quantile of the standard normal distribution is denoted by $z_{\alpha/2}$ and $z_\beta$:

1. Number of diseased individuals

$$n_{se} = \frac{[z_{\alpha/2}\sqrt{V(\theta_{se_0})} + z_{\beta_{se}}\sqrt{V(\theta_{se_1})}]^2}{(\theta_{se_0} - \theta_{se_1})^2} = 508$$

Number of non-diseased individuals

$$n_{sp} = \frac{[z_{\alpha/2}\sqrt{V(\theta_{sp_0})} + z_{\beta_{sp}}\sqrt{V(\theta_{sp_1})}]^2}{(\theta_{sp_0} - \theta_{sp_1})^2} = 683$$

2. Total sample size including at least $n_{se}$ diseased individuals

$$N_{se} = n_{se}/\pi = 508/0.3 \approx 1693$$

Total sample size including at least $n_{sp}$ non-diseased individuals

$$N_{sp} = n_{sp}/(1 - \pi) = 683/(1 - 0.3) \approx 976$$

3. $N = \max(N_{se}, N_{sp}) = 1693$

As the example shows, if the prevalence is low, the total sample size of the sensitivity determines the final sample size. Hence, more people than needed are included to show the specificity which often leads to an overpowered study. If the prevalence was high, the same problem would arise. But in this case, the specificity would probably determine the final sample size and the endpoint of the sensitivity would be overpowered now.

## 2.2. Optimal sample size calculation

To overcome the problem of an overpowered diagnostic accuracy study, an approach for the calculation of an optimal sample size is proposed. This approach ensures the desired overall power which is perfectly adjusted to the prevalence. The sample size is optimal in the way that it is the smallest representative sample that achieves the advertised overall power. The approach is based on the idea to individually split the overall power to the endpoint of the sensitivity and specificity. Hence, an individual type II error is assigned to each of both endpoints so that the required sample sizes of both endpoints are equal. To reach an overall power of 80%, the individual power of each endpoint cannot be smaller than 80%. In conclusion, none of both endpoints is overpowered which leads to a correct empirical overall power. As this method is developed for a confirmatory setting, the true disease status of the patients is unknown at the time of enrolment into the study. In analogy to the conventional sample size calculation, assumptions about the prevalence have to be made.

The mathematical definition of this approach is again exemplified through the single-arm design

$$\mathrm{N_{se}} \stackrel{!}{=} \mathrm{N_{sp}} \tag{2}$$

$$\frac{\mathrm{n_{se}}}{\pi} \stackrel{!}{=} \frac{\mathrm{n_{sp}}}{(1-\pi)} \tag{3}$$

$$\frac{[z_{\alpha/2}\sqrt{V(\theta_{\mathrm{se}_0})} + z_{\beta_{\mathrm{se}}}\sqrt{V(\theta_{\mathrm{se}_1})}]^2}{(\theta_{\mathrm{se}_0} - \theta_{\mathrm{se}_1})^2 \cdot \pi} \stackrel{!}{=} \frac{[z_{\alpha/2}\sqrt{V(\theta_{\mathrm{sp}_0})} + z_{\beta_{\mathrm{sp}}}\sqrt{V(\theta_{\mathrm{sp}_1})}]^2}{(\theta_{\mathrm{sp}_0} - \theta_{\mathrm{sp}_1})^2 \cdot (1-\pi)} \tag{4}$$

$$\begin{aligned} &z_{\beta_{\mathrm{se}}}\sqrt{V(\theta_{\mathrm{se}_1})}(\theta_{\mathrm{sp}_0} - \theta_{\mathrm{sp}_1})\sqrt{(1-\pi)} - z_{\beta_{\mathrm{sp}}}\sqrt{V(\theta_{\mathrm{sp}_1})}(\theta_{\mathrm{se}_0} - \theta_{\mathrm{se}_1})\sqrt{\pi} \\ &\stackrel{!}{=} z_{\alpha/2}\sqrt{V(\theta_{\mathrm{sp}_0})}(\theta_{\mathrm{se}_0} - \theta_{\mathrm{se}_1})\sqrt{\pi} - z_{\alpha/2}\sqrt{V(\theta_{\mathrm{se}_0})}(\theta_{\mathrm{sp}_0} - \theta_{\mathrm{sp}_1})\sqrt{(1-\pi)} \end{aligned} \tag{5}$$

Under the condition

$$\mathrm{Power_{se}} \cdot \mathrm{Power_{sp}} \stackrel{!}{=} \mathrm{Power_t} \tag{6}$$

$$(1-\beta_{\mathrm{se}}) \cdot (1-\beta_{\mathrm{sp}}) \stackrel{!}{=} \mathrm{Power_t} \tag{7}$$

$$\beta_{\mathrm{sp}} = \frac{1 - \beta_{\mathrm{se}} - \mathrm{Power_t}}{1-\beta_{\mathrm{se}}} \tag{8}$$

Plug the condition into the sample size calculation

$$\begin{aligned} &z_{\beta_{\mathrm{se}}}\sqrt{V(\theta_{\mathrm{se}_1})}(\theta_{\mathrm{sp}_0} - \theta_{\mathrm{sp}_1})\sqrt{(1-\pi)} - z_{\frac{1-\beta_{\mathrm{se}}-\mathrm{Power_t}}{1-\beta_{\mathrm{se}}}}\sqrt{V(\theta_{\mathrm{sp}_1})}(\theta_{\mathrm{se}_0} - \theta_{1_{\mathrm{se}}})\sqrt{\pi} \\ &\stackrel{!}{=} z_{\alpha/2}\sqrt{V(\theta_{\mathrm{sp}_0})}(\theta_{\mathrm{se}_0} - \theta_{\mathrm{se}_1})\sqrt{\pi} - z_{\alpha/2}\sqrt{V(\theta_{\mathrm{se}_0})}(\theta_{\mathrm{sp}_0} - \theta_{\mathrm{sp}_1})\sqrt{(1-\pi)} \end{aligned} \tag{9}$$

Equation (9) cannot be solved analytically with respect to $\beta_{\mathrm{se}}$ or $\beta_{\mathrm{sp}}$ and is therefore solved by the software R.[18] The R-code for this sample size calculation is given in the supplement materials.

The analysis of a study based on this optimal sample size calculation is proposed to be done by the logit confidence interval. It is defined as

$$\mathrm{expit}\left(\ln\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right) \pm z_{\alpha/2} \cdot \frac{1}{\sqrt{n \cdot \hat{\theta} \cdot (1-\hat{\theta})}}\right) \tag{10}$$

with $\mathrm{expit}(x) = \frac{e^x}{1+e^x}$[19], $\hat{\theta}$ as $\hat{\theta}_{\mathrm{se}_1}$ or $\hat{\theta}_{\mathrm{sp}_1}$ and n as $\mathrm{n_{se}}$ or $\mathrm{n_{sp}}$, respectively. The individual null hypothesis of each endpoint will be rejected, if $\theta_{\mathrm{se}_0}$ or $\theta_{\mathrm{sp}_0}$ does not fall into this two-sided 1-alpha confidence interval. If the study

was analyzed with the Wald confidence interval, the empirical power would be lower than the theoretical one. The empirical power would also be lower than the theoretical one, if the optimal sample size calculation was based on the logit confidence interval and if the evaluation was done with it. Fleiss et al.[20] address this problem in the context of binomial confidence intervals which are based on the normal approximation. They recommend to use a sample size formula with continuity correction to increase the sample size. They show that the empirical power is now a little higher than the theoretical one. Using the procedure proposed in this paper, the empirical evaluations given in Section 4 suggest the theoretical power is achieved. This is caused by the fact that the sample size of the logit confidence interval is smaller than the one of the Wald confidence interval. The left part of equation (11) represents the sample size of the logit confidence interval and the right part shows the sample size of the Wald confidence interval. The numerator of the sample size of the logit interval is smaller than the numerator of the sample size of the Wald interval. The denumerator of the sample size of the logit interval is larger than the denumerator of the sample size of the Wald interval. Hence, the analysis with the logit confidence interval based on the larger sample size of the Wald confidence interval ensures to reach the theoretical power. In Appendix 1, the derivation of the sample size of the logit interval is given.

$$\frac{\left[\frac{z_{\alpha/2}}{\sqrt{\theta_0(1-\theta_0)}}+\frac{z_{\beta}}{\sqrt{\theta_1(1-\theta_1)}}\right]^2}{\left[\ln\left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)\right]^2} < \frac{\left[z_{\alpha/2}\sqrt{\theta_0(1-\theta_0)}+z_{\beta}\sqrt{\theta_1(1-\theta_1)}\right]^2}{(\theta_0-\theta_1)^2} \tag{11}$$

## 2.3 Application of the optimal sample size calculation

The optimal sample size calculation method is now applied to the single-arm diagnostic accuracy study for the diagnosis of pancreatic cancer already used in the context of the conventional sample size calculation in Section 2.1. Both sample size calculations are based on the requirement to reach an overall power of 80% and a maximal type I error rate per endpoint of 5% (two-sided). Figure 1 compares the empirical overall power and the sample size between the conventional and optimal sample size calculation for a varying prevalence $\pi$. The sample sizes of both approaches in Figure 1 on the right are almost equal if the prevalence is balanced. But with a decreasing or increasing prevalence, the sample sizes of both approaches differ. Due to the individual split of the overall power

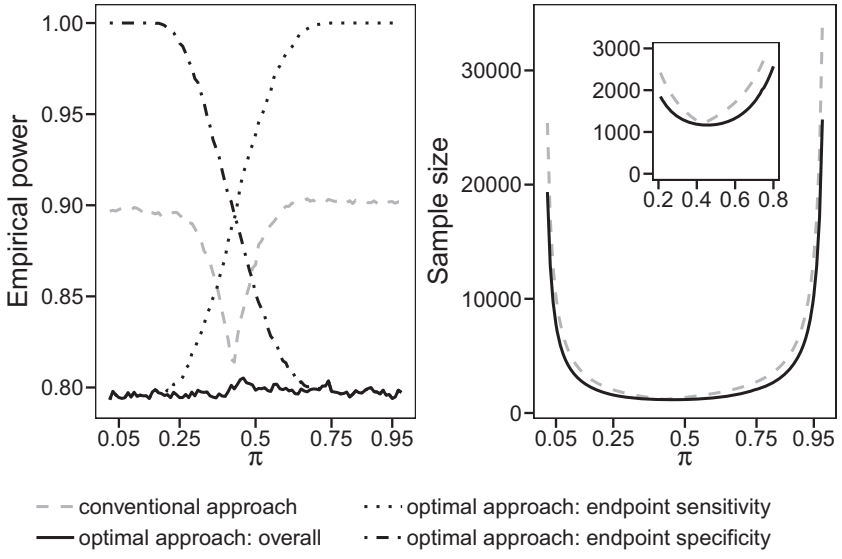



**Figure 1.** Comparison of the conventional and optimal sample size calculation with respect to a varying prevalence $\pi$ and the resulting empirical overall power or sample size. The parameters of this example are as follows: $\alpha = 0.05$ (two-sided), $\theta_{se_0} = 0.75$, $\theta_{se_1} = 0.81$, $\theta_{sp_0} = 0.6$ and $\theta_{sp_1} = 0.66$. The sample size calculation in the conventional approach is based on $\beta_{se} = \beta_{sp} = 0.1$. In the optimal approach, the overall power is aimed to be 80% and is individually split to both endpoints which is depending on the prevalence.In the figure on the right showing the sample size, an enlarged picture inset is given between $\pi = [0.2, 0.8]$ to highlight the difference in the sample sizes between both approaches.

to both endpoints in the optimal approach, the sample size of the optimal approach is smaller than the one of the conventional approach. The study under the conventional procedure is highly overpowered in the case of a low or high prevalence. With a balanced prevalence, the empirical power of the conventional approach is closer to the desired theoretical power of 80%. Adapted to the prevalence, the empirical power of a study conducted with the optimal sample size does not relevantly differ from the theoretical power of 80%. Additionally, Figure 1 shows the empirical individual power of both endpoints which varies complementary between 80% and almost 100%. The individual power of one endpoint cannot become smaller than the advertised overall power of 80%.

To reveal the importance of a correct assumption about the prevalence, the discrepancy between the initial sample size based on a wrongly assumed prevalence and the sample size based on the true prevalence is considered. The sample sizes are calculated with the optimal sample size calculation procedure. In the context of the chosen example, the initially wrongly assumed prevalence is 0.3 with a resulting initial sample size of 1367 individuals. Table 1 shows several scenarios with a variation of the true prevalence $\pi_{\text{true}}$ and the corresponding true sample sizes (true N). Furthermore, Table 1 contains the individual power of each endpoint using the true sample size. The true prevalence varies between 0.1 and 0.6. The largest discrepancy between the initial and true sample size is revealed in the case of a true prevalence of 0.1. With a small prevalence, the sensitivity determines the sample size. If the assumed prevalence is larger than the true one, the initially calculated sample size will be too small. Referring to McCray et al.,[17] the true prevalence of the chosen example equals 0.47. If this is true, this will lead to an overpowered study as the true sample size of 1165 is smaller than the initial sample size. If the true prevalence is 0.6, a true sample size of 1325 will result which is similar to the initial one. This can be explained again by referring to Figure 1 which depicts the symmetry of the sample size around a prevalence of approximately 0.42. As the prevalence of 0.3 and 0.6 is approximately equally distant from 0.42, both corresponding sample sizes do not differ a lot.

The comparison of the initial and true sample size gives evidence for a re-estimation design of the prevalence during a confirmatory diagnostic accuracy study. In this context, a wrongly assumed prevalence can be re-evaluated and consequently the sample size can be adjusted. The following section introduces the procedure of the blinded sample size re-estimation based on the prevalence, using the optimal sample size calculation approach.

## 3 Blinded sample size re-estimation

In a fixed design without an internal pilot study, the sample size is calculated based on assumptions of a preceding study and is not adjusted before the final analysis. The process of an internal pilot design with a one-time re-estimation of the prevalence also starts with this initial sample size calculation but runs through five phases:[21]

1. Calculation of the initial sample size with the optimal procedure (e.g. based on assumptions of a preceding study)
2. Recruitment of patients until the predetermined size of the internal pilot study is reached
3. Re-estimation of the prevalence and recalculation of the sample size with the optimal procedure
4. If the recalculated sample size is larger than the already recruited sample size, further patients will be recruited until the adjusted sample size will be reached. Otherwise, no further recruitment is necessary.
5. Analysis of the study based on the unadjusted type I error level due to the blinded character of the re-estimation procedure

In the repeated prevalence re-estimation design, the prevalence and the sample size are re-estimated several times based on a steadily growing sample. The recruited sample increases during each run by a predetermined size. The re-estimation procedure is finished as soon as the already recruited sample is too large to not exceed the

**Table 1.** Application of the optimal sample size calculation approach: Highlighting the discrepancy between the initial sample size of 1367 people based on the wrongly assumed prevalence of 0.3 and the true sample sizes (true N) based on a varying true prevalence $\pi_{\text{true}}$. The parameters of the scenarios are $\alpha = 0.05$ (two-sided), power $= 0.8$, $\theta_{\text{se}_0} = 0.75$, $\theta_{\text{se}_1} = 0.81$, $\theta_{\text{sp}_0} = 0.6$, $\theta_{\text{sp}_1} = 0.66$.

| $\pi_{\text{true}}$ | 0.10 | 0.20 | 0.40 | 0.47 | 0.50 | 0.60 |
|---|---|---|---|---|---|---|
| true N | 3870 | 1940 | 1185 | 1165 | 1178 | 1325 |
| $\text{Power}_{\text{se}}$ | 0.801 | 0.802 | 0.878 | 0.923 | 0.939 | 0.984 |
| $\text{Power}_{\text{sp}}$ | 1 | 0.998 | 0.912 | 0.869 | 0.852 | 0.813 |

recalculated sample size after the addition of the next fraction of patients. Hence, the repeated re-estimation design iterates between step 2 and 3 before it proceeds to step 4 and 5.

The prevalence is re-estimated by the well-known maximum likelihood estimator of a binomial proportion[22]

$$\hat{\pi} = \frac{X}{n} \tag{12}$$

$X$ denotes the number of diseased patients in the sample and $n$ represents the sample size on which the re-estimation is based. The prevalence represents a nuisance parameter in a diagnostic trial. Consequently, the recalculation of the sample size based on the re-estimated prevalence is defined as a blinded adaptive design.[23] In the context of a blinded sample size recalculation in a diagnostic study, the sensitivity or the specificity of the experimental test is kept a secret because they are not of interest during the interim analysis. Therefore, the type I error is expected to be not inflated which will be explored by the simulation study.

To evaluate the appropriate size of the internal pilot study in the context of the one-time re-estimation design, the quotient $R$ is used. It is defined as[24]

$$R = \frac{E(n_1)}{n_F(\pi_{\text{true}})} \tag{13}$$

$E(n_1)$ represents the simulated adjusted sample size after re-estimation of the prevalence. $n_F(\pi_{\text{true}})$ denotes the correct sample size initially calculated with the true prevalence. Values of $R$ which are close to 1 represent an efficient size of the internal pilot study.

## 4 Simulation study

The simulation study aims to evaluate the type I error rate, the power, and the bias of the design with the one-time as well as the repeated re-estimation of the prevalence, each in comparison to the fixed design. Furthermore, the appropriate size of the internal pilot study for the one-time re-estimation is proposed. The mean squared errors (MSEs) of the re-estimated prevalence and of the adjusted sample size are compared between both adaptive designs. For the design with a one-time re-estimation 3888 scenarios, and for the repeated re-estimation design 1296 scenarios are simulated. They are given in Table 2. Per scenario, 100,000 replications are performed.

### 4.1 One-time re-estimation of the prevalence

At first, the results of the simulations concerning the type I error rate, power, and bias of the design with the one-time re-estimation of the prevalence are given. The results of the design with the repeated re-estimation of the prevalence are similar to those of the one-time re-estimation design and are therefore not shown in the text. The

**Table 2.** Simulated scenarios.

| | One-time re-estimation | Repeated re-estimation |
|---|---|---|
| True prevalence $\pi_{\text{true}}$ | 0.2, 0.4, 0.6, 0.8 | |
| Assumed prevalence $\pi_{\text{ass.}}$ | $\pi_{\text{true}} - 0.1$, $\pi_{\text{true}} + 0.1$ | |
| Minimum sensitivity $\theta_{\text{se}_0}$ | 0.6, 0.7, 0.8 | |
| Minimum specificity $\theta_{\text{sp}_0}$ | 0.6, 0.7, 0.8 | |
| Under the null $H_0: \theta_0 = \theta_1$ | | |
| Fraction for re-estimation $\psi$ | 0.02, 0.1, 0.3, 0.5, 0.7 | 0.1 |
| Significance level $\alpha$ per endpoint | 0.05 (two-sided) | |
| Sensitivity experimental test $\theta_{\text{se}_1}$ | 0.6, 0.7, 0.8 | |
| Specificity experimental test $\theta_{\text{sp}_1}$ | 0.6, 0.7, 0.8 | |
| Under the alternative $H_1: \theta_0 \neq \theta_1$ | | |
| Fraction for re-estimation $\psi$ | 0.5 | 0.1 |
| Overall power $1 - \beta$ | 0.8 | |
| $\theta_{\text{se}_1}$ | $\theta_{0_{\text{se}}} + 0.05$, $\theta_{0_{\text{se}}} + 0.1$, $\theta_{0_{\text{se}}} + 0.15$ | |
| $\theta_{\text{sp}_1}$ | $\theta_{0_{\text{sp}}} + 0.05$, $\theta_{0_{\text{sp}}} + 0.1$, $\theta_{0_{\text{sp}}} + 0.15$ | |

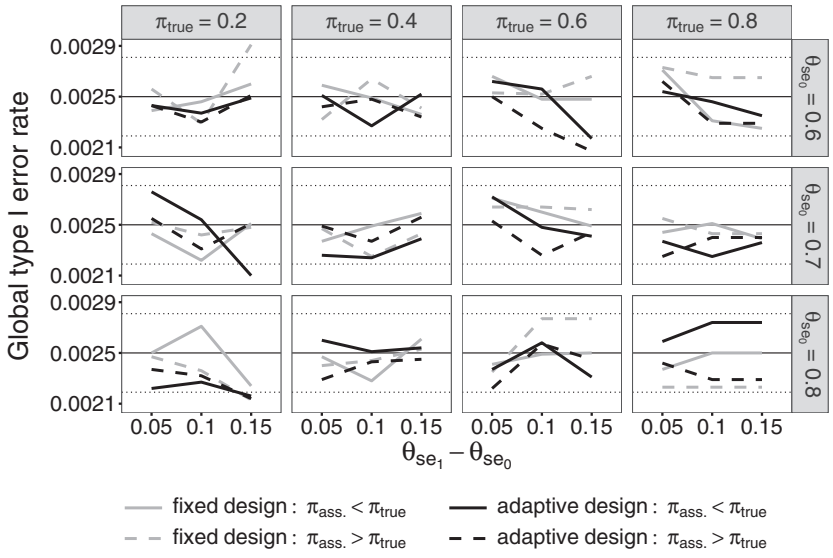



**Figure 2.** Comparison of the global type I error rates of the fixed design and the adaptive design containing a one-time re-estimation of the prevalence with $\theta_{sp_0} = 0.6$, $\theta_{sp_1} - \theta_{sp_0} = 0.1$ and the size of the internal pilot study $\psi = 0.5$. The initially assumed prevalence is either over- or underestimated. The black dotted lines mark the interval of the Monte Carlo error due to simulations.

results of all simulated scenarios are given in the supplement materials as tables. Figure 2 shows the global type I error rate of the one-time re-estimation and the fixed design for the scenarios with $\theta_{sp_0} = 0.6$, $\theta_{sp_1} - \theta_{sp_0} = 0.1$ and the size of the internal pilot study $\psi = 0.5$. Furthermore, the Monte Carlo error due to simulations ($1.96 \times SE = 0.00016$) is depicted. The sample size is calculated with a significance level of each endpoint of 0.05 (two-sided) which leads to a global significance level of 0.0025 (two-sided). As mentioned in Section 2.1, the global type I error rate results as the product of both individual type I error rates. Figure 2 reveals that the global type I error rate is sufficiently controlled in this adaptive design as well as in the fixed design. This is also the case for the individual type I error rates.

A figure containing the individual type I error rate for sensitivity and specificity for the same scenarios as for the global type I error rate is given in the supplement materials. In the following, the results of the individual type I error rate of the endpoint of the sensitivity will be explained. In the scenarios with a small prevalence, a high minimum sensitivity, and a much higher sensitivity of the experimental test, the type I error rate is smaller than 0.05 irrespective of whether the true prevalence is initially over- or underestimated. Kottas et al.[25] show that the logit interval is conservative in the case of a small sample size. In the named scenarios, the small sample is represented through a small number of diseased patients due to the low prevalence. The high minimum sensitivity and the high sensitivity of the experimental test additionally diminish the sample size. The decision of a potential rejection of the null hypothesis is based on this small number of diseased patients who are diagnosed correctly with a high probability. Hence, the use of the logit confidence interval leads to the conservative type I error rates in these scenarios. The individual type I error rates of the specificity reveal the same results in the corresponding scenarios with a high prevalence.

Figure 3 contains the results of the overall power simulations of the scenario with the same parameters as described above in the context of the type I error rate. The results reveal the effect of a wrongly assumed prevalence during sample size calculations in the fixed design. The fixed design is either over- or underpowered depending on the difference between the true and the initially assumed prevalence. If the true prevalence is assumed to be too low, the study will be overpowered in the case of a low prevalence. In this context, the individual sample size of the sensitivity is the maximum to choose. But it is divided by a too small assumed prevalence which leads to a too large sample size. This causes an overpowered study. If the true prevalence is high, the individual sample size of the specificity will determine the total sample size. If the true prevalence is underestimated, the sample size of the specificity will be divided by a too large proportion of the non-diseased. Hence, the sample size and the power are too low. This mechanism will be reversed if the true prevalence is overestimated. Additionally to the true and assumed prevalence, the difference between minimum sensitivity and the sensitivity of

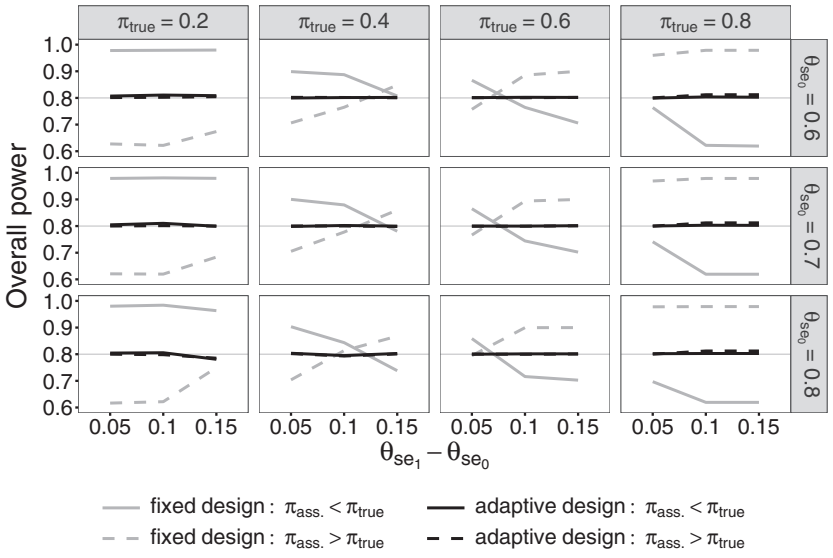



**Figure 3.** Comparison of the overall power of the fixed design and the adaptive design containing a one-time re-estimation of the prevalence with $\theta_{sp_0} = 0.6$, $\theta_{sp_1} - \theta_{sp_0} = 0.1$ and the size of the internal pilot study $\psi = 0.5$ with either an initially over- or underestimated true prevalence.

the experimental test $\theta_{se_1} - \theta_{se_0}$ influences the overall power of the fixed design. If the true prevalence is low, a large difference will diminish the over- or underpowering. Otherwise, if the true prevalence is high, a high difference between the assumed and minimum sensitivity will intensify the consequences of an initially wrongly assumed prevalence.

The overall power of the one-time re-estimation design reaches exactly the desired power of 80%. Due to the re-estimation of the prevalence, the effect of an initially wrongly specified prevalence can be absorbed. This is valid for all other simulated scenarios.

The bias of the estimated prevalence decreases the larger the size of the internal pilot study $\psi$ is. But with $\psi = 0.1$, the prevalence is already re-estimated without any bias. A figure containing the relative bias is given in the supplement materials.

## 4.2 The size of the internal pilot study in the one-time re-estimation design

The appropriate size of the internal pilot study is explored by simulating the quotient $R$ of the adjusted sample size after re-estimation $E(n_1)$ divided by the correct sample size initially calculated with the true prevalence $n_F(\pi_{true})$. Values of $R$ which are equal to 1 indicate a correct sample size re-estimation. Figure 4 depicts the quotient $R$ in dependence of $\psi$ for the scenarios under $\theta_{se_0} = 0.8$, $\theta_{sp_0} = 0.8$ and $\theta_{sp_1} - \theta_{sp_0} = 0.15$ with either an initially over- or underestimated prevalence. If $\psi = 0.1$, $R$ will differ clearly from the optimum 1. If $\psi$ increases up to 0.3, the correct sample size will be overestimated about approximately 10%, especially with an unbalanced true prevalence. This proportion of overestimation is reduced with $\psi = 0.5$. If $\psi$ becomes larger than 0.5 up to 0.9, the quotient $R$ is not relevantly closer to the value of 1. The results for $\psi = 0.9$ are not depicted in Figure 4 as they provide no further information about the appropriate size of the internal pilot study. Hence, the appropriate size of the internal pilot study is considered to be $\psi = 0.5$. These thoughts are only valid for the simulated scenarios as just the prevalence is wrongly assumed. In reality, there might be further parameters that are wrongly assumed during sample size calculation (e.g. the sensitivity or the specificity of the experimental test). Consequently, $\psi = 0.5$ might not be the appropriate fraction for such scenarios.

## 4.3 Comparison of the design with one-time and repeated re-estimation of the prevalence

This section compares the designs with the one-time and repeated re-estimation of the prevalence with respect to the MSE. The MSE measures the squared mean difference between the true prevalence $\pi_{true}$ and the re-estimated

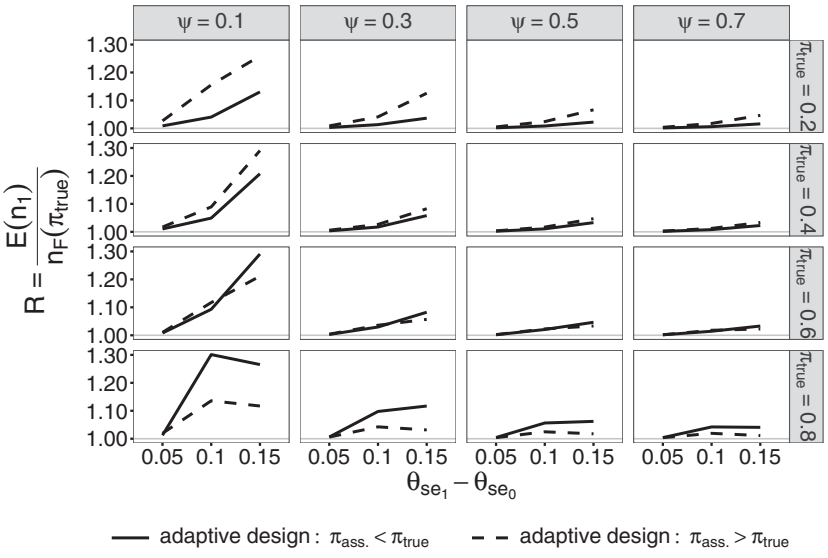



**Figure 4.** *R* in dependence of the size of the internal pilot study $\psi$ for the scenarios with $\theta_{se_0} = 0.8$, $\theta_{sp_0} = 0.8$, and $\theta_{sp_1} - \theta_{sp_0} = 0.15$ with either an initially over- or underestimated prevalence. $E(n_1)$ denotes the simulated adjusted sample size after re-estimation of the prevalence. $n_F(\pi_{true})$ represents the correct sample size initially calculated with the true prevalence. Values of *R* equal to 1 indicate a correct sample size re-estimation.

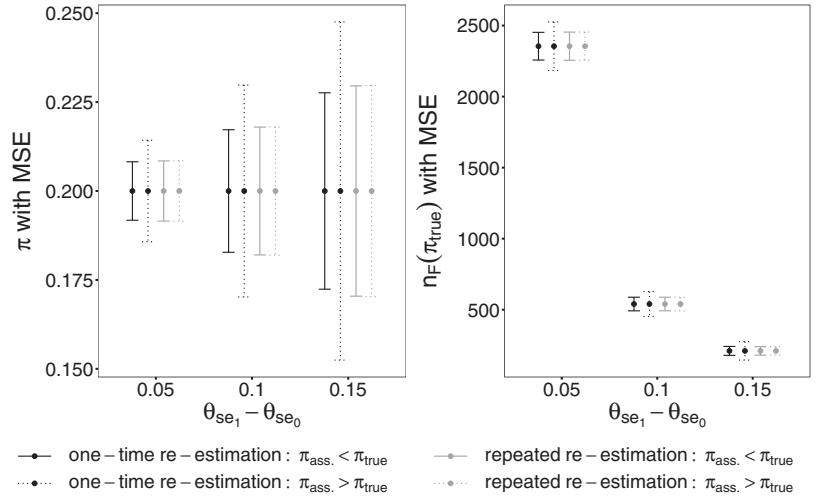



**Figure 5.** Comparison of the design with one-time and repeated re-estimation of the prevalence regarding the MSE of the re-estimated prevalence $\hat{\pi}$ and of the adjusted sample size $E(n_1)$ for the scenarios with $\theta_{se_0} = 0.8$, $\theta_{sp_0} = 0.8$, $\theta_{sp_1} = 0.95$, and $\pi_{true} = 0.2$. The appropriate size of the internal pilot study of the one-time re-estimation design is $\psi_o = 0.5$. The fraction for re-estimation in the repeated re-estimation design is $\psi_R = 0.1$. MSE: mean squared error.

prevalence, or the squared mean difference between the true sample size $n_F(\pi_{true})$ and the adjusted sample size, respectively. Figure 5 shows these results for the scenarios $\theta_{se_0} = 0.8$, $\theta_{sp_0} = 0.8$, $\theta_{sp_1} = 0.95$, $\pi_{true} = 0.2$. The internal pilot study of the one-time re-estimation design has the appropriate size with $\psi = 0.5$. The fraction for re-estimation in the repeated re-estimation design is $\psi_R = 0.1$. The graphic on the left in Figure 5 refers to the MSE of the re-estimated prevalence; the graphic on the right refers to the MSE of the adjusted sample size. This figure reveals that the one-time re-estimation design has no relevant disadvantage compared to the repeated re-estimation design concerning the MSE.

## 5 Discussion

This paper deals with two aspects of confirmatory diagnostic accuracy studies. First, it presents an improved method for the sample size calculation. This approach allows to calculate the sample size by individually splitting the overall power to each endpoint depending on the prevalence. Consequently, the study will not be overpowered. This approach can be generalized for all sample size calculations combining two co-primary endpoints which are based on independent data. Its idea is to get the same sample size for each endpoint for the purpose of not needing to choose a maximum out of them. Hence, it is not limited to diagnostic studies.

Second, this paper evaluates two designs to re-estimate the prevalence and to adjust the sample size in a confirmatory diagnostic accuracy study: the one-time re-estimation and repeated re-estimation design. For both designs, we propose the optimal approach for the initial and adjusted sample size calculations. Both blinded sample size re-estimation designs do not inflate the type I error rate and re-estimate the prevalence without any bias. These two re-estimation procedures compensate a wrongly assumed prevalence and its consequences on the initial sample size. Consequently, the empirical overall power equals the desired theoretical one.

As chosen in the simulation study, a difference between the assumed and true prevalence of 10% is considered to be a realistic deviation. The assumptions about the prevalence in a confirmatory diagnostic accuracy study come in general from preceding studies. Hence, the assumed difference in the confirmatory diagnostic accuracy study should not differ heavily from the true one.

The repeated re-estimation design reveals no relevant advantage concerning the MSE of the re-estimated prevalence or of the adjusted sample size. In both designs, an initially wrongly assumed prevalence can be efficiently corrected.

Hence, we recommend the application of a one-time re-estimation design in a confirmatory diagnostic accuracy study. A unique re-estimation of the prevalence is sufficient. It shows no disadvantage concerning the precision of the estimation but causes less effort compared to the repeated re-estimation procedure.

The appropriate size of the internal pilot study in the one-time re-estimation design is evaluated to be 50% of the initially calculated sample size. The sample size in diagnostic accuracy studies strongly varies from hundred to several thousands participants. The prevalence and the effect size of the sensitivity and of the specificity mainly cause this large range. To be able to make common statements about the appropriate size of the internal pilot study despite of the large range of sample sizes, the size of the internal pilot study is indicated as a proportion.

## 6 Conclusion

In this paper, a new method for the calculation of the sample size in a confirmatory diagnostic accuracy study with independent co-primary endpoints, the sensitivity and the specificity, is developed. With this method, it is possible to avoid overpowered diagnostic studies which often appear with the conventional approach of sample size calculation. The idea of the optimal sample size calculation is to individually split the overall power to both endpoints in dependence of the prevalence. Furthermore, two blinded designs for the re-estimation of the sample size based on the prevalence are presented either with a one-time or a repeated re-estimation. These designs are evaluated in a simulation study under various parameter combinations. Due to the blinded re-estimation, the type I error rate is not inflated. An initially wrongly assumed prevalence can be compensated and the desired overall power is reached. The design with a one-time re-estimation reveals no disadvantages concerning the MSE of the re-estimated prevalence or adjusted sample size compared to the repeated re-estimation design. Therefore, it is recommended for a confirmatory diagnostic accuracy study. The re-estimation of the prevalence has practical relevance to avoid over- or underpowered studies with wrongly specified sample sizes. Hence, an unnecessary burden of participants in a confirmatory diagnostic trial can be inhibited.

### Acknowledgements

We thank Werner Brannath for his inspiring ideas in the context of the development of the new methods. Our thanks also go to the unknown reviewers who really helped to improve the manuscript. Furthermore, we acknowledge the Deutsche Forschungsgemeinschaft for financing the project "Flexible designs for diagnostic studies" to which this article belongs.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article is supported by the Deutsche Forschungsgemeinschaft (ZA 687/1-1).

## ORCID iD


Maria Stark https://orcid.org/0000-0002-1463-0885

## Supplemental material

Supplement material is available online for this article.

## References

1. Committee for Medicinal Products for Human Use (CHMP), et al. *Guideline on clinical evaluation of diagnostic agents*. London: European Medicines Agency, https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-evaluation-diagnostic-agents_en.pdf (2009, accessed 25 July 2018).
2. Jones S, Carley S and Harrison M. An introduction to power and sample size estimation. *Emerg Med J* 2003; **20**: 453–458.
3. Bachmann LM, Puhan MA, Ter Riet G, et al. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ* 2006; **332**: 1127–1129.
4. Bochmann F, Johnson Z and Azuara-Blanco A. Sample size in studies on diagnostic accuracy in ophthalmology: a literature survey. *Br J Ophthalmol* 2007; **91**: 898–900.
5. Agresti A and Coull BA. Approximate is better than exact for interval estimation of binomial proportions. *Am Stat* 1998; **52**: 119–126.
6. Piegorsch WW. Sample sizes for improved binomial confidence intervals. *Comput Stat Data Anal* 2004; **46**: 309–316.
7. Wei L and Hutson AD. A comment on sample size calculations for binomial confidence intervals. *J Appl Stat* 2013; **40**: 311–319.
8. Friede T and Kieser M. Blinded sample size re-estimation in superiority and noninferiority trials: bias versus variance in variance estimation. *Pharm Stat* 2013; **12**: 141–146.
9. Sander A, Rauch G and Kieser M. Blinded sample size recalculation in clinical trials with binary composite endpoints. *J Biopharm Stat* 2017; **27**: 705–715.
10. Proschan MA. Sample size re-estimation in clinical trials. *Biom J* 2009; **51**: 348–357.
11. Asakura K, Hamasaki T and Evans SR. Interim evaluation of efficacy or futility in group-sequential trials with multiple co-primary endpoints. *Biom J* 2017; **59**: 703–731.
12. Flahault A, Cadilhac M and Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol* 2005; **58**: 859–862.
13. Zhou XH, McClish DK and Obuchowski NA. *Statistical methods in diagnostic medicine*. **Vol. 569**. Hoboken, NJ: John Wiley & Sons, 2009.
14. Obuchowski NA. Sample size calculations in studies of test accuracy. *Stat Methods Med Res* 1998; **7**: 371–392.
15. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. *J Biomed Inform* 2014; **48**: 193–204.
16. Buderer NMF. Statistical methodology: I. incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Acad Emerg Med* 1996; **3**: 895–900.
17. McCray GP, Titman AC, Ghaneh P, et al. Sample size re-estimation in paired comparative diagnostic accuracy studies with a binary response. *BMC Med Res Methodol* 2017; **17**: 102.
18. R Core Team. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, http://www.R-project.org/ (2014, accessed 8 August 2018).
19. Held L and Sabanés Bové D. *Applied statistical inference: Likelihood and Bayes*. Berlin: Springer, 2014.
20. Fleiss JL, Levin B and Paik MC. *Statistical methods for rates and proportions*. Hoboken, NJ: John Wiley & Sons, 2013.
21. Brinton JT, Ringham BM and Glueck DH. An internal pilot design for prospective cancer screening trials with unknown disease prevalence. *Trials* 2015; **16**: 458.
22. Brown LD, Cai TT and DasGupta A. Interval estimation for a binomial proportion. *Stat Sci* 2001; **16**: 101–117.
23. Wassmer G and Brannath W. *Group sequential and confirmatory adaptive designs in clinical trials*. Cham: Springer International Publishing AG, 2016.
24. Denne JS and Jennison C. Estimating the sample size for a t-test using an internal pilot. *Stat Med* 1999; **18**: 1575–1585.
25. Kottas M, Kuss O and Zapf A. A modified Wald interval for the area under the ROC curve (AUC) in diagnostic case-control studies. *BMC Med Res Methodol* 2014; **14**: 26.
26. Chow SC, Shao J, Wang H, et al. *Sample size calculations in clinical research*. New York: Chapman and Hall/CRC, 2017.
27. HyLown Consulting LLC. Deriving Z-test formulas: 1-sample, 1-sided, http://powerandsamplesize.com/Knowledge/derive-z-test-1-sample-1-sided (2013, accessed 17 April 2019).

## Appendix 1

### A.1. A derivation of the sample size of the logit confidence interval

Find the sample size $n$ so that P(Reject $H_0$ | $H_0$ is true) $\leq \alpha$ and P(Reject $H_0$ | $H_0$ is false) $\geq 1-\beta$.[20,26] $\hat{\theta}$ is the maximum likelihood estimator of a true proportion $\theta$ and is approximately normally distributed for large $n$[19]

$$\hat{\theta} \overset{a}{\sim} N\left(\theta, \frac{\theta(1-\theta)}{n}\right), \quad \text{with} \quad \frac{\theta(1-\theta)}{n} = \frac{\sigma^2}{n}$$

- Definition of the type I error rate with the critical value $\zeta$[27]

$$\begin{aligned}
\alpha/2 &= P(\theta \geq \zeta | H_0) \\
&= 1 - P\left(\frac{\theta-\theta_0}{\sigma^2/\sqrt{n}} \leq \frac{\zeta-\theta_0}{\sigma^2/\sqrt{n}} | H_0\right) \\
&= 1 - \phi\left(\frac{\zeta-\theta_0}{\sigma^2/\sqrt{n}}\right) \\
z_{1-\alpha/2} &= -z_{\alpha/2} = \frac{\zeta-\theta_0}{\sigma^2/\sqrt{n}} \\
\Rightarrow \zeta &= \theta_0 - z_{\alpha/2}\sigma^2/\sqrt{n}
\end{aligned}$$

- Definition of the type I error rate for the logit interval

$$\begin{aligned}
\zeta &= \frac{\exp\left(\ln\left(\frac{\theta_0}{1-\theta_0}\right) - z_{\alpha/2}\frac{1}{\sqrt{n\theta_0(1-\theta_0)}}\right)}{1+\exp\left(\ln\left(\frac{\theta_0}{1-\theta_0}\right) - z_{\alpha/2}\frac{1}{\sqrt{n\theta_0(1-\theta_0)}}\right)} \\
&= \frac{\frac{\theta_0}{1-\theta_0}}{\exp\left(\frac{z_{\alpha/2}}{\sqrt{n\theta_0(1-\theta_0)}}\right)} \\
&\quad 1 + \frac{\frac{\theta_0}{1-\theta_0}}{\exp\left(\frac{z_{\alpha/2}}{\sqrt{n\theta_0(1-\theta_0)}}\right)}
\end{aligned}$$

- Definition of the power in general[27]

$$\begin{aligned}
\text{Power} &= P(\theta \geq \zeta | H_1) \\
&= 1 - P(\theta \leq \zeta | H_1) \\
&= 1 - P\left(\frac{\theta-\theta_1}{\sigma^2/\sqrt{n}} \leq \frac{\zeta-\theta_1}{\sigma^2/\sqrt{n}} | H_1\right) \\
1-\beta &= 1 - \phi\left(\frac{\zeta-\theta_1}{\sigma^2/\sqrt{n}}\right) \\
\beta &= \phi\left(\frac{\zeta-\theta_1}{\sigma^2/\sqrt{n}}\right) \\
z_\beta &= \frac{\zeta-\theta_1}{\sigma^2/\sqrt{n}} \\
\Rightarrow \zeta &= \theta_1 + z_\beta\sigma^2/\sqrt{n}
\end{aligned}$$

Definition of the power for the logit interval

$$\zeta = \frac{\exp\left(\ln\left(\frac{\theta_1}{1-\theta_1}\right) + z_\beta \frac{1}{\sqrt{n\theta_1(1-\theta_1)}}\right)}{1+\exp\left(\ln\left(\frac{\theta_1}{1-\theta_1}\right) + z_\beta \frac{1}{\sqrt{n\theta_1(1-\theta_1)}}\right)} = \frac{\left(\frac{\theta_1}{1-\theta_1}\right)\cdot \exp\left(\frac{z_\beta}{\sqrt{n\theta_1(1-\theta_1)}}\right)}{1+\left(\frac{\theta_1}{1-\theta_1}\right)\cdot \exp\left(\frac{z_\beta}{\sqrt{n\theta_1(1-\theta_1)}}\right)}$$

$$\frac{\frac{\frac{\theta_0}{1-\theta_0}}{\exp\left(\frac{z_{\alpha/2}}{\sqrt{n\theta_0(1-\theta_0)}}\right)}}{1+\frac{\frac{\theta_0}{1-\theta_0}}{\exp\left(\frac{z_{\alpha/2}}{\sqrt{n\theta_0(1-\theta_0)}}\right)}} = \frac{\left(\frac{\theta_1}{1-\theta_1}\right)\cdot \exp\left(\frac{z_\beta}{\sqrt{n\theta_1(1-\theta_1)}}\right)}{1+\left(\frac{\theta_1}{1-\theta_1}\right)\cdot \exp\left(\frac{z_\beta}{\sqrt{n\theta_1(1-\theta_1)}}\right)}$$

$$\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)} = \exp\left(\frac{z_{\alpha/2}}{\sqrt{n\theta_0(1-\theta_0)}} + \frac{z_\beta}{\sqrt{n\theta_1(1-\theta_1)}}\right)$$

$$\sqrt{n} = \frac{\frac{z_{\alpha/2}}{\sqrt{\theta_0(1-\theta_0)}} + \frac{z_\beta}{\sqrt{\theta_1(1-\theta_1)}}}{\ln\left(\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right)}$$

$$n = \frac{\left(\frac{z_{\alpha/2}}{\sqrt{\theta_0(1-\theta_0)}} + \frac{z_\beta}{\sqrt{\theta_1(1-\theta_1)}}\right)^2}{\left(\ln\left(\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right)\right)^2}$$

### 2.2.2 Online Supplement Material

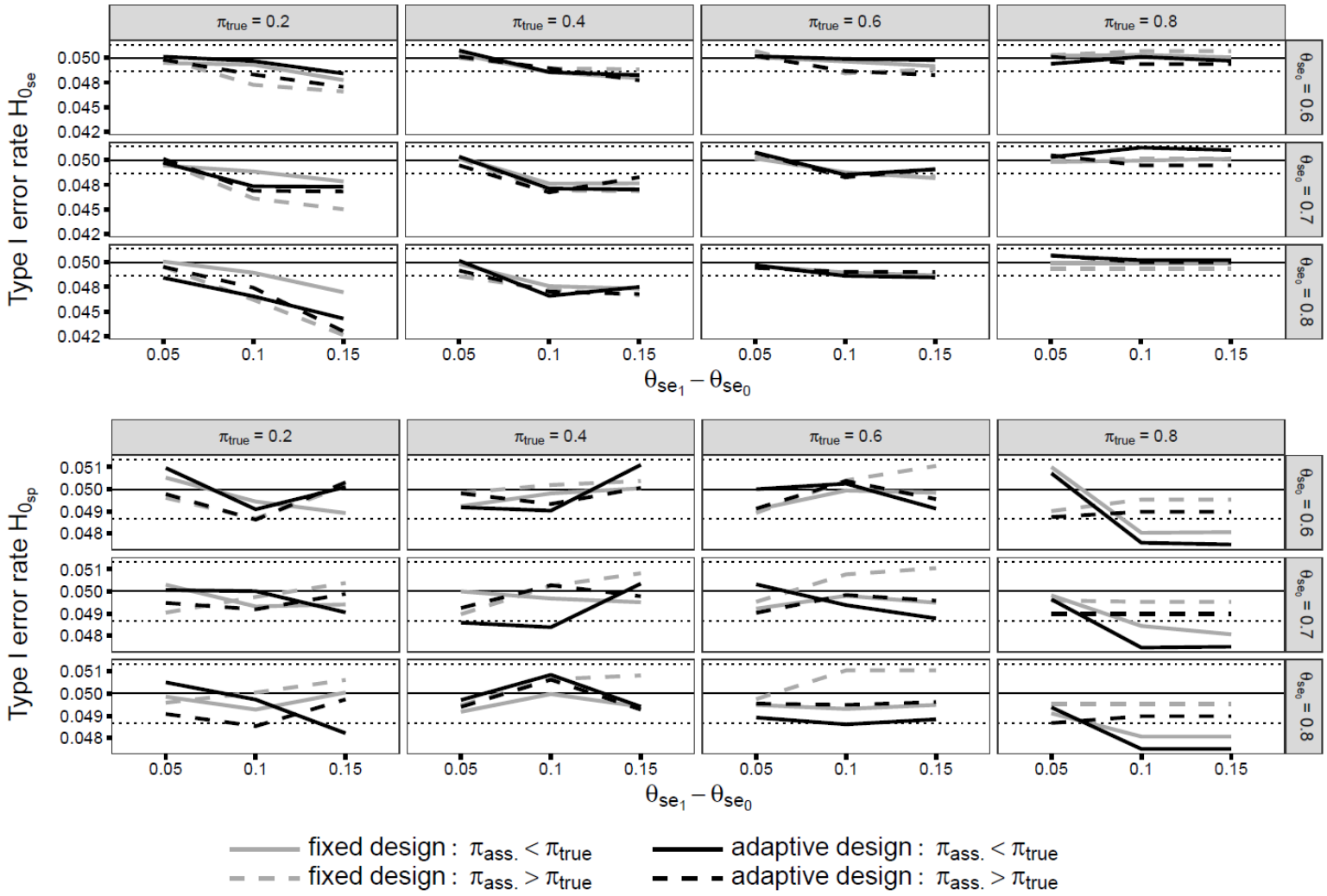
There is online supplement material available for *Thesis Article* 2 which I partially show in this section.

Online supplement material included in this section:

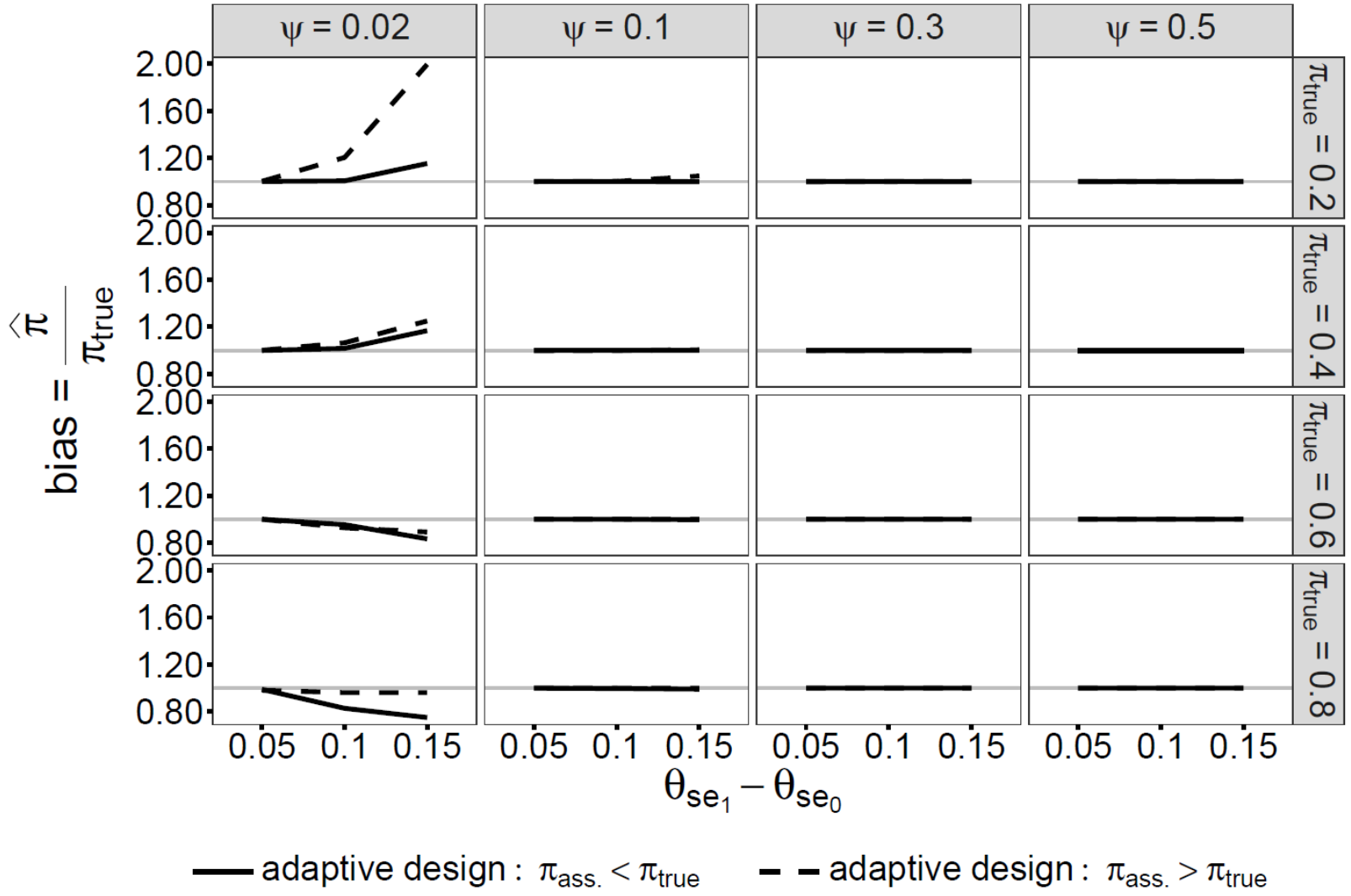
- Figure with individual type I error rates regarding endpoints of sensitivity and specificity in the fixed and adaptive single-test design
- Figure with the relative bias of the estimated prevalence
- R-code of the optimal sample size calculation in the single-test design

Further online supplement material not included in this section, but available online:

- Simulated type I error rates for all scenarios
- Simulated powers for all scenarios
- Simulated sample sizes for all scenarios




Individual type I error rate for sensitivity and specificity in the fixed and adaptive design depending on the true prevalence ($\pi_{\text{true}}$), the assumed prevalence ($\pi_{\text{ass.}}$), the minimum sensitivity ($\Theta_{se_0}$) and the sensitivity of the experimental test ($\Theta_{se_1}$).




Relative bias of the estimated prevalence ($\hat{\pi}$) in relation to the true prevalence ($\pi_{\text{true}}$) depending on the size of the internal pilot study ($\psi$), the assumed prevalence ($\pi_{\text{ass.}}$), the minimum sensitivity ($\Theta_{\text{se}_0}$) and the sensitivity of the experimental test ($\Theta_{\text{se}_1}$).

# R-Code: optimal sample size calculation

```
ss.exact <- function(alpha, power, se, se.ni, sp, sp.ni, prev) {
  # parameter description
  # alpha: desired type I error rate per endpoint
  # power: desired overall power
  # se: sensitivity of the index test
  # se.ni: sensitivity to which the index test is compared
  # sp: specificity of the index test
  # sp.ni: specificity to which the index test is compared
  # prev: prevalence

  # help functions
  # the function v estimates the variance of theta based on the binomial
      distribution
  v = function(theta){
    return(theta * (1 - theta))
  }
  # sample size for one endpoint
  single = function(alpha, beta, theta.0, theta){
    n = ceiling((qnorm(alpha / 2) * sqrt(v(theta.0)) + qnorm(beta) * sqrt(v(
        theta))) ^ 2 / (theta.0 - theta) ^ 2)
    return(n)
  }
  # calculate power for one endpoint
  calculate.power <- function(n, alpha, theta.0, theta) {
    # the function v estimates the variance of the sensitivity based on the
        binomial distribution
    v = function(theta){
      return(theta * (1 - theta))
    }
    z <- (sqrt(n) * (theta.0 - theta) - qnorm(alpha / 2) * sqrt(v(theta.0)))
        / sqrt(v(theta))
    power <- 1 - pnorm(z)
    return(power)
  }
  # function for the equal sample size for both endpoints
  f <- function(alpha, power, beta.1, se, se.ni, sp, sp.ni, prev) {
    diff.n <- qnorm(beta.1) * sqrt(v(se)) * (sp.ni - sp) * sqrt(1 - prev) -
        qnorm(1 - (power / (1 - beta.1))) * sqrt(v(sp)) * (se.ni - se) * sqrt
        (prev) - qnorm(alpha / 2) *
      sqrt(v(sp.ni)) * (se.ni - se) * sqrt(prev) + qnorm(alpha / 2) * sqrt(v(
          se.ni)) * (sp.ni - sp) * sqrt(1 - prev)
    return(diff.n)
  }
  # solve the sample size for beta.1 and then calculate beta.2
  beta.1 <- uniroot(f, alpha = alpha, power = power, se = se, se.ni = se.ni,
      sp = sp, sp.ni = sp.ni, prev = prev, lower = 0, upper = 1 - power)$root
  beta.2 <- (power + beta.1 - 1) / (beta.1 - 1)
  # calculate sample size with the known beta.1 and beta.2
  n.se <- single(alpha = alpha, beta = beta.1, theta.0 = se.ni, theta = se) #
      sample size sensitivity
```

```
  n.sp <- single(alpha = alpha, beta = beta.2, theta.0 = sp.ni, theta = sp) #
      sample size specificity
  N.se <- n.se / prev
  N.sp <- n.sp / (1 - prev)
  N <- ceiling(max(N.se, N.sp)) #total sample size
  # calculate total power
  power.se <- calculate.power(n = N * prev, alpha = alpha, theta.0 = se.ni,
      theta = se)
  power.sp <- calculate.power(n = N * (1 - prev), alpha = alpha, theta.0 = sp
      .ni, theta = sp)
  power.total <- power.se * power.sp
  return(list(N = N, N.se = N.se, N.sp = N.sp, power.total = power.total,
      power.se = power.se, power.sp = power.sp))
}
```

## 2.3 Thesis Article 3

### 2.3.1 Main Document

**Stark, M.**, Hesse, M., Brannath, W., & Zapf, A. (2022). Blinded sample size re-estimation in a comparative diagnostic accuracy study. *BMC Medical Research Methodology*, *22*, Article 115. https://doi.org/10.1186/s12874-022-01564-2

Stark *et al.*
*BMC Medical Research Methodology* (2022) 22:115
https://doi.org/10.1186/s12874-022-01564-2

BMC Medical Research Methodology

RESEARCH

Open Access

# Blinded sample size re-estimation in a comparative diagnostic accuracy study

Maria Stark[1*], Mailin Hesse[2], Werner Brannath[3] and Antonia Zapf[1]

## Abstract

**Background:** The sample size calculation in a confirmatory diagnostic accuracy study is performed for co-primary endpoints because sensitivity and specificity are considered simultaneously. The initial sample size calculation in an unpaired and paired diagnostic study is based on assumptions about, among others, the prevalence of the disease and, in the paired design, the proportion of discordant test results between the experimental and the comparator test. The choice of the power for the individual endpoints impacts the sample size and overall power. Uncertain assumptions about the nuisance parameters can additionally affect the sample size.

**Methods:** We develop an optimal sample size calculation considering co-primary endpoints to avoid an overpowered study in the unpaired and paired design. To adjust assumptions about the nuisance parameters during the study period, we introduce a blinded adaptive design for sample size re-estimation for the unpaired and the paired study design. A simulation study compares the adaptive design to the fixed design. For the paired design, the new approach is compared to an existing approach using an example study.

**Results:** Due to blinding, the adaptive design does not inflate type I error rates. The adaptive design reaches the target power and re-estimates nuisance parameters without any relevant bias. Compared to the existing approach, the proposed methods lead to a smaller sample size.

**Conclusions:** We recommend the application of the optimal sample size calculation and a blinded adaptive design in a confirmatory diagnostic accuracy study. They compensate inefficiencies of the sample size calculation and support to reach the study aim.

**Keywords:** Adaptive design, Co-primary endpoints, Sensitivity, Specificity, Unpaired design, Paired design

## Background

In a diagnostic accuracy trial the experimental test is compared to the reference standard, which defines the true disease status. Either the evaluation is limited to the comparison with the reference standard (single-test design) or another test is considered in addition (comparative design) [1]. The present article puts the focus on comparative study designs in which the experimental test is compared to an already evaluated comparator test. In the unpaired design, either the experimental test or the comparator test is assigned randomly to study participants in addition to the reference standard [2]. In contrast, in the paired design, participants undergo all three diagnostic procedures [3]. Due to the within-subject comparison of the diagnostic tests in the paired design, the variability of the study results will be diminished [4]. For this reason, the paired design is preferred to the unpaired design if technically feasible and ethically justifiable [4]. Hence, the focus of this article is especially on the paired design. Figure 1 gives an overview about the different designs.

Independent of the chosen study design, sensitivity and specificity are used as co-primary endpoints

*Correspondence: m.stark@uke.de
[1] University Medical Center Hamburg-Eppendorf, Institute of Medical Biometry and Epidemiology, Martinistr. 52, 20246 Hamburg, Germany
Full list of author information is available at the end of the article

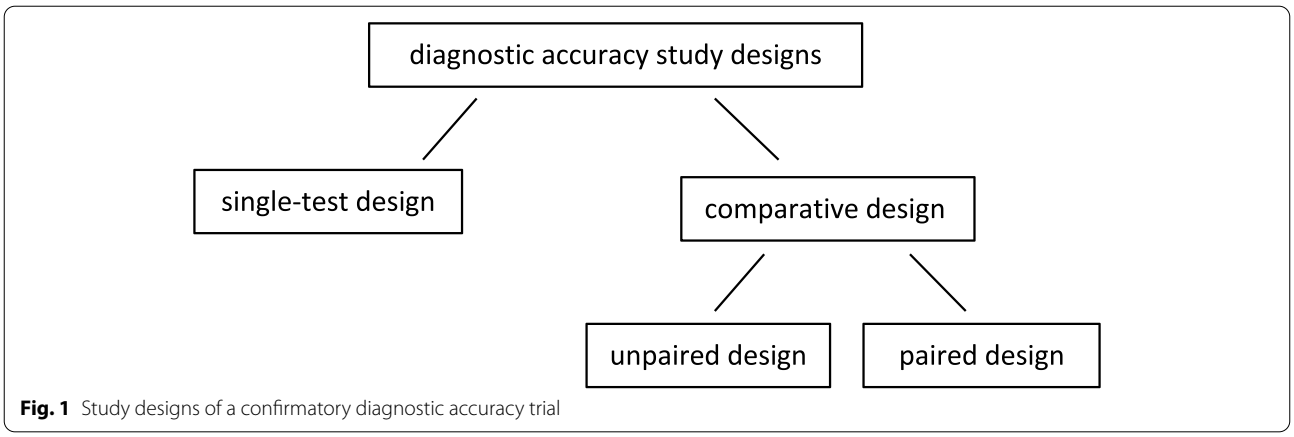



**Fig. 1** Study designs of a confirmatory diagnostic accuracy trial

in a confirmatory diagnostic accuracy trial [4, 5]. Both endpoints are combined via a joint hypothesis which is evaluated by the Intersection-Union Test [6, 7]. In this context, Stark et al. [8] developed an approach to calculate the sample size considering the prevalence. The advantage of this optimal sample size calculation is to avoid an overpowered study as it is often the case with the conventional approach. We will extend this approach to the unpaired and paired comparative study design. Hereby, the study might either aim to show superiority, non-inferiority or a combination of both regarding the co-primary endpoints.

To adjust the sample size during the course of the study, an adaptive design can be applied. Zapf et al. [9] reveal that adaptive designs including group-sequential designs are hardly developed and rarely applied in diagnostic studies. Stark et al. [8] introduce a blinded adaptive design for sample size re-estimation in the single-test design. Focusing on comparative study designs, Mazumdar et al. [10] propose a group-sequential design, but restricted to the area under the receiver operating characteristic curve as endpoint. McCray et al. [11] developed a blinded sample size re-estimation procedure in the paired study design regarding sensitivity and specificity. Their approach is based on the re-estimation of the proportion of concordant test results and the prevalence. To further develop the approaches of McCray et al. [11] and Stark et al. [8], we transfer the blinded adaptive design in the single-test design using the optimal sample size calculation to both comparative study designs. Hence, novel aspects in the present work are first, the development of the optimal sample size calculation in the unpaired as well as paired design aiming to show superiority, non-inferiority or a combination of both regarding the co-primary endpoints and second, the implementation of a blinded-sample size re-estimation procedure in the unpaired and paired design based on the optimal sample size calculation.

The present article is structured the following way: at first, we introduce the optimal sample size calculation in the unpaired and paired study design aiming to show superiority, non-inferiority or a combination of both. Second, we describe the procedure of the blinded sample size re-estimation in the unpaired and paired study design. Third, we compare the blinded adaptive design in a paired trial to the approach of McCray et al. [11] using an exemplary trial. Then, we present the results of a simulation study investigating the blinded adaptive design compared to a fixed design in an unpaired and paired study. Finally, we discuss the results and offer a conclusion.

## Methods

### Sample size calculation in a comparative diagnostic study

In this section, we introduce the optimal sample size calculation for a comparative diagnostic study, which is already developed by Stark et al. [8] for the single-test design. In a comparative diagnostic study, sensitivity and specificity of the experimental test can be tested for superiority, non-inferiority or the combination of superiority and non-inferiority against the comparator test. For the motivation and application of the optimal sample size calculation, we focus on the paired design testing for superiority regarding both endpoints because the paired design is the more relevant design in comparative studies [4]. However, the advantages of the optimal sample size calculation are also valid in the unpaired design. Furthermore, we provide formulas for the optimal approach in the unpaired and paired design.

In confirmatory diagnostic studies, sensitivity and specificity are combined as co-primary endpoints via the Intersection-Union test [8]. The null hypothesis of the Intersection-Union-Test is the union of the

individual null hypothesis regarding sensitivity and the individual null hypothesis regarding specificity [6]. The overall power of this Intersection-Union test is calculated by the product of the power of each individual hypothesis. To show superiority of the experimental test regarding sensitivity and specificity against the comparator test, the global null hypothesis $H_{0_{\text{global}}}$ for equality is given by:

$$\begin{aligned} \mathrm{H}_{0_{\mathrm{Se}}} : \mathrm{Se}_{\mathrm{E}} = \mathrm{Se}_{\mathrm{C}} \text{ and } \mathrm{H}_{0_{\mathrm{Sp}}} : \mathrm{Sp}_{\mathrm{E}} = \mathrm{Sp}_{\mathrm{C}} \\ \mathrm{H}_{0_{\text{global}}} = \mathrm{H}_{0_{\mathrm{Se}}} \cup \mathrm{H}_{0_{\mathrm{Sp}}} \end{aligned} \tag{1}$$

$\mathrm{Se}_{\mathrm{E}}$ and $\mathrm{Sp}_{\mathrm{E}}$ denote the sensitivity and specificity of the experimental test. $\mathrm{Se}_{\mathrm{C}}$ and $\mathrm{Sp}_{\mathrm{C}}$ represent the sensitivity and specificity of the comparator test. $\mathrm{H}_{0_{\text{global}}}$ is only rejected if both $\mathrm{H}_{0_{\mathrm{Se}}}$ and $\mathrm{H}_{0_{\mathrm{Sp}}}$ are rejected simultaneously. Superiority of the experimental test regarding sensitivity and specificity against the comparator test can be concluded from point estimates and $p$-values or confidence intervals. Sensitivity and specificity represent the success probabilities of a binomial distribution which follow an asymptotic normality in the case of a large sample [12]. For the analysis based on confidence intervals, we propose to use approximate $100 \cdot (1-\alpha)\%$ confidence intervals for the difference of two proportions.

### Conventional sample size calculation

To motivate the advantage of the optimal sample size calculation, we show the problems related to the procedure of the conventional sample size calculation in a confirmatory diagnostic study in the context of the paired design.

The conventional sample size calculation consists of three steps: calculate the needed number of diseased and non-diseased individuals, refer these numbers to the prevalence to receive numbers needed to show sensitivity and specificity and, choose the maximum to determine the final sample size [13–15].

We now perform these three steps for a paired diagnostic study mentioned in McCray et al. [11]. The example study compares the experimental combination of Positron Emission Tomography (PET) and computed tomography (CT) against CT alone to diagnose pancreatic cancer. The goal is to show superiority of the experimental test against the comparator test. The biopsy defines the true disease status. Table 1 shows the assumptions for sample size calculation used in this example. The disease prevalence $\pi$ represents the proportion of diseased individuals on all individuals. Parameters $\psi_{\mathrm{D}}$ and $\psi_{\mathrm{ND}}$ denote the proportion of discordant test results in the diseased and non-diseased population, hence those proportions in which both diagnostic tests lead to different test results. The conventional approach plans the sample size for each endpoint with a power of 90% which theoretically leads in the product to an overall target power of approximately 80%. The significance level $\alpha$ is set to 5% per endpoint. The $1-\alpha/2$ and $1-\beta$ quantile of the standard normal distribution is denoted by $z_{1-\alpha/2}$ and $z_{1-\beta}$. The individual steps are as follows:

1. Sample size of diseased individuals based on the formula of Miettinen et al. [16]:

$$n_{\mathrm{D}} = \frac{\left(z_{1-\alpha/2} \cdot \psi_{\mathrm{D}} + z_{1-\beta_{\mathrm{Se}}} \sqrt{\psi_{\mathrm{D}}^2 - \frac{1}{4}\left(\mathrm{Se}_{\mathrm{C}} - \mathrm{Se}_{\mathrm{E}}\right)^2 \left(3 + \psi_{\mathrm{D}}\right)}\right)^2}{\psi_{\mathrm{D}} \left(\mathrm{Se}_{\mathrm{C}} - \mathrm{Se}_{\mathrm{E}}\right)^2} = 74$$

   Sample size of non-diseased individuals:

$$n_{\mathrm{ND}} = \frac{\left(z_{1-\alpha/2} \cdot \psi_{\mathrm{ND}} + z_{1-\beta_{\mathrm{Sp}}} \sqrt{\psi_{\mathrm{ND}}^2 - \frac{1}{4}\left(\mathrm{Sp}_{\mathrm{C}} - \mathrm{Sp}_{\mathrm{E}}\right)^2 \left(3 + \psi_{\mathrm{ND}}\right)}\right)^2}{\psi_{\mathrm{ND}} \left(\mathrm{Sp}_{\mathrm{C}} - \mathrm{Sp}_{\mathrm{E}}\right)^2} = 47$$

2. Total sample size including at least $n_{\mathrm{Se}}$ diseased individuals:

$$N_{\mathrm{Se}} = \frac{n_{\mathrm{Se}}}{\pi} = \frac{74}{0.47} = 157$$

   Total sample size including at least $n_{\mathrm{Sp}}$ non-diseased individuals:

**Table 1** Assumptions of the paired diagnostic accuracy trial for the comparison of the experimental Positron Emission Tomography (PET) combined with the computed tomography (CT) against the comparator test PET

| **General input parameters:**<br>**Significance level per endpoint:** $\alpha = 0.05$ (two − sided),<br>**Overall Power:** $\text{Power}_{\text{overall}} = 1 - \beta_{\text{overall}} = 0.8$<br>**Power per endpoint:** $\text{Power}_{\mathrm{Se}} = \text{Power}_{\mathrm{Sp}} = 1 - \beta_{\mathrm{Se}} = 1 - \beta_{\mathrm{Sp}} = 0.9$ | | | |
|---|---|---|---|
| **Prevalence:**<br>$\pi = 0.47$ | **Comparator test (CT)** | **Experimental test (PET/CT)** | **Proportion of discordant test results** |
| Diseased population | $\mathrm{Se}_{\mathrm{C}} = 0.81$ | $\mathrm{Se}_{\mathrm{E}} = 0.90$ | $\psi_{\mathrm{D}} = 0.09$ |
| Non-diseased population | $\mathrm{Sp}_{\mathrm{C}} = 0.66$ | $\mathrm{Sp}_{\mathrm{E}} = 0.80$ | $\psi_{\mathrm{ND}} = 0.14$ |

$$N_{Sp} = \frac{n_{Sp}}{1-\pi} = \frac{47}{1-0.47} = 88$$

3.

$$N = \max\left(N_{Se}, N_{Sp}\right) = 157$$

The study recruits more individuals than would be necessary to show the specificity because the sensitivity determines the final sample size in this scenario. This can result in an overpowered study. If the prevalence was smaller, the difference between $N_{Se}$ and $N_{Sp}$ would be even larger. Vice versa, if the prevalence was larger, $N_{Sp}$ would determine the final sample size. These discrepancies between the sample sizes of both endpoints can result in an overpowered study. To face this problem, we propose the optimal sample size calculation explained in the next section.

### Optimal sample size calculation

At first, we present the general idea of the optimal sample size calculation. Then, we expand the optimal sample size calculation in the single-test design developed by Stark et al. [8] to an unpaired and paired study. Furthermore, we provide formulas testing for superiority regarding both endpoints in the unpaired and paired design. In additional materials, we show hypotheses and sample size formulas testing for non-inferiority or combinations of superiority and non-inferiority [see Additional file 1]. Furthermore, we offer R-Code for the optimal sample size calculation considering superiority in both endpoints in additional materials [see Additional file 2].

The general idea behind the optimal sample size calculation consists of the individual splitting of the overall power ($\text{Power}_{\text{overall}}$) to both endpoints, so that $N_{Se}$ and $N_{Sp}$ are equal. In this case, we won't need to select a maximum from both sample sizes. Consequently, the final sample size is the smallest representative sample which allows to reach the desired overall power. We calculate the final sample size with the following equation in which the symbol " $\stackrel{!}{=}$ " denotes that terms on both sides must be equal:

$$N_{Se} \stackrel{!}{=} N_{Sp} \tag{2}$$

$$\frac{n_{Se}}{\pi} \stackrel{!}{=} \frac{n_{Sp}}{1-\pi} \tag{3}$$

Under the condition:

$$\text{Power}_{Se} \cdot \text{Power}_{Sp} = \text{Power}_{\text{overall}} \tag{4}$$

$$(1-\beta_{Se}) \cdot \left(1-\beta_{Sp}\right) = \text{Power}_{\text{overall}} \tag{5}$$

$$\beta_{Sp} = \frac{1-\beta_{Se}-\text{Power}_{\text{overall}}}{1-\beta_{Se}} = 1 - \frac{\text{Power}_{\text{overall}}}{1-\beta_{Se}} \tag{6}$$

In the following subsections, we plug the condition into the sample size calculation; noting that the resulting equations cannot be solved analytically respect to $\beta_{Se}$.

### Unpaired design

In the unpaired design, the optimal sample size calculation uses the formula for the comparison of two independent proportions following Zhou et al. [1]:

$$\frac{\left(z_{\alpha/2}\sqrt{V_0(\text{Se}_C-\text{Se}_E)} + z_{\beta_{Se}}\sqrt{V_A(\text{Se}_C-\text{Se}_E)}\right)^2}{(\text{Se}_C-\text{Se}_E)^2 \cdot \pi} \stackrel{!}{=}$$

$$\frac{\left(z_{\alpha/2}\sqrt{V_0(\text{Sp}_C-\text{Sp}_E)} + z_{\frac{1-\beta_{Se}-\text{Power}_{\text{overall}}}{1-\beta_{Se}}}\sqrt{V_A(\text{Sp}_C-\text{Sp}_E)}\right)^2}{(\text{Sp}_C-\text{Sp}_E)^2 \cdot (1-\pi)} \tag{7}$$

where $V_0(\text{Se}_C-\text{Se}_E)$ and $V_A(\text{Se}_C-\text{Se}_E)$ represent the variance of the difference between $\text{Se}_C$ and $\text{Se}_E$ under the null and alternative hypothesis, respectively. In the unpaired design, the variance $V(\text{Se}_C-\text{Se}_E)$ is defined as [1]:

$$V(\text{Se}_C-\text{Se}_E) = \text{Se}_C \cdot (1-\text{Se}_C) + \text{Se}_E \cdot (1-\text{Se}_E) \tag{8}$$

The variance $V(\text{Sp}_C-\text{Sp}_E)$ is calculated in analogy.

Although the sample size formula in Eq. (7) fits to the Wald confidence interval for the difference of two independent proportions, we propose to analyse the unpaired design with the two-sided 1- α Score confidence interval for the difference of two independent proportions [17]. The coverage probability of the Score confidence interval is closer to the nominal level compared to the Wald confidence interval [18–20].

### Paired design

In the paired design, the optimal sample size is based on the formula of Miettinen et al. [16]:

$$\frac{\left(z_{1-\alpha/2}\cdot\psi_D + z_{1-\beta_{Se}}\sqrt{\psi_D^2 - \frac{1}{4}(\text{Se}_C-\text{Se}_E)^2(3+\psi_D)}\right)^2}{\psi_D(\text{Se}_C-\text{Se}_E)^2\pi} \stackrel{!}{=}$$

$$\frac{\left(z_{1-\alpha/2}\cdot\psi_{ND} + z_{\frac{\text{Power}_{\text{overall}}}{1-\beta_{Se}}}\sqrt{\psi_{ND}^2 - \frac{1}{4}(\text{Sp}_C-\text{Sp}_E)^2(3+\psi_{ND})}\right)^2}{\psi_{ND}(\text{Sp}_C-\text{Sp}_E)^2(1-\pi)} \tag{9}$$

with $\psi_D$ as the proportion of discordant test results in the diseased sample, which varies between [16, 21]:

$$|\text{Se}_C - \text{Se}_E| \leq \psi_D \leq \text{Se}_C + \text{Se}_E - 2\cdot\text{Se}_C\cdot\text{Se}_E \tag{10}$$

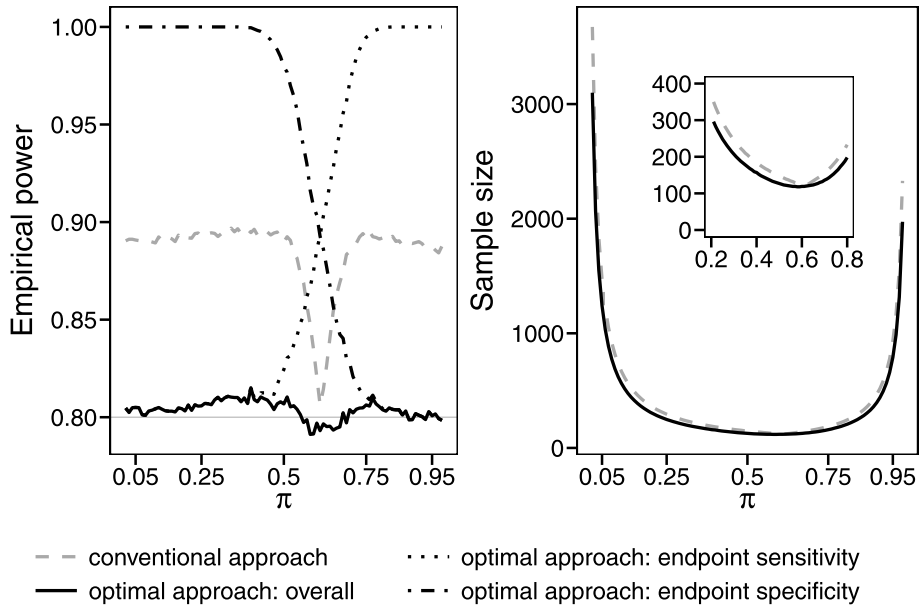



**Fig. 2** Empirical power and sample size of the conventional and optimal sample size calculation. Simulations are based on the example study given in Table 1 with a varying prevalence $\pi$. The figure considering the sample size contains an enlarged image section so that the differences between both approaches are highlighted

The interval of the proportion of discordant test results in the non-diseased sample $\psi_{ND}$ is calculated in analogy by considering $Sp_C$ and $Sp_E$.

For two different proportions of discordant test results in the diseased $(\psi_{D_1}, \psi_{D_2})$ and non-diseased $(\psi_{ND_1}, \psi_{ND_2})$ population, the total sample size $N(\psi_D, \psi_{ND})$ in Eq. (9) is monotone increasing:

$$\begin{gathered}\psi_{D_1}, \psi_{D_2} \in \left[\left|Se_C - Se_E\right|; Se_C + Se_E - 2 \cdot Se_C \cdot Se_E\right] \text{ and} \\ \psi_{ND_1}, \psi_{ND_2} \in \left[\left|Sp_C - Sp_E\right|; Sp_C + Sp_E - 2 \cdot Sp_C \cdot Sp_E\right] \\ \psi_{D_1} \leq \psi_{D_2} \text{and } \psi_{ND_1} \leq \psi_{ND_2} \Rightarrow N(\psi_{D_1}, \psi_{ND_1}) \leq N(\psi_{D_2}, \psi_{ND_2})\end{gathered} \tag{11}$$

In analogy to the unpaired design, we propose to analyse the paired design with the two-sided 1- α Tango's asymptotic score confidence interval for the difference of two matched proportions [22, 23]. We recommend this based on the reason given above. Furthermore, the Wald confidence is not range preserving [24].

### Application of the optimal sample size calculation in the paired design

We apply the optimal sample size approach to the example study introduced in Table 1 and compare the results to those of the conventional approach. For this purpose, we simulate, based on 10,000 simulation runs, the empirical power of both approaches for a varying prevalence $\pi$ and calculate the sample size. Figure 2 shows the results. In most cases, the conventional approach is highly overpowered due to the choice of the maximum sample size of both endpoints in the third step. If the prevalence is in the range between 0.5 and 0.75, the empirical power will be closer to the target power of 80%. The empirical power will be the closest to the target power, if the prevalence equals 0.6 as the discrepancy between $N_{Se}$ and $N_{Sp}$ is the smallest.

The optimal approach splits the overall power to both endpoints depending on the prevalence, so that the product of the empirical power of both endpoints comes close to the target power of 80%.

Considering the sample size, the optimal approach will lead to a smaller sample size than the conventional approach if the prevalence is unbalanced. Figure 2 contains an enlarged image section of the sample size so that the differences between both approaches are highlighted.

### Blinded sample size re-estimation

The procedure of a blinded sample size adjustment based on the re-estimation of nuisance parameters basically follows five phases named by Stark et al. [8]. In Fig. 3, these five steps are explained in context of the unpaired and paired study design. The nuisance parameters re-estimated during the study are the prevalence and additionally proportions of discordant test results in the paired design. The main difference between the adaptive designs in the unpaired and paired study design consists of the sample size for the interim analysis. In the unpaired design, the prevalence is estimated based on 50% of the initially calculated sample size. In the paired design, both, the initial sample size and the sample size for the interim

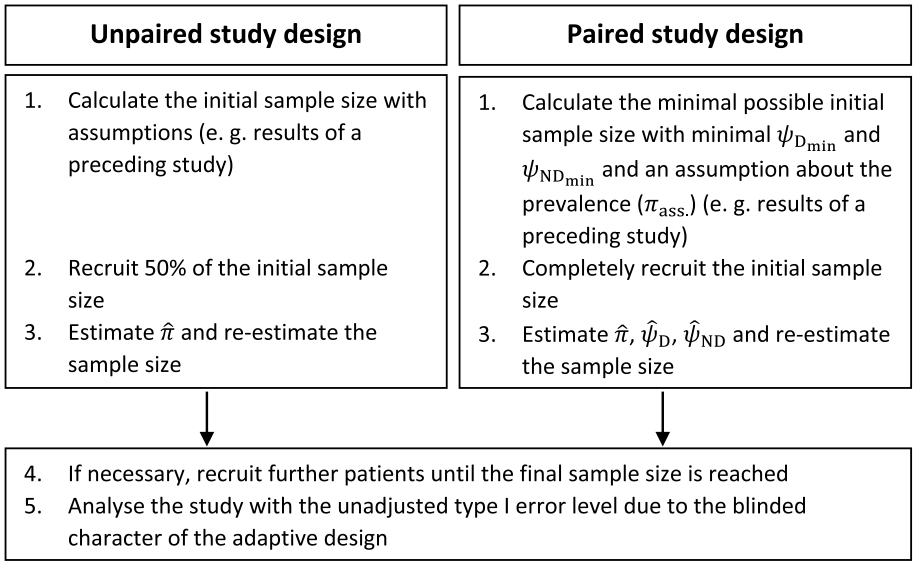



**Fig. 3** Procedure of the blinded adaptive design in an unpaired and paired diagnostic trial

analysis equal the minimal sample size [11]. The minimal sample size is received with the minimal possible proportion of discordant test results in the diseased ($\psi_{D_{min}}$) and non-diseased population ($\psi_{ND_{min}}$). Assumptions about the sensitivity and the specificity of the comparator and experimental test determine the minimal possible proportion of discordant test results. Following Eq. (10), the minimal proportion of discordant test results are calculated with:

$$\begin{aligned} \psi_{D_{min}} &= |Se_C - Se_E| \\ \psi_{ND_{min}} &= |Sp_C - Sp_E| \end{aligned} \tag{12}$$

Furthermore, the calculation of the minimal sample size requires assumptions about the prevalence.

During interim analysis, the prevalence is estimated by the maximum likelihood estimator of a binomial proportion [25]:

$$\hat{\pi} = \frac{n_D}{n} \tag{13}$$

The number of diseased individuals involved in the interim analysis is represented by $n_D$, and the sample size used for interim analysis is denoted by $n$.

In analogy, the proportion of discordant test results is estimated by the maximum likelihood estimator of a multinomial distribution [26]:

$$\hat{\psi}_D = \frac{n_{D10} + n_{D01}}{n_D} \tag{14}$$

**Table 2** Results in a paired diagnostic study

| **Diseased $n_D$** | | | |
|---|---|---|---|
| | | **Comparator Test** | |
| | | **True Positive ($TP_C$)** | **False Negative ($FN_C$)** |
| **Experimental Test** | **True Positive ($TP_E$)** | $n_{D_{11}}$ | $n_{D_{10}}$ |
| | **False Negative ($FN_E$)** | $n_{D_{01}}$ | $n_{D_{00}}$ |
| **Non-diseased $n_{ND}$** | | | |
| | | **Comparator Test** | |
| | | **False Positive ($FP_C$)** | **True Negative ($TN_C$)** |
| **Experimental Test** | **False Positive ($FP_E$)** | $n_{ND_{11}}$ | $n_{ND_{10}}$ |
| | **True Negative ($TN_E$)** | $n_{ND_{01}}$ | $n_{ND_{00}}$ |

$$\hat{\psi}_{ND} = \frac{n_{ND10} + n_{ND01}}{n_{ND}} \tag{15}$$

Table 2 shows the parameters needed to re-estimate the proportions of discordant test results.

The estimation of nuisance parameters represents a blinded adaptive design because the sensitivity and the specificity of the experimental test are not revealed. Hence, the type I error rate will not be inflated by definition.

**Table 3** Comparison of the blinded adaptive design procedure with McCray et al. [11]

| | | McCray et al. (2017) | Our approach |
|---|---|---|---|
| General information | **Endpoint** | $\frac{Se_E}{Se_C}$ and $\frac{Sp_E}{Sp_C}$ | $Se_E - Se_C$ and $Sp_E - Sp_C$ |
| | **$H_{0_{global}}$** | $H_{0_{Se}} : \frac{Se_E}{Se_C} = 1 \cup$<br>$H_{0_{Sp}} : \frac{Sp_E}{Sp_C} = 1$ | $H_{0_{Se}} : Se_E - Se_C = 0 \cup$<br>$H_{0_{Sp}} : Sp_E - Sp_C = 0$ |
| | **Sample size calculation** | Conventional approach<br>$\alpha$ per endpoint: 0.05 (two-sided)<br>Power per endpoint: 0.8 | Optimal approach<br>$\alpha$ per endpoint: 0.05 (two-sided)<br>Overall power: 0.8 |
| | **Parameter of dependency between both tests** | $TPPR = \frac{n_{D11}}{n_D}$<br>$TNNR = \frac{n_{ND00}}{n_{ND}}$ | $\psi_D = \frac{n_{D10}+n_{D01}}{n_D}$<br>$\psi_{ND} = \frac{n_{ND10}+n_{ND01}}{n_{ND}}$ |
| Initial sample size calculation | **Size of internal pilot study** | $TPPR_{max}$ and $TNNR_{max}$ correspond to $\psi_{D_{min}}$ and $\psi_{ND_{min}}$ | |
| | **Parameter of dependency between both tests for initial sample size calculation** | $TPPR_{max} = Se_C = 0.81$<br>$TNNR_{max} = Sp_C = 0.66$ | $\psi_{D_{min}} = \lvert Se_C - Se_E \rvert = 0.09$<br>$\psi_{ND_{min}} = \lvert Sp_C - Sp_E \rvert = 0.14$ |
| | **Initial sample size, size of internal pilot study** | 186 | 133 |
| Sample size re-estimation | **Estimation of nuisance parameters** | $\hat{\pi} = 0.44$<br>$\hat{TPPR} = 0.80$ \|<br>$\hat{TNNR} = 0.66$ | $\hat{\pi} = 0.44$<br>$\hat{\psi}_D = 0.11$ \|<br>$\hat{\psi}_{ND} = 0.14$ |
| | **Re-estimated sample size** | 242 | 200 |

## Results

### Application of the blinded sample size re-estimation in the example study

This section serves for illustration of the blinded sample size re-estimation in the paired study design. For this purpose, we compare the approach of McCray et al. [11] to the adaptive design procedure described in this article by taking up the example of a paired diagnostic accuracy study already introduced in Table 1. The main progress of our new approach compared to McCray et al. [11] is to implement the optimal sample size calculation. We reveal the advantage of the optimal sample size calculation in this context again.

Table 3 compares the theoretical aspects and the results of both adaptive design procedures. They differ in the definition of endpoints, hypothesis and in the way the sample size calculation is performed. McCray et al. [11] work with the quotient of sensitivities and the quotient of specificities of both diagnostic tests as endpoints. They use sample size formulas which rely on the true-positive-positive rate (TPPR) and true-negative-negative-rate (TNNR) [27]. TPPR denotes the proportion of test results in which both, the comparator test and the experimental test correctly diagnose a diseased individual. Vice versa, TNNR represents the proportion of test results in which both tests correctly return a negative test result. For initial sample size calculation, $TPPR_{max}$ and $TNNR_{max}$ are used, which represent the maximal possible TPPR and TNNR, respectively.

McCray et al. [11] perform the sample size calculation based on the conventional three steps by planning the sample size calculation with a power of 80% per endpoint. This leads to a theoretical overall power of 64%.

In contrast to McCray et al. [11], our approach uses the optimal sample size calculation. It is based on sample size formulas considering the difference of sensitivities and the proportion of discordant test results in the diseased population or the difference of specificities of both tests and the proportion of discordant test results in the non-diseased population, respectively [1]. In contrast to McCray et al. [11], we choose the differences as endpoint measurement because the guideline on clinical evaluation of diagnostic agents suggests this [4]. Furthermore, we perform the optimal sample size calculation to reach an overall power of 80%.

Table 3 shows the initial sample size, the sample size for interim analysis and the re-estimated sample size of both adaptive design procedures. Due to the optimal approach, sample sizes resulting from our adaptive design are lower than those of McCray et al. [11]. The optimal sample size calculation avoids that one of both co-primary endpoints is overpowered which leads to smaller sample sizes.

The difference between both approaches regarding sample sizes will be even more extensive if the prevalence is unbalanced. A figure in additional materials, which depicts the simulated empirical overall power based on 10,000 simulations runs and the calculated sample size, illustrates this difference between both approaches for

**Table 4** Simulated scenarios in the unpaired and paired study design testing for superiority in both endpoints. The proportion of discordant test results is only relevant in the paired design

| | **10,000 simulation runs per scenario** | |
|---|---|---|
| Nominal significance level α per endpoint | 0.05 (two-sided) | |
| Nominal overall target power | 0.8 | |
| | **Initial scenario** | **Variation of initial scenario** |
| Sensitivity comparator test $Se_C$ | 0.8 | 0.6, 0.7 |
| Specificity comparator test $Sp_C$ | 0.7 | 0.6, 0.8 |
| True prevalence $\boldsymbol{\pi}_{true}$ | 0.2 | 0.4, 0.6, 0.8 |
| Assumed prevalence $\boldsymbol{\pi}_{\mathbf{ass.}}$ | $\pi_{true} + 0.1$ | $\pi_{true}$ - 0.1<br>$\pi_{true} + 0.2$<br>$\pi_{true} + 0.3$ |
| True discordant results diseased population $\psi_{D_{true}}$ | 0.11<br>(0.15, if:<br>$Se_E - Se_C = 0.15$) | 0.18, 0.26 |
| Assumed discordant results diseased population $\psi_{D_{ass.}}$ | 0.18 | |
| True discordant results non-diseased population $\psi_{ND_{true}}$ | 0.14<br>(0.15, if:<br>$Sp_E - Sp_C = 0.15$) | 0.24, 0.38 |
| Assumed discordant results non-diseased population $\psi_{ND_{ass.}}$ | 0.24 | |
| Sensitivity experimental test $Se_E$ | $\hat{=}Se_C$ | |
| Specificity experimental test $Sp_E$ | $\hat{=}Sp_C$ | |
| Sensitivity experimental test $Se_E$ | $Se_C + 0.1$ | $Se_C + 0.05$<br>$Se_C + 0.15$ |
| Specificity experimental test $Sp_E$ | $Sp_C + 0.1$ | $Sp_C + 0.05$<br>$Sp_C + 0.15$ |

the initial sample size calculation based on $\psi_{D_{min}}$ and $\psi_{ND_{min}}$ by varying $\pi$ [see Additional file 3]. This figure reveals that the approach of McCray et al [11]. is highly overpowered although they plan with a power of 80% per endpoint. This theoretically leads to a theoretical overall power of 64%. In this example, the dependence between both diagnostic tests is almost maximal because $\psi_D$ and $\psi_{ND}$ are almost minimal. In this case, the underlying assumptions of sample size formulas and confidence intervals are not valid [11]. Hence, the approach of McCray et al. [11] is highly overpowered.

In contrast, the optimal sample size calculation enables to reach an overall power of 80% independent of the prevalence.

### Simulation study

We perform a simulation study to evaluate type I error rates, statistical power, sample sizes and bias of the adaptive design based on re-estimated nuisance parameters in the unpaired and paired study design. We compare results of the adaptive design to those of the fixed design which gets by without re-estimation of the sample size. Table 4 shows the simulated scenarios testing for superiority in both endpoints. Based on the example of a paired diagnostic accuracy study used by McCray et al. [11], we choose one initial scenario. Starting from the initial scenario, we vary one parameter in each further scenario. That results in 15 scenarios in the unpaired design and 19 scenarios in the paired design, each simulated with 10,000 simulation runs. In analogy to these scenarios, we perform simulations testing for non-inferiority in both endpoints, or the combinations of superiority and non-inferiority, respectively. In this section, we focus on the results of those scenarios testing for superiority in both endpoints because the other results are comparable to them. For completeness, we make the remaining simulated scenarios and their results available in the online supplement materials [see Additional files 4 and 5].

Table 5 shows distributions involved in the data generation mechanism. We use the statistical software *R* version 4.0.5 to perform the simulations with the default random number generator Mersenne-Twister, but with the own initialization methods of *R* [28, 29].

Figure 4 shows type I error rates with according Monte Carlo errors due to simulations (1.96 x SE = 0.00098), power and true sample sizes ($N_{true}$) with root-mean-squared-error of the re-estimated sample size (RMSE) under $H_1$ and additionally the mean of the re-estimated samples sizes per scenario ($N_{mean}$) of those scenarios containing the minimal, medium and maximal $\psi_{D_{true}}$ in the paired study design**.** The depicted

**Table 5** Description of the data generation mechanism of the unpaired and paired design in the simulation study (*Bin*: binomial distribution, *MVBin*: multivariate binomial distribution, *k*: number of trials, *p*: success probability, *ρ*: dependence between both tests, *N*: total sample size, $n_{D_E}$: diseased individuals in experimental group, $n_{D_C}$: diseased individuals in comparator group)

| | **Unpaired design** | **Paired design** |
|---|---|---|
| Diseased individuals ($n_D$) according to reference standard | $n_{D_E} \sim Bin(k = N, p = \pi_{true})$<br>$n_{D_C} \sim Bin(k = N, p = \pi_{true})$ | $n_D \sim Bin(k = N, p = \pi_{true})$ |
| True Positive Results (TP) | $TP_E \sim Bin(k = n_{D_E}, p = Se_E)$<br>$TP_C \sim Bin(k = n_{D_C}, p = Se_C)$ | $(TP_E, TP_C) \sim MVBin(k_E = n_{D_E}, k_C = n_{D_C}, p_E = Se_E, p_C = Se_C, \rho = TPPR)$ |
| True Negative Results (TN) | $TN_E \sim Bin(k = N - n_{D_E}, p = Sp_E)$<br>$TN_C \sim Bin(k = N - n_{D_C}, p = Sp_C)$ | $(TN_E, TN_C) \sim MVBin(k_E = N - n_{D_E}, k_C = N - n_{D_C}, p_E = Sp_E, p_C = Sp_C, \rho = TNNR)$ |

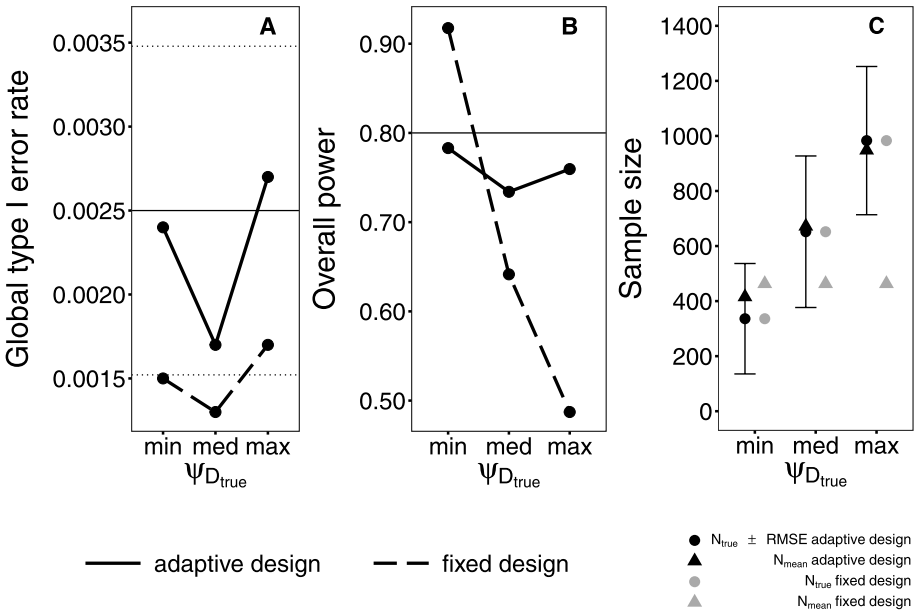



**Fig. 4** Global type I error, overall power and sample sizes of the fixed and adaptive paired design. Simulations are based on the initial scenario and a variation of the true proportion of discordant test results in the diseased population ($\psi_{D_{true}}$). In Fig. A, black dotted lines mark the interval of Monte Carlo error due to simulations. In Fig. B, the target power equals 0.8

results offer some characteristics which can be generalized to other scenarios in the paired and unpaired design. Referring to Fig. A, one important aspect is that scenarios preserve type I error rates. In analogy to the overall power of the Intersection-Union Test explained in section 2, global type I error rates result as the product of the individual type I error rates of each endpoint (0.05 two-sided each). Due to the analysis with the score confidence interval in this scenario with small prevalence, results are conservative [24].

Considering Fig. B and C, the overall power of the fixed design decreases with increasing $\psi_{D_{true}}$. The larger $\psi_{D_{true}}$ is, the smaller the dependence between both tests is. The smaller the dependence between both tests is, the larger $N_{true}$ becomes. The discrepancy between $N_{true}$ and $N_{mean}$ in the fixed design increases, if $\psi_{D_{true}}$ increases. If $\psi_{D_{true}}$ is medium, the assumption about this parameter in the fixed design equals the true parameter. But the assumption about the prevalence is larger than the prevalence is in truth. Therefore, $N_{mean}$ is smaller than $N_{true}$ and the overall power is smaller than the target power of 80%.

The adaptive design compensates wrong assumptions about nuisance parameters. The discrepancy between $N_{true}$ and $N_{mean}$ of the adaptive design is small. Hence, the overall power comes close to the target power. The adaptive design re-estimates $\psi_{D_{true}}$, $\psi_{ND_{true}}$ and $\pi_{true}$ without any relevant bias. In those scenarios based on the initial prevalence of 20%, relative bias of $\hat{\psi}_D$ is little higher than relative bias of $\hat{\psi}_{ND}$. Due to this prevalence, there is only a small number of diseased patients in the sample which can be consulted for the re-estimation of $\psi_{D_{true}}$. Supplement materials show simulations results of the bias.

Figure 5 compares the overall power depending on the true prevalence $\pi_{true}$ in the unpaired and paired design. If $\pi_{true}$ is low, the power in both fixed designs is the lowest. The power becomes larger with increasing prevalence. In the depicted scenarios, the assumed prevalence is larger than the true prevalence. A low true prevalence represents a small number of diseased individuals. In this case, the number of diseased individuals is the determining aspect for sample size calculation to show the sensitivity. In the fixed unpaired design, a higher number of diseased individuals is wrongly assumed which results in a too small sample size and power. Vice versa, a high true prevalence leads to a too large sample size and power. The number of non-diseased individuals now determines the sample size to show the specificity. Due to the wrongly assumed prevalence, a too small number of non-diseased individuals is expected. The sample size is calculated too large. The fixed paired design is highly overpowered, independent of $\pi_{true}$. Both proportions of discordant test results are assumed higher than in truth. The sample size is calculated too large.

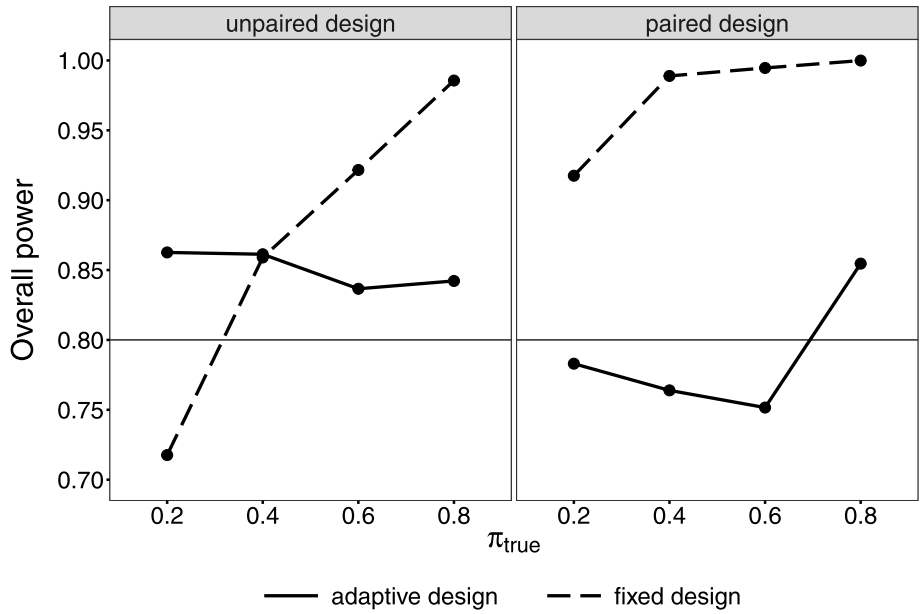



**Fig. 5** Overall power of the fixed and adaptive design in an unpaired and paired diagnostic study. Simulations are based on the initial scenario and a variation of the true prevalence ($\pi_{true}$). The target power equals 0.8

In contrast to the fixed designs, both adaptive designs reveal a power closer to the target power of 80%. If $\pi_{true}$ equals 80%, the overall power of the adaptive paired design stands out. In this scenario, the proportion of non-diseased individuals is initially assumed smaller than in truth. Hence, the sample size used for the re-estimation of nuisance parameters is already larger than the true sample size. The overall power is higher compared to scenarios with a lower $\pi_{true}$.

## Discussion

In this article, we present an approach for blinded sample size re-estimation in a comparative diagnostic accuracy study. This allows the sample size to be revised for incorrect assumptions during the course of the study, so that the study is neither over- nor underpowered. We use an example and simulation study to show that the approach does not inflate type I error rates, reach the target power and re-estimate nuisance parameters without any relevant bias.

One strength of our simulation study is that it is based on a realistic initial scenario. Therefore, the simulation study covers the results of realistic as well as of extreme parameter combinations. But of course the simulation study does not depict all possible parameter combinations.

One general weakness of our proposed approach is that the sample size calculation and the confidence intervals used for evaluation are not based on the same formulas.

McCray et al. [11] use a sample size calculation and an evaluation method which belong together. Due to different endpoints in the approach of McCray et al. [11] and our approach, we don't compare both approaches within an extensive simulation study. However, we compare both approaches within the example study. We show that our approach requires a smaller sample size and comes closer to the target power than the approach of McCray et al. [11], if the dependence between both diagnostic tests is maximal. In contrast to our work, McCray et al. [11] do not extend their approach to show non-inferiority or a combination of superiority and non-inferiority in both diagnostic tests.

We recommend to apply blinded adaptive designs in comparative diagnostic accuracy studies, especially if the nuisance parameters are extremely small or large. The reason for this is that a blinded adaptive design can correct extremely small or large sample sizes based on wrong assumptions.

Our work creates some space for further research. One important unanswered question asks about the consequences of the re-estimation of the prevalence on the blinding if predictive values are chosen as co-primary endpoints. Both, the positive and negative predictive value depend on the prevalence. Hence, the analysis is not blinded in the strong sense. Furthermore, it is of interest to develop unblinded adaptive designs in comparative diagnostic accuracy studies to allow for early stopping due to futility or efficacy [9].

## Conclusions

A confirmatory diagnostic accuracy study can either be performed as a single-test or a comparative study design. Comparative study designs are distinguished between an unpaired and paired study design. Stark et al. [8] introduce the optimal sample size calculation and the blinded adaptive design to re-estimate the sample size in the single-test design. This approach avoids an overpowered diagnostic accuracy study by calculating the sample size for two co-primary endpoints sensitivity and specificity in dependence of the prevalence of the disease.

In this article, we transfer the optimal sample size calculation to both comparative study designs. Furthermore, we propose blinded adaptive designs for an unpaired and paired diagnostic accuracy study. In the unpaired design, the adaptive design re-estimates the prevalence whereas, in the paired design, it additionally re-estimates the proportions of discordant test results. Subsequent to the re-estimation of these nuisance parameters, the sample size is re-calculated. Due to the blinded character of the adaptive designs, type I error rates are not inflated. Both approaches reach the target power and re-estimate nuisance parameters without any relevant bias.

We recommend to apply the optimal sample size calculation and a blinded adaptive design in a confirmatory diagnostic accuracy trial. Both approaches support to calculate the necessary sample size to achieve the targeted power without much additional effort.

#### Abbreviations

Bin: Binomial distribution; CT: Computed Tomography; MVBin: Multivariate Binomial distribution; PET: Positron-Emission Tomography; RMSE: Root-Mean-Squared-Error; $Se_C$: Sensitivity of the comparator test; $Se_E$: Sensitivity of the experimental test; $Sp_C$: Specificity of the comparator test; $Sp_E$: Specificity of the experimental test; TNNR: True-Negative-Negative-Rate; TPPR: True-Positive-Positive-Rate.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01564-2.

**Additional file 1.** Formulas for the optimal sample size calculation.

**Additional file 2.** R-Code for the optimal sample size calculation testing for superiority in both endpoints in the unpaired and paired design.

**Additional file 3** Figure containing the comparison of the optimal sample size calculation with the approach of McCray et al. [11].

**Additional file 4.** Simulation results of the blinded sample size re-estimation in the unpaired design.

**Additional file 5.** Simulation results of the blinded sample size re-estimation in the paired design.

#### Acknowledgments

We acknowledge the German Research Foundation for financing the project" Flexible designs for diagnostic studies" to which this article belongs (ZA 687/1-1).

#### Authors' contributions

All authors read and approved the final version of the manuscript. Their specific contributions are as follows: MS implemented the statistical methods, wrote the initial and final drafts of the manuscript and revised the manuscript for important intellectual content. MH provided R-Code for the simulation study in the adaptive unpaired design. MH and WB critically reviewed and commented the draft of the manuscript and made intellectual contribution to its content. AZ provided the idea for the content of the manuscript and the overall supervision and administration for this project; critically reviewed and commented on multiple drafts of the manuscript and made intellectual contribution to its content.

#### Funding

Open Access funding enabled and organized by Projekt DEAL. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article is supported by the Deutsche Forschungsgemeinschaft (ZA 687/1–1).

#### Availability of data and materials

All simulations results used to illustrate the method can be found in online additional material of this article. This additional material is available online for the article:
- Additional file 1 ("Additional_file_1_pdf): Formulas for the optimal sample size calculation
- Additional file 2 ("Additional_file_2.pdf"): R-Code for the optimal sample size calculation testing for superiority in both endpoints in the unpaired and paired design
- Additional file 3 ("Additional_file_3.pdf"): Figure containing the comparison of the optimal sample size calculation with the approach of McCray et al. [11]
- Additional file 4 ("Additional_file_4.pdf"): Simulation results of the blinded sample size re-estimation in the unpaired design
- Additional file 5 ("Additional_file_5.pdf"): Simulation results of the blinded sample size re-estimation in the paired design

## Declarations

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

#### Author details

[1]University Medical Center Hamburg-Eppendorf, Institute of Medical Biometry and Epidemiology, Martinistr. 52, 20246 Hamburg, Germany. [2]Abbott GmbH, Wiesbaden, Germany. [3]University of Bremen, Institute of Statistics, Bremen, Germany.

Received: 13 December 2021 Accepted: 28 February 2022
Published online: 19 April 2022

#### References

1. Zhou X-H, McClish DK, Obuchowski NA. Statistical methods in diagnostic medicine, vol. 569. 2nd ed. Hoboken: John Wiley & Sons; 2011.
2. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2003.
3. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332:1089–92.
4. Committee for Medicinal Products for Human Use (CHMP). Guideline on clinical evaluation of diagnostic agents. London: European Medicines Agency, https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-evaluation-diagnostic-agents_en.pdf. Accessed 21 March 2021.

5. U.S. Food and Drug Administration (FDA). Guidance for industry and FDA staff: statistical guidance on reporting results from studies evaluating diagnostic tests. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/statistical-guidance-reporting-results-studies-evaluating-diagnostic-tests-guidance-industry-and-fda. Accessed 21 March 2021.
6. Hamasaki T, Evans SR, Asakura K. Design, data monitoring, and analysis of clinical trials with co-primary endpoints: a review. J Biopharm Stat. 2018;28:28–51.
7. Korevaar DA, Gopalakrishna G, Cohen JF, Bossuyt PM. Targeted test evaluation: a framework for designing diagnostic accuracy studies with clear study hypotheses. Diagn Prognostic Res. 2019;3:1–10.
8. Stark M, Zapf A. Sample size calculation and re-estimation based on the prevalence in a single-arm confirmatory diagnostic accuracy study. Stat Methods Med Res. 2020;29:2958–71.
9. Zapf A, Stark M, Gerke O, Ehret C, Benda N, Bossuyt P, et al. Adaptive trial designs in diagnostic accuracy research. Stat Med. 2020;39:591–601.
10. Mazumdar M, Liu A. Group sequential design for comparative diagnostic accuracy studies. Stat Med. 2003;22:727–39.
11. McCray GP, Titman AC, Ghaneh P, Lancaster GA. Sample size re-estimation in paired comparative diagnostic accuracy studies with a binary response. BMC Med Res Methodol. 2017;17:102–13.
12. Thomopoulos NT. Statistical distributions. Cham: Springer International Publishing; 2017.
13. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biopharm Inform. 2014;48:193–204.
14. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. J Clin Epidemiol. 2005;58:859–62.
15. Buderer NM. Statistical methodology: I. incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. Acad Emerg Med. 1996;3:895–900.
16. Miettinen OS. The matched pairs design in the case of all-or-none responses. Biometrics. 1968;24:339–52.
17. Miettinen O, Nurminen M. Comparative analysis of two rates. Stat Med. 1985;4:213–26.
18. Agresti A. Categorical data analysis, vol. 482. 3rd ed. Hoboken: John Wiley & Sons; 2013.
19. Agresti A, Caffo B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Am Stat. 2000;54:280–8.
20. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat. 1998;52:119–26.
21. Connor RJ. Sample size for testing differences in proportions for the paired-sample design. Biometrics. 1987;43:207–11.
22. Agresti A, Min Y. Simple improved confidence intervals for comparing matched proportions. Stat Med. 2005;24:729–40.
23. Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Stat Med. 1998;17:891–908.
24. Fagerland MW, Lydersen S, Laake P. Recommended tests and confidence intervals for paired binomial proportions. Stat Med. 2014;33:2850–75.
25. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Stat Sci. 2001;16:101–17.
26. Held L, Sabanés BD. Applied statistical inference, vol. 10. Berlin: Springer; 2014.
27. Alonzo TA, Pepe MS, Moskowitz CS. Sample size calculations for comparative studies of medical tests for detecting presence of disease. Stat Med. 2002;21:835–52.
28. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans Model Comput Simulation (TOMACS). 1998;8:3–30.
29. R Core Team: R. A language and environment for statistical computing. In. Vienna: R Foundation for Statistical Computing; 2021.


## Publisher's Note

### 2.3.2 Online Supplement Material

There is online supplement material available for *Thesis Article* 2 which I partially show in this section.

Online supplement material included in this section:

- Formulas for the optimal sample size calculation
- R-Code for the optimal sample size calculation testing for superiority in both endpoints in the unpaired and paired design
- Figure containing the comparison of the optimal sample size calculation with the approach of McCray et al. (2017)

Further online supplement material not included in this section, but available online:

- Simulation results in the unpaired design
- Simulation results in the paired design

## Appendix A Optimal sample size calculation

### A.I. Optimal sample size calculation to show superiority in sensitivity and non-inferiority in specificity

To investigate if the experimental test is superior to the comparator test regarding sensitivity and non-inferior regarding specificity, $\mathrm{H}_{0_{\mathrm{global}}}$ is defined as:

$$\mathrm{H}_{0_{\mathrm{global}}}: \ \mathrm{H}_{0_{\mathrm{Se}}}: \mathrm{Se_E} = \mathrm{Se_C} \ \cup \ \mathrm{H}_{0_{\mathrm{Sp}}}: \mathrm{Sp_E} \leq \mathrm{Sp_C} - \Delta$$

The positive non-inferiority margin is denoted by Δ.

#### A.I.I. Unpaired design

To calculate the corresponding sample size in the unpaired design, the formulas for superiority and non-inferiority following Zhou et al. [1] are combined:

$$\frac{\left(z_{\alpha/2}\sqrt{V_0(\mathrm{Se_C} - \mathrm{Se_E})} + z_{\beta_{\mathrm{Se}}}\sqrt{V_A(\mathrm{Se_C} - \mathrm{Se_E})}\right)^2}{(\mathrm{Se_C} - \mathrm{Se_E})^2 \cdot \pi} \stackrel{!}{=}$$

$$\frac{\left(\mathrm{z}_{\alpha/2} + z_{\beta_{\frac{1-\beta_{\mathrm{Se}}-\mathrm{Power_{overall}}}{1-\beta_{\mathrm{Se}}}}}\right)^2 \cdot V_A(\mathrm{Sp_C} - \mathrm{Sp_E})}{(\mathrm{Sp_C} - \mathrm{Sp_E} - \Delta)^2 \cdot (1-\pi)}$$

In the following formulas, α denotes the two-sided type I error rate.

#### A.I.II. Paired design

In the paired design, the optimal sample size calculation combines the formula for superiority of Miettinen et al. [2] and the formula for non-inferiority of Liu et al. [3]. Liu et

al. [3] report the sample size formula to test for equivalence. In appendix B, the sample size formula to test for non-inferiority based on the power function of Liu et al. [3] is derived.

$$\frac{\left(z_{1-\alpha/2}\cdot\psi_{\mathrm{D}}+z_{1-\beta_{\mathrm{Se}}}\sqrt{\psi_{\mathrm{D}}^2-\frac{1}{4}(\mathrm{Se_C}-\mathrm{Se_E})^2(3+\psi_{\mathrm{D}})}\right)^2}{\psi_D(\mathrm{Se_C}-\mathrm{Se_E})^2\pi}\stackrel{!}{=}$$

$$\frac{(\psi_{\mathrm{ND}}-(\mathrm{Sp_E}-\mathrm{Sp_C})^2)\left(\frac{z_{\alpha/2}}{w_{u_{\mathrm{Sp}}}}+z_{\frac{1-\beta_{\mathrm{Se}}-\mathrm{Power_{overall}}}{1-\beta_{\mathrm{Se}}}}\right)^2}{\left(-\Delta-(\mathrm{Sp_E}-\mathrm{Sp_C})\right)^2\cdot(1-\pi)}$$

With

$$w_{u_{\mathrm{Sp}}}=\frac{\sqrt{2p_{01}+(\mathrm{Sp_E}-\mathrm{Sp_C})-(\mathrm{Sp_E}-\mathrm{Sp_C})^2}}{\sqrt{2\cdot\overline{p}_{u,01}-\Delta-\Delta^2}}$$

$$\overline{p}_{u,01}=\frac{\left(-a_u+\sqrt{a_u^2-8b_u}\right)}{4}$$

$$a_u=-(\mathrm{Sp_E}-\mathrm{Sp_C})(1-\Delta)-2(p_{01}+\Delta)$$

$$b_u=\Delta(1+\Delta)p_{01}$$

$$p_{01}=\frac{\psi-\mathrm{Sp_E}+\mathrm{Sp_C}}{2}$$

## A.II. Optimal sample size calculation to show non-inferiority in sensitivity and superiority in specificity

To investigate if the experimental test is non-inferior to the comparator test regarding sensitivity and superior regarding specificity, $\mathrm{H}_{0_{\mathrm{global}}}$ is defined as:

$$\mathrm{H}_{0_{\mathrm{global}}}:\ \ \mathrm{H}_{0_{\mathrm{Se}}}:\mathrm{Se_E}\leq\mathrm{Se_C}-\Delta\cup\ \ \mathrm{H}_{0_{\mathrm{Sp}}}:\mathrm{Sp_E}=\mathrm{Sp_C}$$

The positive non-inferiority margin is denoted by Δ.

### A.II.I. Unpaired design

$$\frac{\left(\mathrm{z}_{\alpha/2} + z_{\beta_{\mathrm{Se}}}\right)^2 \cdot V_A(\mathrm{Se_C} - \mathrm{Se_E})}{(\mathrm{Se_C} - \mathrm{Se_E} - \Delta)^2 \cdot \pi} \stackrel{!}{=}$$

$$\frac{\left(z_{\alpha/2}\sqrt{V_0(\mathrm{Sp_C} - \mathrm{Sp_E})} + z_{\frac{1-\beta_{\mathrm{Se}}-\mathrm{Power_{overall}}}{1-\beta_{\mathrm{Se}}}}\sqrt{V_A(\mathrm{Sp_C} - \mathrm{Sp_E})}\right)^2}{(\mathrm{Sp_C} - \mathrm{Sp_E})^2 \cdot (1 - \pi)}$$

### A.II.II. Paired design

$$\frac{(\psi_{\mathrm{D}} - (\mathrm{Se_E} - \mathrm{Se_C})^2)\left(\frac{z_{\alpha/2}}{w_{u_{\mathrm{Se}}}} + z_{\beta_{\mathrm{Se}}}\right)^2}{\left(-\Delta - (\mathrm{Se_E} - \mathrm{Se_C})\right)^2 \cdot \pi} \stackrel{!}{=}$$

$$\frac{\left(z_{1-\alpha/2} \cdot \psi_{\mathrm{ND}} + z_{1-\frac{1-\beta_{\mathrm{Se}}-\mathrm{Power_{overall}}}{1-\beta_{\mathrm{Se}}}}\sqrt{\psi_{\mathrm{ND}}^2 - \frac{1}{4}(\mathrm{Sp_C} - \mathrm{Sp_E})^2(3 + \psi_{\mathrm{ND}})}\right)^2}{\psi_{\mathrm{ND}}(\mathrm{Sp_C} - \mathrm{Sp_E})^2(1 - \pi)}$$

The parameter $w_{u_{\mathrm{Se}}}$ is defined in analogy to $w_{u_{\mathrm{Sp}}}$ above.

### A.III. Optimal sample size calculation to show non-inferiority in both endpoints

To investigate if the experimental test is non-inferior to the comparator test regarding sensitivity and specificity, $\mathrm{H}_{0_{\mathrm{global}}}$ is defined as:

$$\mathrm{H}_{0_{\mathrm{global}}}: \ \mathrm{H}_{0_{\mathrm{Se}}}: \mathrm{Se_E} \leq \mathrm{Se_C} - \Delta_{\mathrm{Se}} \cup \ \mathrm{H}_{0_{\mathrm{Sp}}}: \mathrm{Sp_E} \leq \mathrm{Sp_C} - \Delta_{\mathrm{Sp}}$$

The positive non-inferiority margins are denoted by $\Delta_{\mathrm{Se}}$ and $\Delta_{\mathrm{Sp}}$.

### A.III.I. Unpaired design

$$\frac{\left(\mathrm{z}_{\alpha/2} + z_{\beta_{\mathrm{Se}}}\right)^2 \cdot V_A(\mathrm{Se_C} - \mathrm{Se_E})}{(\mathrm{Se_C} - \mathrm{Se_E} - \Delta)^2 \cdot \pi} \stackrel{!}{=} \frac{\left(\mathrm{z}_{\alpha/2} + z_{\beta_{\frac{1-\beta_{\mathrm{Se}}-\mathrm{Power_{overall}}}{1-\beta_{\mathrm{Se}}}}}\right)^2 \cdot V_A(\mathrm{Sp_C} - \mathrm{Sp_E})}{(\mathrm{Sp_C} - \mathrm{Sp_E} - \Delta)^2 \cdot \pi}$$

### A.III.II. Paired design

$$\frac{(\psi_{\mathrm{D}} - (\mathrm{Se_E} - \mathrm{Se_C})^2)\left(\frac{z_{\alpha/2}}{w_{u_{\mathrm{Se}}}} + z_{\beta_{\mathrm{Se}}}\right)^2}{\left(-\Delta_{\mathrm{Se}} - (\mathrm{Se_E} - \mathrm{Se_C})\right)^2 \cdot \pi} \overset{!}{=} \frac{(\psi_{\mathrm{ND}} - (\mathrm{Sp_E} - \mathrm{Sp_C})^2)\left(\frac{z_{\alpha/2}}{w_{u_{\mathrm{Sp}}}} + z_{\frac{1-\beta_{\mathrm{se}}-\mathrm{Power_{overall}}}{1-\beta_{\mathrm{se}}}}\right)^2}{\left(-\Delta_{\mathrm{Sp}} - (\mathrm{Sp_E} - \mathrm{Sp_C})\right)^2 \cdot (1-\pi)}$$

## Appendix B Derivation of the sample size for testing for non-inferiority in the paired design

Liu et al. [3] report the asymptotic power function to test for non-inferiority between the sensitivity or specificity of experimental test ($\theta_{\mathrm{E}}$) and comparator test ($\theta_{\mathrm{C}}$). The asymptotic power function is solved for the sample size.

$$1 - \phi\left(-\frac{z_{\alpha/2}}{w_u} - \frac{\Delta + (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})}{\sigma}\right) = 1 - \beta = \mathrm{Power}$$

$$\phi\left(-\frac{z_{\alpha/2}}{w_u} - \frac{\Delta + (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})}{\sqrt{\frac{\psi - (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})^2}{n}}}\right) = \beta$$

$$-\frac{z_{\alpha/2}}{w_u} - \frac{\Delta + (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})}{\sqrt{\frac{\psi - (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})^2}{n}}} = z_\beta$$

$$\frac{-\Delta - (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})}{\sqrt{\frac{\psi - (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})^2}{n}}} = z_\beta + \frac{z_{\alpha/2}}{w_u}$$

$$\sqrt{\frac{\psi - (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})^2}{n}} = \frac{-\Delta - (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})}{z_\beta + \frac{z_{\alpha/2}}{w_u}}$$

$$\sqrt{n} = \frac{\sqrt{\psi - (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})^2}\left(z_\beta + \frac{z_{\alpha/2}}{w_u}\right)}{-\Delta - (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})}$$

$$n = (\psi - (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})^2)\left(\frac{\frac{z_{\alpha/2}}{w_u} + z_\beta}{-\Delta - (\theta_{\mathrm{E}} - \theta_{\mathrm{C}})}\right)^2$$

With

$$\Delta > 0$$
$$w_u = \frac{\sqrt{2p_{01} + (\theta_E - \theta_C) - (\theta_E - \theta_C)^2}}{\sqrt{2 \cdot \overline{p}_{u,01} - \Delta - \Delta^2}}$$
$$\overline{p}_{u,01} = \frac{\left(-a_u + \sqrt{a_u^2 - 8b_u}\right)}{4}$$
$$a_u = -(\theta_E - \theta_C)(1 - \Delta) - 2(p_{01} + \Delta)$$
$$b_u = \Delta(1 + \Delta)p_{01}$$
$$p_{01} = \frac{\psi - \theta_E + \theta_C}{2}$$

**References**


1. Zhou X-H, McClish DK, Obuchowski NA. Statistical methods in diagnostic medicine. Vol. 569. 2nd ed. Hoboken, New Jersey: John Wiley & Sons; 2011.
2. Miettinen OS. The matched pairs design in the case of all-or-none responses. Biometrics. 1968;24:339-352.
3. Liu Jp, Hsueh Hm, Hsieh E, Chen JJ. Tests for equivalence or non-inferiority for paired binary data. Stat Med. 2002;21:231-245.

## R-Code

### Optimal sample size calculation testing for superiority in sensitivity and specificity

### 1. Unpaired design

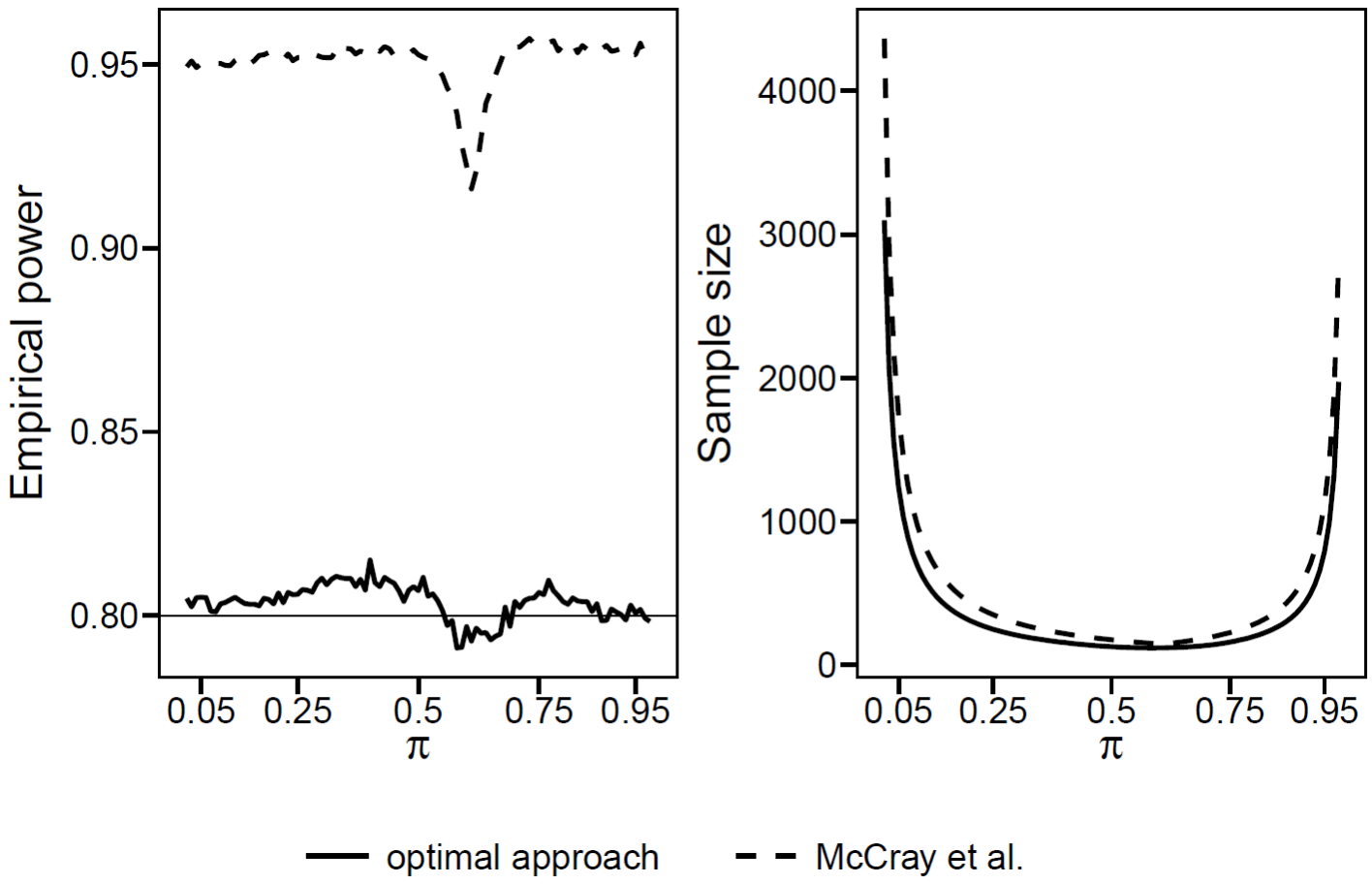
```
unpaired_superiority <- function (alpha, power, theta.se.c, theta.se.e, theta.sp.c, theta.sp.e, prev) {
 # parameter description
 # alpha: desired type I error rate per endpoint
 # power: desired overall power
 # theta.se.c: sensitivity of the comparator test
 # theta.se.e: sensitivity of the experimental test
 # theta.sp.c: specificity of the comparator test
 # theta.sp.e: specificity of the experimental test
 # prev:  prevalence

 # the variance function
 v <- function(theta.c, theta.e) {
  variance <- theta.c * (1 - theta.c) + theta.e * (1 - theta.e)
  return(variance)
 }

 # sample size calculation for one endpoint following Zhou et al. (2011)
 unpaired <- function(alpha, beta, theta.c, theta.e) {
  n <- ceiling((qnorm(alpha / 2) * sqrt(v(theta.c, theta.c)) + qnorm(beta) * sqrt(v(theta.c, theta.e))) ^
       2 / (theta.c - theta.e) ^ 2)
 }

 # calculate power for one endpoint
 calculate.power.unpaired <- function(n, alpha, theta.c, theta.e) {
  z <- (sqrt(n) * (theta.c - theta.e) - qnorm(alpha / 2) * sqrt(v(theta.c, theta.c))) / sqrt(v(theta.c,
       theta.e))
  power <- 1 - pnorm(z)
  return(power)
 }

 # function for the optimal sample size
 f <- function(alpha, power, beta.se, theta.se.c, theta.se.e, theta.sp.c, theta.sp.e, prev) {
  diff.n <- qnorm(beta.se) * sqrt(v(theta.se.c, theta.se.e)) * sqrt(1 - prev) * (theta.sp.c - theta.sp.e) -
            qnorm(1 - (power / (1 - beta.se))) * sqrt(v(theta.sp.c, theta.sp.e)) * sqrt(prev) * (theta.se.c
            - theta.se.e) - qnorm(alpha / 2) * sqrt(v(theta.sp.c, theta.sp.c)) * sqrt(prev) * (theta.se.c -
            theta.se.e) + qnorm(alpha / 2)*sqrt(v(theta.se.c, theta.se.c)) * sqrt(1 - prev) * (theta.sp.c -
            theta.sp.e)

  return(diff.n)
 }
```

```
  # approximate the sample size for beta.se and then calculate beta.sp
  beta.se <- uniroot(f, alpha = alpha, power = power, theta.se.c = theta.se.c, theta.se.e = theta.se.e,
                theta.sp.c = theta.sp.c, theta.sp.e = theta.sp.e, prev = prev,
                 lower = 0, upper = 1-power)$root
  beta.sp <- (power+beta.se-1)/(beta.se-1)
  power.total <- (1-beta.se)*(1-beta.sp)

  # calculate the sample size with beta.se and beta.sp
  n.se <- unpaired(alpha = alpha, beta = beta.se, theta.c = theta.se.c, theta.e = theta.se.e)
  n.sp <- unpaired(alpha = alpha, beta = beta.sp, theta.c = theta.sp.c, theta.e = theta.sp.e)
  N.se <- n.se / prev
  N.sp <- n.sp / (1 - prev)
  N <- ceiling(max(N.se, N.sp)) # total sample size per group

  # calculate total power
  power.se <- calculate.power.unpaired(n = N * prev, alpha = alpha, theta.c = theta.se.c, theta.e =
                                         theta.se.e)
  power.sp <- calculate.power.unpaired(n = N * (1 - prev), alpha = alpha, theta.c = theta.sp.c, theta.e
                                         = theta.sp.e)
  power.total <- power.se * power.sp
  return(list(N = N, N.se = N.se, N.sp = N.sp, power.total = power.total, power.se = power.se,
          power.sp = power.sp))
}
```

**2. Paired design**

```
paired_superiority <- function(alpha, power, theta.se.c, theta.se.e, theta.sp.c, theta.sp.e,
                               psi.d, psi.nd, prev) {

 # sample size calculation for one endpoint following Miettinen (1968)
sample.size.paired.one.endpoint <- function(alpha, beta, theta.c, theta.e, psi){
   delta <- abs(theta.c-theta.e)
   n <- (qnorm(1-alpha/2)*psi+qnorm(1-beta)*sqrt((psi^2)-0.25*(delta^2)*(3+psi)))^2 /
        (psi*(delta^2))
   return(n)
 }

 # calculate power for one endpoint
 calculate.power <- function(n, alpha, theta.c, theta.e, psi) {
   delta <- abs(theta.c-theta.e)
   z <- (sqrt(n*psi)*delta - qnorm(1-alpha/2)*psi)/sqrt((psi^2)-0.25*(delta^2)*(3+psi))
   power <- pnorm(z)
   return(power)
 }

 # function for the equal sample size for both endpoints
 f <- function(alpha, power, beta.1, theta.se.c, theta.se.e, theta.sp.c, theta.sp.e, psi.d, psi.nd, prev) {
   delta.se <- abs(theta.se.c-theta.se.e)
   delta.sp <- abs(theta.sp.c-theta.sp.e)
   diff.n <- qnorm(1-beta.1)*sqrt((psi.d^2)-0.25*(delta.se^2)*(3+psi.d))*sqrt(psi.nd*(1-
              prev))*delta.sp - qnorm(power/(1-beta.1))*sqrt((psi.nd^2)-
              0.25*(delta.sp^2)*(3+psi.nd))*sqrt(psi.d*prev)*delta.se -
              qnorm(1-alpha/2)*psi.nd*sqrt(psi.d*prev)*delta.se + qnorm(1-
              alpha/2)*psi.d*sqrt(psi.nd*(1-prev))*delta.sp
   return(diff.n)
 }

 # solve the sample size for beta.1 and then calculate beta.2
 beta.1 <- uniroot(f, alpha = alpha, power = power, theta.se.c = theta.se.c, theta.se.e = theta.se.e,
              theta.sp.c = theta.sp.c, theta.sp.e = theta.sp.e, psi.d = psi.d, psi.nd = psi.nd,
              prev = prev, lower = 0, upper = 1-power)$root
 beta.2 <- (power+beta.1-1)/(beta.1-1)
 power.total.theoretical <- (1-beta.1)*(1-beta.2)

 # calculate sample size with known beta.1 and beta.2
 n.se <- sample.size.paired.one.endpoint(alpha = alpha, beta = beta.1, theta.c = theta.se.c, theta.e =
          theta.se.e, psi = psi.d)
 n.sp <- sample.size.paired.one.endpoint(alpha = alpha, beta = beta.2, theta.c = theta.sp.c, theta.e =
          theta.sp.e, psi = psi.nd)
 N.se <- n.se/prev
 N.sp <- n.sp/(1-prev)
 N <- ceiling(max(N.se, N.sp))
```

```
# calculate power
power.se <- calculate.power(n= N*prev, alpha = alpha, theta.c = theta.se.c, theta.e = theta.se.e, psi =
 psi.d)
power.sp <- calculate.power(n= N*(1-prev), alpha = alpha, theta.c = theta.sp.c, theta.e = theta.sp.e,
psi = psi.nd)
power.total <- power.se * power.sp

return(list(N = N, N.se = N.se, N.sp = N.sp, power.total = power.total, power.se = power.se, power.sp
= power.sp, beta.1 = beta.1, beta.2 = beta.2))
}
```




Comparison of the optimal sample size calculation with the approach of McCray et al. (2017) by varying the prevalence ($\pi$) regarding the simulated empirical overall power based on 10,000 simulations runs and the calculated sample size.

## 2.4 Statement of Own Contributions to Thesis Articles

### Thesis Article 1

Zapf, A., **Stark, M.**, Gerke, O., Ehret, C., Benda, N., Bossuyt, P., Deeks, J., Reitsma, J., Alonzo, T., & Friede, T. (2020). Adaptive trial designs in diagnostic accuracy research. *Statistics in Medicine, 39*(5), 591-601. https://doi.org/10.1002/sim.8430

- Involvement in literature review and derivation of potential uses of adaptive designs in diagnostic studies
- Participation in writing the manuscript draft
- Participation in revising and finalizing the manuscript

### Thesis Article 2

**Stark, M.**, & Zapf, A. (2020). Sample size calculation and re-estimation based on the prevalence in a single-arm confirmatory diagnostic accuracy study. *Statistical Methods in Medical Research*, *29*(10), 2958-2971. https://doi.org/10.1177/0962280220913588

- Development of the optimal sample size calculation and blinded sample size re-estimation in the single-test design
- Implementation of the simulation study in R
- Writing of the manuscript draft
- Revising and finalizing the manuscript

### Thesis Article 3

**Stark, M.**, Hesse, M., Brannath, W., & Zapf, A. (2022). Blinded sample size re-estimation in a comparative diagnostic accuracy study. *BMC Medical Research Methodology*, *22*, Article 115. https://doi.org/10.1186/s12874-022-01564-2

- Development of the optimal sample size calculation and blinded sample size re-estimation in comparative design
- Implementation of the simulation study in R
- Writing of the manuscript draft
- Revising and finalizing the manuscript

# 3 Curriculum Vitae

*entfällt aus datenschutzrechtlichen Gründen*

*entfällt aus datenschutzrechtlichen Gründen*

# 4 List of Scientific Contributions

## 4.1 List of Publications

**Thesis Articles:**

**Stark, M.**, Hesse, M., Brannath, W., & Zapf, A. (2022). Blinded sample size re-estimation in a comparative diagnostic accuracy study. *BMC Medical Research Methodology*, *22*, Article 115. https://doi.org/10.1186/s12874-022-01564-2

**Stark, M.**, & Zapf, A. (2020). Sample size calculation and re-estimation based on the prevalence in a single-arm confirmatory diagnostic accuracy study. *Statistical Methods in Medical Research*, *29*(10), 2958-2971. https://doi.org/10.1177/0962280220913588

Zapf, A., **Stark, M.**, Gerke, O., Ehret, C., Benda, N., Bossuyt, P., Deeks, J., Reitsma, J., Alonzo, T., & Friede, T. (2020). Adaptive trial designs in diagnostic accuracy research. *Statistics in Medicine, 39*(5), 591-601. https://doi.org/10.1002/sim.8430

**Further Articles:**

Kohse, E. K., Siebert, H. K., Sasu, P. B., Loock, K., Dohrmann, T., Breitfeld, P., Barclay-Steuart, A., **Stark, M.**, Sehner, S., Zöllner, C., & Petzoldt, M. (2022). A modelt o predict difficult airway alerts after videolaryngoscopy in adults with anticipated difficult airways. *Anaesthesia*. https://doi:10.1111/anae.15841

Well, L., Careddu, A., **Stark, M.**, Farschtschi, S., Bannas, P., Adam, G., Mautner, V., & Salamon, J. (2021). Phenotyping spinal abnormalities in patients with Neurofibromatosis type 1 using whole-body MRI. *Scientific reports*, *11*(1), 1-13. https://doi.org/10.1038/s41598-021-96310-x

Thaler, C., Kyselyova, A. A., Faizy, T. D., Nawka, M. T., Jespersen, S., Hansen, B., Stellmann, J.-P., Heesen, C., Stürner, K. H., **Stark, M.**, Fiehler, J., Bester, M., & Gellißen, S. (2021). Heterogeneity of multiple sclerosis lesions in fast diffusional kurtosis imaging. *Plos one*, *16*(2), e0245844. https://doi.org/10.1371/journal.pone.0245844

## 4.2 List of Presentations and Posters

**Presentations related to Thesis:**

**2022** *'*Sample size re-estimation in a paired diagnostic study' at the 6th Conference of the German Consortium in Statistics (DAGStat)

**2021** 'Sample size re-estimation in a paired diagnostic study'
- at the 67th Biometric Colloquium
- at the 66th Annual Conference of the German Association for Medical Informatics, Biometry and Epidemiology (GMDS) e.V.

'Application of adaptive designs to the HEDOS study' at the 3rd workshop on the DFG project Flexible designs for diagnostic studies, joint talk together with Antonia Zapf and Amra Hot

**2019** 'Re-estimation of the prevalence in a confirmatory diagnostic accuracy study' at the
- at the 5th Conference of the German Consortium in Statistics (DAGStat)
- at the 2nd workshop on the DFG project 'Flexible designs for diagnostic studies'
- 64th Annual Conference of the German Association for Medical Informatics, Biometry and Epidemiology (GMDS) e.V.

'Neuberechnung der Fallzahl in einer zweiarmig gepaarten Diagnosestudie' at the Annual Autumn Workshop of the German Association for Medical Informatics, Biometry and Epidemiology (GMDS) e.V.

**2018** 'Prävalenzneuschätzung in konfirmatorischen Diagnosegütestudien' at the Annual Autumn Workshop of the German Association for Medical Informatics, Biometry and Epidemiology (GMDS) e.V.

**Posters related to the Thesis**

**2019** 'Fallzahlplanung und Prävalenzneuschätzung in einer konfirmatorischen Diagnosestudie' at the 3rd Symposium of the Hamburger Netzwerk für Versorgungsforschung (HAM-NET)

# Acknowledgment

*entfällt aus datenschutzrechtlichen Gründen*

# Eidesstattliche Versicherung

Ich versichere ausdrücklich, dass ich die Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die aus den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen einzeln nach Ausgabe (Auflage und Jahr des Erscheinens), Band und Seite des benutzten Werkes kenntlich gemacht habe.
Ferner versichere ich, dass ich die Dissertation bisher nicht einem Fachvertreter an einer anderen Hochschule zur Überprüfung vorgelegt oder mich anderweitig um Zulassung zur Promotion beworben habe.
Ich erkläre mich einverstanden, dass meine Dissertation vom Dekanat der Medizinischen Fakultät mit einer gängigen Software zur Erkennung von Plagiaten überprüft werden kann.

Unterschrift: .....