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

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...
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 over powering 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 over powering 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 \(H_{0\text{global}}\) is defined as the union of the null hypothesis of the sensitivity \(H_{0\text{se}}\) and the null hypothesis of the specificity \(H_{0\text{sp}}\)

\[
H_{0\text{global}} : \quad \theta_{\text{se}} = \theta_{\text{se}1} \cup H_{0\text{sp}} : \quad \theta_{\text{sp}} = \theta_{\text{sp}1}\tag{1}
\]

\(\theta_{\text{se}1}\) and \(\theta_{\text{sp}1}\) represent the sensitivity and specificity of the experimental test. \(\theta_{\text{se}0}\) and \(\theta_{\text{sp}0}\) denote the minimum sensitivity and minimum specificity to which the experimental test is compared. \(H_{0\text{global}}\) can only be rejected if \(H_{0\text{se}}\) and \(H_{0\text{sp}}\) 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. 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 \( N_{sp} \). This must be done to obtain a representative sample with the correct ratio between cases and controls. 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.

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. 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. is not considered here. The conventional sample size calculation is done so that an overall power \( \left( \frac{1}{C_0} \right) \) 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 \( \theta_{se1} = 0.81 \) and it should be shown that it is larger than \( \theta_{se0} = 0.75 \). The specificity of the CT is expected to be \( \theta_{sp1} = 0.66 \) and the study aims to show that it is larger than \( \theta_{sp0} = 0.6 \). 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 \( V(\theta) = \theta \cdot (1 - \theta) \). The upper \( z/2 \) and \( \beta \) quantile of the standard normal distribution is denoted by \( z_{\alpha/2} \) and \( z_{\beta} \):

1. Number of diseased individuals

\[
\begin{align*}
n_{se} &= \frac{\left[ z_{\alpha/2} \sqrt{V(\theta_{se0})} + z_{\beta_{se}} \sqrt{V(\theta_{se1})} \right]^2}{(\theta_{se0} - \theta_{se1})^2} = 508
\end{align*}
\]

Number of non-diseased individuals

\[
\begin{align*}
n_{sp} &= \frac{\left[ z_{\alpha/2} \sqrt{V(\theta_{sp0})} + z_{\beta_{sp}} \sqrt{V(\theta_{sp1})} \right]^2}{(\theta_{sp0} - \theta_{sp1})^2} = 683
\end{align*}
\]

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

$$N_{se} = N_{sp}$$ (2)

$$n_{se} = \frac{n_{sp}}{(1 - \pi)}$$ (3)

$$\frac{z_{1/2} \sqrt{V(\theta_{se})} + z_{\beta_{se}} \sqrt{V(\theta_{se})}}{(\theta_{se} - \theta_{se})^2 \cdot \pi} = \frac{z_{1/2} \sqrt{V(\theta_{sp})} + z_{\beta_{sp}} \sqrt{V(\theta_{sp})}}{(\theta_{sp} - \theta_{sp})^2 \cdot (1 - \pi)}$$ (4)

$$z_{\beta_{se}} \sqrt{V(\theta_{se})} (\theta_{sp} - \theta_{sp}) \sqrt{(1 - \pi)} - z_{\beta_{sp}} \sqrt{V(\theta_{sp})} (\theta_{se} - \theta_{se}) \sqrt{\pi}$$

$$= z_{1/2} \sqrt{V(\theta_{sp})} (\theta_{se} - \theta_{se}) \sqrt{\pi} - z_{1/2} \sqrt{V(\theta_{se})} (\theta_{sp} - \theta_{sp}) \sqrt{(1 - \pi)}$$ (5)

Under the condition

$$\text{Power}_{se} \cdot \text{Power}_{sp} = \text{Power}_{t}$$ (6)

$$(1 - \beta_{se}) \cdot (1 - \beta_{sp}) = \text{Power}_{t}$$ (7)

$$\beta_{sp} = \frac{1 - \beta_{se} - \text{Power}_{t}}{1 - \beta_{se}}$$ (8)

Plug the condition into the sample size calculation

$$z_{\beta_{se}} \sqrt{V(\theta_{se})} (\theta_{sp} - \theta_{sp}) \sqrt{(1 - \pi)} - z_{\beta_{sp} \cdot \text{Power}_{t}} \sqrt{V(\theta_{sp})} (\theta_{se} - \theta_{se}) \sqrt{\pi}$$

$$= z_{1/2} \sqrt{V(\theta_{sp})} (\theta_{se} - \theta_{se}) \sqrt{\pi} - z_{1/2} \sqrt{V(\theta_{se})} (\theta_{sp} - \theta_{sp}) \sqrt{(1 - \pi)}$$ (9)

Equation (9) cannot be solved analytically with respect to $\beta_{se}$ or $\beta_{sp}$ 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

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

with $\expit(x) = \frac{e^x}{1 + e^x}$, $\hat{\theta}$ as $\hat{\theta}_{se}$ or $\hat{\theta}_{sp}$ and n as n_{se} or n_{sp}, respectively. The individual null hypothesis of each endpoint will be rejected, if $\theta_{se}$ or $\theta_{sp}$ 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.\textsuperscript{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.

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

### 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](image-url)
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 consid-
ered. 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 \( p_{\text{true}} \) 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., 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 approxi-
mately 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:

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
\( p_{\text{true}} \). The parameters of the scenarios are \( \alpha = 0.05 \) (two-sided), power = 0.8, \( \theta_{\text{se}0} = 0.75, \theta_{\text{se}1} = 0.81, \theta_{\text{sp}0} = 0.6, \theta_{\text{sp}1} = 0.66 \).

| \( p_{\text{true}} \) | 0.10 | 0.20 | 0.40 | 0.47 | 0.50 | 0.60 |
|-----------------|------|------|------|------|------|------|
| true N          | 3870 | 1940 | 1185 | 1165 | 1178 | 1325 |
| Power_{se}      | 0.801| 0.802| 0.878| 0.923| 0.939| 0.984|
| Power_{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\(^{22}\)

\[
\hat{\pi} = \frac{X}{n}
\]  

\(X\) 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\(^{24}\)

\[
R = \frac{E(n_1)}{n_F(\pi_{true})}
\]  

\(E(n_1)\) represents the simulated adjusted sample size after re-estimation of the prevalence. \(n_F(\pi_{true})\) 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.

| Table 2. Simulated scenarios. |
|------------------------------|
| **One-time re-estimation**   |
| **Repeated re-estimation**   |
| True prevalence \(\pi_{true}\) | 0.2, 0.4, 0.6, 0.8 | \(\pi_{true} - 0.1, \pi_{true} + 0.1\) |
| Assumed prevalence \(\pi_{ass}\) | \(\pi_{true} - 0.1, \pi_{true} + 0.1\) | \(\pi_{true} - 0.1, \pi_{true} + 0.1\) |
| Minimum sensitivity \(\theta_{se0}\) | 0.6, 0.7, 0.8 | 0.6, 0.7, 0.8 |
| Minimum specificity \(\theta_{sp0}\) | 0.6, 0.7, 0.8 | 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) | \(\theta_{se1}\) | \(\theta_{se1} + 0.05, \theta_{se1} + 0.1, \theta_{se1} + 0.15\) |
| Sensitivity experimental test \(\theta_{se1}\) | 0.6, 0.7, 0.8 | \(\theta_{sp1}\) | \(\theta_{sp1} + 0.05, \theta_{sp1} + 0.1, \theta_{sp1} + 0.15\) |
| Specificity experimental test \(\theta_{sp1}\) | 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 - \(\beta\) | \(\theta_{se1}\) | \(\theta_{se1} + 0.05, \theta_{se1} + 0.1, \theta_{se1} + 0.15\) | \(\theta_{sp1}\) | \(\theta_{sp1} + 0.05, \theta_{sp1} + 0.1, \theta_{sp1} + 0.15\) |
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 $h_{sp0} = 0.6$, $h_{sp1} - h_{sp0} = 0.1$ 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.

![Figure 2](image.png)

**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 $h_{sp0} = 0.6$, $h_{sp1} - h_{sp0} = 0.1$ 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.

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.\textsuperscript{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
the experimental test $\theta_{se} - \theta_{se0}$ 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 $\psi$. 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 $\theta_{se0} = 0.8$, $\theta_{sp} = 0.8$ and $\theta_{sp} - \theta_{se0} = 0.15$ with either an initially over- or underestimated prevalence. If $\psi = 0.1$, $R$ will differ clearly from the optimum 1. If $\psi$ increases 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 3.** Comparison of the overall power of the fixed design and the adaptive design containing a one-time re-estimation of the prevalence with $\theta_{sp0} = 0.6$, $\theta_{sp1} - \theta_{sp0} = 0.1$ and the size of the internal pilot study $\psi = 0.5$ with either an initially over- or underestimated true prevalence.
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 \( \theta_{se0} = 0.8, \theta_{sp0} = 0.8, \theta_{sp1} = 0.95, \pi_{true} = 0.2 \). 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 estimate 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. Pieporsch 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. Plahault A, Cadilhac M and Thomas G. Sample size evaluation 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 Sabanéès 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 \( P(\text{Reject } H_0 \mid H_0 \text{ is true}) \leq \alpha \) and \( P(\text{Reject } H_0 \mid H_0 \text{ is false}) \geq 1 - \beta. \) \( \hat{\theta} \) is the maximum likelihood estimator of a true proportion \( \theta \) and is approximately normally distributed for large \( n \):

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

- Definition of the type I error rate with the critical value \( \zeta \):

\[
z_{\alpha/2} = P(\theta \geq \zeta \mid H_0) = 1 - P \left( \frac{\theta - \theta_0}{\sigma^2/\sqrt{n}} \leq \frac{\zeta - \theta_0}{\sigma^2/\sqrt{n}} \mid H_0 \right)
\]

\[
= 1 - \Phi \left( \frac{\zeta - \theta_0}{\sigma^2/\sqrt{n}} \right)
\]

\[
z_{1-\alpha/2} = -z_{\alpha/2} = \frac{\zeta - \theta_0}{\sigma^2/\sqrt{n}}
\]

\[
\Rightarrow \zeta = \theta_0 - z_{\alpha/2} \frac{\sigma^2}{\sqrt{n}}
\]

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

\[
\zeta = \frac{\exp \left( \ln \left( \frac{\theta_0}{1-\theta_0} \right) - z_{\alpha/2} \frac{1}{\sqrt{n\theta_0(1-\theta_0)}} \right)}{1 + \exp \left( \ln \left( \frac{\theta_0}{1-\theta_0} \right) - z_{\alpha/2} \frac{1}{\sqrt{n\theta_0(1-\theta_0)}} \right)}
\]

\[
\exp \left( \zeta \frac{z_{\alpha/2}}{\sqrt{n\theta_0(1-\theta_0)}} \right) = \frac{1 - \theta_0}{\theta_0} \frac{z_{\alpha/2}}{\sqrt{n\theta_0(1-\theta_0)}}
\]

\[
\Rightarrow \zeta = \theta_0 + z_{\alpha/2} \frac{\sigma^2}{\sqrt{n}}
\]

- Definition of the power in general

\[
\text{Power} = P(\theta \geq \zeta \mid H_1) = 1 - P(\theta \leq \zeta \mid H_1)
\]

\[
= 1 - P \left( \theta - \theta_1 \leq \zeta - \theta_1 \mid H_1 \right)
\]

\[
1 - \beta = 1 - \Phi \left( \frac{\zeta - \theta_1}{\sigma^2/\sqrt{n}} \right)
\]

\[
\beta = \Phi \left( \frac{\zeta - \theta_1}{\sigma^2/\sqrt{n}} \right)
\]

\[
z_\beta = \frac{\zeta - \theta_1}{\sigma^2/\sqrt{n}}
\]

\[
\Rightarrow \zeta = \theta_1 + z_\beta \frac{\sigma^2}{\sqrt{n}}
\]
Definition of the power for the logit interval

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

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

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

\[
\sqrt{n} = \frac{\sqrt{\theta_0(1 - \theta_0)}}{\sqrt{\theta_1(1 - \theta_1)}} + \frac{z_{\beta}}{\ln \left( \frac{\theta_0(1 - \theta_1)}{\theta_1(1 - \theta_0)} \right)}
\]

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