

RESEARCH ARTICLE

Adaptive trial designs in diagnostic accuracy research

Antonia Zapf¹  | Maria Stark¹ | Oke Gerke² | Christoph Ehret³ | Norbert Benda^{4,5} |
Patrick Bossuyt⁶ | Jon Deeks^{7,8} | Johannes Reitsma⁹ | Todd Alonzo¹⁰ | Tim Friede⁵ 

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

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

³Roche Diagnostics GmbH, Penzberg, Germany

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

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

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

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

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

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

¹⁰Keck 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 al³⁵ and Brannath et al³⁶).

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 al³⁷ 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 example³⁸⁻⁴⁰). 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 AUC	Paired difference ³⁶	Parametric ³⁶
		Single group ⁴¹	Parametric ^{37,38}
		Paired difference ^{37-42,44}	Semiparametric ⁴⁰
		Unpaired difference ⁴¹	Nonparametric ^{37-42,44}
	Sensitivity, specificity PPV, NPV	Single group ^{49,50}	Parametric ^{49,50} Nonparametric ⁵⁰
Adaptation of sample size	AUC	Paired difference ^{41-43,45-48}	Parametric ^{45,47} Nonparametric ^{46,48}
			Parametric ⁵¹
	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 al⁵⁴ 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 al⁵⁵ 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 al⁵⁷ 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 Feng⁵⁸ as well as Tayob et al⁵⁹ 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 al⁶⁰ 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 al⁶¹ 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 al⁶² 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 α . 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 Anderson⁶³), 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 designs⁷² 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, Høiland-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, Høiland-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, Høiland-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-PANOC: 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>