Received: 18 September 2019 | Revised: 6 July 2020 | Accepted: 17 July 2020

DOI: 10.1002/sim.8719

RESEARCH ARTICLE

Statistics in Medicine WILEY

Assessing consistency in clinical trials with two subgroups and binary endpoints: A new test within the logistic regression model

Susann Grill[1,2] | Arne Ring[3,4] | Werner Brannath[2] | Martin Scharpenberg[2]

[1] Department Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS GmbH, Bremen, Germany

[2] Competence Center for Clinical Trials Bremen, University of Bremen, Bremen, Germany

[3] medac GmbH, Wedel, Germany

[4] Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein, South Africa

Correspondence
Martin Scharpenberg, Competence Center for Clinical Trials Bremen, University of Bremen, Linzer Straße 4, 28359 Bremen, Germany.
Email: mscharpenberg@uni-bremen.de

Abstract

In late stage drug development, the experimental drug is tested in a diverse study population within the relevant indication. In order to receive marketing authorization, robust evidence for the therapeutic efficacy is crucial requiring investigation of treatment effects in well-defined subgroups. Conventionally, consistency analyses in subgroups have been performed by means of interaction tests. However, the interaction test can only reject the null hypothesis of equivalence and not confirm consistency. Simulation studies suggest that the interaction test has low power but can also be oversensitive depending on sample size—leading in combination with the actually ill-posed null hypothesis to findings regardless of clinical relevance. In order to overcome these disadvantages in the setup of binary endpoints, we propose to use a consistency test based on the interval inclusion principle, which is able to reject heterogeneity and confirm consistency of subgroup-specific treatment effects while controlling the type I error. This homogeneity test is based upon the deviation between overall treatment effect and subgroup-specific effects on the odds ratio scale and is compared with an equivalence test based on the ratio of both subgroup-specific effects. Performance of these consistency tests is assessed in a simulation study. In addition, the consistency tests are outlined for the relative risk regression. The proposed homogeneity test reaches sufficient power in realistic scenarios with small interactions. As expected, power decreases for unbalanced subgroups, lower sample sizes, and narrower margins. Severe interactions are covered by the null hypothesis and are more likely to be rejected the stronger they are.

KEYWORDS

consistency test, homogeneity test, subgroup analysis, treatment-by-subgroup interaction

1 | INTRODUCTION

1.1 | Regulatory view

Phase III clinical trials are performed to reach decisions on treatment recommendation and risk-benefit ratio in late-stage drug development. Assessment of these confirmatory trials comprises analyses of the therapeutic efficacy as well as of the safety profile of the drug in the whole study population.[1,2]

For marketing authorization of an experimental drug, "robust evidence for therapeutic efficacy" in a representative study population for the indication is inevitable according to the European Medicines Agency (EMA) *Guideline on the investigation of subgroups in confirmatory clinical trials*.[1] Typically, the treatment effect is to be proven in the whole trial population. In personalized medicine, however, tailoring clinical trials examine the overall population as well as subpopulations allowing tailored population labels for drugs that seem effective in subpopulations only and broad labels for global effects.[3]

The composition of the trial population is mostly stipulated by the target population. Furthermore, the extent of heterogeneity can be controlled by the definition of inclusion and exclusion criteria. Since it is in the interest of both the industry and public health to make the drug broadly accessible and avoid withholding effective treatment from patients often a relatively broad study population is recruited. This divers group of patients varies regarding baseline characteristics like, for example, demographic or disease parameters. Hence, treatment effects are also expected to vary in different subgroups of the study population, making investigation of possibly inconsistent treatment effects necessary.[2,4]

A differential, inconsistent, or heterogeneous treatment effect associated with a covariate is a treatment-by-covariate interaction.[1] In contrast, an absent interaction is referred to as homogeneity, equivalence or consistency. Note, that in this manuscript, the term consistency is not used in its statistical meaning describing the characteristic of an estimator. For existing differential treatment effects in the subgroups quantitative and qualitative interactions can be distinguished—subgroup-specific treatment effects differ in magnitude only (quantitative interaction) or in direction and possibly magnitude (qualitative interaction).[5]

Among other criteria, the EMA considers internal consistency robust evidence, thus similar treatment effects within relevant subgroups of the trial are assumed to confirm the treatment effect to be corroborated. However, the null hypothesis of the commonly used treatment-by-subgroup interaction test claims absence of interaction, thus treatment effect consistency can only be rejected but not confirmed. Yet, a nonsignificant interaction test cannot be considered sufficient evidence for consistency.

In a simulation study with normally distributed endpoints by Brooks et al, the conventional interaction test was found to have high power for very clear interactions, but to lose power quickly for smaller interactions, which are more likely to occur in practice.[5] Thus, the test for interaction has low power for small sample sizes and might be oversensitive for large study populations[2,6] in a way that a small irrelevant interaction is likely to be found significant. These common characteristics of a statistical test lead in combination with the ill-posed null hypothesis to decisions disregarding clinical relevance. In addition, the power of the interaction test depends on the relative size of the subgroups; the more unbalanced the subgroups the less powerful the test.[1]

Confirmatory clinical trials are typically planned and powered to prove superiority or non-inferiority of a new drug within the whole study population for a particular primary endpoint. Nevertheless, in planning as well as in analysis and inference, possible differential subgroup-specific treatment effects must be considered. If inconsistency is to be expected already during planning, further increase of sample size for a clinical trial can be legitimate in order to reach reasonable power for consistency assessment in subgroups.[1]

Conducting subgroup analyses to identify differences or confirm consistency brings along several disadvantages: low power to demonstrate treatment effects in subgroups, uncontrolled type I error rates, data-driven tests, and clinically unsound results if tests are not pre-specified.[7]

As in the EMA guideline, in this investigation, the term "subgroup" is used to refer to "a subset of the clinical trial population" including all patients showing the same level of one or more descriptive factors obtained prior to treatment assignment.[1] Millen et al use the term "subpopulation."[3]

1.2 | Demonstrating consistency

Assessment of differences in effect estimations between subgroups needs to be done with respect to the medical relevance of the difference, not only based on significance as done by the interaction test. For this reason, Ring et al have developed a consistency test for normally distributed endpoints as an alternative to the classical interaction test for trials with two treatment groups (denoted by A and B) and two subgroups.[6] This consistency or equivalence test is based on the so-called *consistency ratio*, the difference in subgroup-specific treatment effects scaled by the residual standard deviation, and enables confirmation of consistency regarding the subgroup-specific treatment effects. Formally, they regard the model

$$Y_{ijk} = \mu + \tau_i + \lambda_j + \kappa_{(i*j)} + \epsilon_{ijk},$$

where Y_{ijk} is the response variable, μ is the overall mean, τ_i the treatment effect ($i \in \{A,B\}$), λ_j the subgroup effect ($j \in \{1,2\}$), $\kappa_{(i*j)}$ the treatment-by-subgroup interaction, and ϵ_{ijk} are independent and normally distributed residuals with common standard deviation σ_r. The subgroup-specific treatment effects are then denoted by $\delta_j = (\tau_A + \kappa_{(A*j)}) - (\tau_B + \kappa_{(B*j)})$, $j \in \{1,2\}$, and the consistency ratio is defined as the ratio of the difference in subgroup-specific treatment effects and the residual standard deviation:

$$cr = \frac{\delta_1 - \delta_2}{\sigma_r}.$$

Scaling by the residual variability allows the differential treatment effect to be larger for variable data and vice versa. Ring et al derive a two-sided confidence interval for the consistency ratio, based on the non-central t-distribution. They then employ the two one-sided tests (TOST)[8] procedure, also known as interval inclusion principle.[9] For a pre-specified medically relevant margin θ_c, called *consistency margin*, which should not be exceeded by *cr* in case of equivalence of treatment effects, they simultaneously consider the two one-sided hypotheses

$$H_{0,1} : -\theta_c \geq cr \quad \text{and} \quad H_{0,2} : cr \geq \theta_c,$$

which claim that there is a relevant difference in either direction. If both hypotheses can be rejected at level α, which is equivalent to the two-sided level 2α confidence interval for *cr* being included in the interval $(-\theta_c, \theta_c)$, consistency of the subgroup-specific treatment effects is demonstrated.

Millen et al propose a decision tree for novel clinical trial designs with tailoring objectives, in which two paths lead to broad labels—the other paths result in tailored or enhanced labels.[3] Tests for overall and subpopulation treatment effects are conducted and satisfaction of two conditions, the influence and interaction condition, are examined to make a decision on the label. "The influence condition states that to enable overall population labeling, the beneficial effect of treatment must not be limited to only the predefined subpopulation." The fulfilled influence condition ensures that a harmful effect in one subgroup is not covered by a beneficial effect in the complementary subgroup, thus basically prevents qualitative interactions. However, it does not ensure similar effects in both subgroups which we aim at with our consistency tests. "The interaction condition states that to support enhanced labeling for the predefined subpopulation [...], the treatment effect in the predefined subpopulation [...] should be appreciably greater than the treatment effect in the complementary subpopulation [...]." Therefore, a significant overall treatment effect and a nonsignificant treatment effect in a predefined subpopulation lead to a broad label meaning that the overall effect applies to the whole trial population. On the other hand, a significant overall treatment effect, a significant treatment effect in a predefined subpopulation, the satisfied influence condition, and the non-satisfied interaction condition lead to a broad label.

1.3 | Objectives

Our investigations aim to expand examinations on consistency assessment for normally distributed endpoints to the framework of binary outcomes, such as, for example, response/remission rates in oncological trials. Therefore, we develop a consistency test within the logistic regression model as an alternative to the interaction test to confirm homogeneity of treatment effects in two subgroups. By demonstrating consistency we aim at ruling out qualitative interactions, but also at proving comparable effects in both predefined subgroups. The performance of the proposed test is assessed in a simulation study. In addition, we outline the consistency test for the relative risk regression.

The theoretical background is introduced in Section 2. We focus on the case of two treatments and two subgroups, which is of practical relevance, for example, for assessing gender differences. The dependency of the binary endpoint on the dichotomous covariates for treatment and subgroup as well as on the interaction of the two is modeled by logistic (and relative risk) regression. Focus is on the derivation of a consistency test, aiming to overcome disadvantages of the conventional interaction test. Since no variance term is available in the logistic (and relative risk) model, the factor between overall and subgroup-specific effect is used for homogeneity assessment. The proposed test resembles a consistency test which directly compares both subgroup-specific odds ratios published by Ring et al,[10] but can be applied in less restrictive scenarios with unbalanced subgroups.

Monte Carlo simulations are applied to examine performance of the consistency test. Therefore, a randomized controlled trial (RCT) with two parallel arms is simulated. Each patient is assigned to one of two mutually exclusive and exhaustive subgroups, for example, gender, whose characteristics are suspected to affect the treatment effect. Subgroups can be balanced or unbalanced. Data are always simulated with a true overall treatment effect but with different magnitudes of the interaction.

2 | THEORETICAL FRAMEWORK

2.1 | Statistical model and notation

We consider a clinical trial to compare two treatments T and C (eg, experimental treatment and control) with regard to their effect on a binary outcome variable Y. We focus on the case of two subgroups, denoted by S_1 and S_2. The statistical model for our analysis is:

$$\text{logit}(p_i) = \beta_0 + \beta_T x_{iT} + \beta_S x_{iS} + \beta_{TS} x_{iT} x_{iS}, \tag{1}$$

where $p_i = \mathbb{P}(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i)$ is the event probability of subject i given its covariate values $\mathbf{x}_i$. The coefficient β_T is the treatment effect, β_S is the subgroup effect and β_{TS} is the treatment-by-subgroup interaction. The following coding for the covariates is used:

$$x_{iT} = \begin{cases} 0, & \text{if subject } i \text{ is in treatment group } C \\ 1, & \text{if subject } i \text{ is in treatment group } T \end{cases}$$

$$x_{iS} = \begin{cases} -\pi_2, & \text{if subject } i \text{ is in subgroup } S_1 \\ \pi_1, & \text{if subject } i \text{ is in subgroup } S_2 \end{cases}$$

$$x_{iT}x_{iS} = \begin{cases} 0, & \text{if subject } i \text{ is in treatment group } C \\ -\pi_2, & \text{if subject } i \text{ is in treatment group } T \text{ and subgroup } S_1 \\ \pi_1, & \text{if subject } i \text{ is in treatment group } T \text{ and subgroup } S_2. \end{cases}$$

Here $\pi_k, k \in \{1,2\}$, is the proportion of subjects in subgroup S_k on a population level which can be estimated by $\hat{\pi}_k = \frac{n_k}{N}$ with n_k being the number of subjects in subgroup S_k and N the total sample size. Hence, $\pi_1 + \pi_2 = \hat{\pi}_1 + \hat{\pi}_2 = 1$ applies. This coding is chosen to assure that $E(X_{is}) = 0$. The subgroup-specific treatment effects are defined as the differences of the logits (or log-odds-ratios) of the success probabilities in both treatment groups within the respective subgroup:

$$\delta_1 = \text{logit}[\mathbb{P}(Y_i = 1|X_{iT} = 1, X_{iS} = -\pi_2)] - \text{logit}[\mathbb{P}(Y_i = 1|X_{iT} = 0, X_{iS} = -\pi_2)] = \beta_T - \pi_2\beta_{TS}, \tag{2}$$

respectively

$$\delta_2 = \text{logit}[\mathbb{P}(Y_i = 1|X_{iT} = 1, X_{iS} = \pi_1)] - \text{logit}[\mathbb{P}(Y_i = 1|X_{iT} = 0, X_{iS} = \pi_1)] = \beta_T + \pi_1\beta_{TS}. \tag{3}$$

By definition, we then also have

$$\delta_1 = \log\left(\frac{\text{odds}[\mathbb{P}(Y_i = 1|X_{iT} = 1, X_{iS} = -\pi_2)]}{\text{odds}[\mathbb{P}(Y_i = 1|X_{iT} = 0, X_{iS} = -\pi_2)]}\right) =: \log(OR_1), \tag{4}$$

and

$$\delta_2 = \log\left(\frac{\text{odds}[\mathbb{P}(Y_i = 1|X_{iT} = 1, X_{iS} = \pi_1)]}{\text{odds}[\mathbb{P}(Y_i = 1|X_{iT} = 0, X_{iS} = \pi_1)]}\right) =: \log(OR_2). \tag{5}$$

It follows that the difference in subgroup-specific treatment effects $\delta_2 - \delta_1$ equals the treatment-by-subgroup interaction parameter $\delta_2 - \delta_1 = \beta_{TS}$. The overall treatment effect Δ is defined as the expected difference of the logits of the event probabilities in both treatment groups, which equals the expected log-odds-ratio of T versus C:

$$\Delta = E(\text{logit}[\mathbb{P}(Y_i = 1|X_{iT} = 1)] - \text{logit}[\mathbb{P}(Y_i = 1|X_{iT} = 0)]) = E\left(\log\left(\frac{\text{odds}[\mathbb{P}(Y_i = 1|X_{iT} = 1)]}{\text{odds}[\mathbb{P}(Y_i = 1|X_{iT} = 0)]}\right)\right). \tag{6}$$

Because of the coding of the covariates, this amounts to

$$\Delta = \beta_T. \tag{7}$$

To characterize the magnitude of interaction, we define the following parameter, which was also considered by Ring et al:[6,10]

$$\phi = 1 - \frac{\min(\delta_1, \delta_2)}{\max(\delta_1, \delta_2)}.$$

This parameter indicates by which percentage the smaller subgroup-specific treatment effect is below the larger one. It equals zero if the subgroup-specific treatment effects are equal, indicating equivalence, and is equal to 1 if there is no treatment effect in one of the subgroups. If the treatment effect in one of the subgroups is twice as large as in the other subgroup ϕ equals 0.5. Values of $\phi > 1$ correspond to qualitative interactions. We will not consider this case further and restrict our investigations to quantitative interactions, that is, $\phi \leq 1$. An alternative parameter for the quantification of the interaction, also considered by Brookes et al[5,11] and Ring et al,[10] is defined as the ratio of the difference of subgroup-specific treatment effects and the overall treatment effect

$$\psi = \frac{\delta_2 - \delta_1}{\Delta} = \frac{\beta_{TS}}{\beta_T}.$$

The formula for ψ is simpler than that of ϕ, but since ϕ offers the more convenient interpretation, we will use ϕ in our simulations and analyses to implement and present the underlying "true" interaction. However, if $\delta_2 \geq \delta_1$, which can be achieved by reordering the subgroups, both parameters can be transformed into the other:

$$\phi = \frac{2\psi}{2 + \psi} \quad \text{respectively} \quad \psi = \frac{2\phi}{2 - \phi}.$$

Furthermore, in this case, the formula for ϕ simplifies to

$$\phi = \frac{2\beta_{TS}}{2\beta_T + \beta_{TS}}. \tag{8}$$

The subgroup-specific treatment effects can be expressed via ϕ as follows:

$$\delta_1 = \beta_T\left(1 - \frac{\phi}{2 - \phi}\right) \quad \text{and} \quad \delta_2 = \beta_T\left(1 + \frac{\phi}{2 - \phi}\right).$$

Note that the framework of this manuscript, in particular model (1), can be extended to other distributions and link functions by means of generalized linear models.[10]

2.2 | Definition of the consistency test

We observed that, because of the chosen parameterization of our model, the overall treatment effect Δ differs from the subgroup-specific effects by $-\pi_2\beta_{TS}$ and $\pi_1\beta_{TS}$:

$$\delta_1 = \Delta - \pi_2\beta_{TS} \quad \text{and} \quad \delta_2 = \Delta + \pi_1\beta_{TS}, \tag{9}$$

or expressed on the odds-ratio-scale:

$$OR_1 = exp\{\Delta\}\exp\{-\pi_2\ \beta_{TS}\} \quad \text{and} \quad OR_2 = exp\{\Delta\}\exp\{\pi_1\beta_{TS}\},$$

where $exp\{\Delta\}$ can be seen as an overall effect on the odds ratio scale. Hence, consistency of the treatment effect could be claimed, if

$$-\theta_c < \pi_2\beta_{TS} < \theta_c \quad \text{and} \quad -\theta_c < \pi_1\beta_{TS} < \theta_c, \tag{10}$$

for some consistency margin θ_c. In that case, the subgroup-specific treatment effects would not differ relevantly from the overall treatment effect. However, while statistical considerations can inform the choice of the consistency margin θ_c, the specific value has to be mainly chosen based on medical aspects. This choice should take into account therapy benefits as well as expected average efficacy in the whole population (see subsection 5.1 for more details). The ICH guideline E10[12] describes general aspects to consider for the specification of non-inferiority margins which should be adapted for evaluation of subgroup heterogeneity.

To claim the consistency stated in Equation (10), the TOST principle is employed for each subgroup, that is, the following hypotheses are tested:

$$H^1_{0,1}: \pi_2\beta_{TS} \geq \theta_c \quad H^1_{0,2}: \pi_2\beta_{TS} \leq -\theta_c \quad H^2_{0,1}: \pi_1\beta_{TS} \geq \theta_c \quad H^2_{0,2}: \pi_1\beta_{TS} \leq -\theta_c.$$

Simultaneous rejection of all four null hypotheses would allow to claim Equation (10) and thereby homogeneity of the treatment effect across subgroups. Since all hypotheses need to be rejected for a claim of homogeneity no adjustment for multiplicity is needed and they can all be tested on the one-sided significance level α. We will conduct the tests for the hypotheses by applying the interval inclusion principle. The following Theorem shows how confidence intervals for $\pi_2\beta_{TS}$ and $\pi_1\beta_{TS}$ can be derived.

Theorem 1. *Under the given assumptions, when we have an i.i.d. sample $(Y_i,\mathbf{X}_i)$, $i \in \{1, \dots, N\}$, approximate $1-2\alpha$ confidence intervals for $\pi_2\beta_{TS}$ and $\pi_1\beta_{TS}$ are given by*

$$CI^{2\alpha}_1 = \left[\hat{\pi}_2\hat{\beta}_{TS} \mp z_{1-\alpha}\frac{\hat{\sigma}_{\pi_2\beta_{TS}}}{\sqrt{N}}\right] \quad \text{and} \quad CI^{2\alpha}_2 = \left[\hat{\pi}_1\hat{\beta}_{TS} \mp z_{1-\alpha}\frac{\hat{\sigma}_{\pi_1\beta_{TS}}}{\sqrt{N}}\right].$$

Here $\hat{\pi}_k = \frac{n_k}{N}$ with n_k being the number of subjects in subgroup S_k and N the total sample size. $\hat{\beta}_{TS}$ is the maximum likelihood estimator of β_{TS} from the logistic regression model (1) and

$$\hat{\sigma}^2_{\pi_k\beta_{TS}} = \hat{\pi}^2_k\hat{\sigma}^2_{\beta_{TS}} + \hat{\beta}^2_{TS}\hat{\pi}_k(1-\hat{\pi}_k),$$

where $\hat{\sigma}^2_{\beta_{TS}}$ is the respective element from the inverse of the Fisher matrix $\mathbf{F}^{-1}(\hat{\beta})$. $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution.

Proof. A proof is given in the appendix. ■

Applying the interval inclusion principle, consistency of treatment effects across subgroups can be claimed, if $CI^{2\alpha}_1 \subseteq (-\theta_c, \theta_c)$ and $CI^{2\alpha}_2 \subseteq (-\theta_c, \theta_c)$, that is, if both confidence intervals are included by the consistency margins.

An alternative to the approach of testing for consistency in the setup of a binary endpoint outlined above is to compare both subgroup-specific treatment effects directly. This route was followed by Ring et al[10] for equally sized subgroups ($\pi_1 = \pi_2 = 0.5$). Their approach can easily be generalized to unbalanced subgroups. We can claim treatment effect consistency across subgroups if the subgroup-specific treatment effects δ_1 and δ_2 do not differ too much, that is, not by more than a pre-specified consistency margin $\tilde{\theta}_c$. Since the difference of subgroup-specific treatment effects is given by $\delta_2 - \delta_1 = \beta_{TS}$, this can be formalized by

$$-\tilde{\theta}_c < \beta_{TS} < \tilde{\theta}_c. \tag{11}$$

We can claim Equation (11) and hence treatment effect consistency if the hypotheses

$$H^3_{0,1}\ :\ \beta_{TS} \geq \tilde{\theta}_c \quad H^3_{0,2}\ :\ \beta_{TS} \leq -\tilde{\theta}_c$$

can be rejected. From Ring et al, respectively, Hosmer et al[13] follows that an $1-\alpha$ confidence interval for β_{TS} is given by

$$CI_3^{2\alpha} = \left[\hat{\beta}_{TS} \mp z_{1-\alpha}\frac{\hat{\sigma}_{\beta_{TS}}}{\sqrt{N}}\right], \tag{12}$$

where $\hat{\beta}_{TS}$ is again the maximum likelihood estimator of β_{TS} in model (1), $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution, and $\hat{\sigma}^2_{\beta_{TS}}$ is the respective element of the inverse of the Fisher matrix $\mathbf{F}^{-1}(\hat{\beta})$. Applying the interval inclusion principle we can reject the hypotheses $H^3_{0,1}$ and $H^3_{0,2}$, claiming consistency at significance level α, if $CI_3^{2\alpha} \subseteq (-\tilde{\theta}_c, \tilde{\theta}_c)$, that is, the confidence interval is included by the consistency margins.

The first approach to the testing of subgroup effect homogeneity, trying to prove (10), focuses on comparison of the subgroup-specific treatment effects to the overall treatment effect. The alternative approach (11) compares the subgroup-specific treatment effects directly, ignoring the overall treatment effect. The first kind of test is known as heterogeneity test, while the second test is known as interaction test. As clarified by Dehbi and Hackshaw[14] interaction and heterogeneity tests serve different purposes. The heterogeneity test evaluates the deviation of the subgroup-specific treatment effects to the overall treatment effect, trying to reject the hypothesis of the same treatment effects in the subgroup and the whole study population. The interaction test, on the other hand, compares treatment effects between subgroups, for example, between female and male patients, thus the null hypothesis claims that treatment effects are the same in both subgroups.

Since we aim to show consistency, respectively, the absence of relevant interaction, and thereby consider other null hypotheses than Dehbi and Hackshaw, a more appropriate naming would be homogeneity test for the first approach and equivalence test for the second approach as special cases of consistency tests. As for the heterogeneity and interaction test both tests can be seen to have different purposes. Dehbi and Hackshaw provide examples in which the heterogeneity test and the interaction test can lead to seemingly inconsistent conclusions. They recommend that it should be clarified in advance which test is examined. Otherwise, a two-stage strategy is suggested, where in a first step subgroup effects are examined in relation to the overall effect and in a second step, if evidence of heterogeneity exists, then the subgroup effects should be compared between each other with an interaction test.

In our considerations, we examine both consistency tests independently with regard to their power and type I error in various scenarios. The equivalence test was, on the one hand, simulated again to expand the scenarios examined, and on the other hand, to correct a mistake made by Ring et al.[10] While the theoretical considerations were not affected, here, the confidence interval width was overestimated due to a simulation mistake leading to decreased power of the equivalence test.

2.3 | Consistency tests in the relative risk regression

The odds ratio as an efficacy measure is often criticised for not being logic-respecting, that is, the estimated overall OR might not lie inbetween the two subgroup-specific ORs even though both subgroups amount to the whole study population.[15] However, due to the chosen parametrization our overall effect—which is not equal to the usual overall OR—is by definition logic-respecting as can be comprehended in Equation (9). Both subgroup-specific effects are defined as deviations, lower for one subgroup and higher for the other subgroup, from the overall effect.

Nonetheless, the model can be generalized and the homogeneity test can be applied in relative risk regression. The model for our analysis (1) can be generalized to:

$$g(p_i) = \beta_0 + \beta_T x_{iT} + \beta_S x_{iS} + \beta_{TS} x_{iT} x_{iS}, \tag{13}$$

where g is an invertible link function. Besides the logit-link ($g(p_i) = \text{logit}(p_i)$) from the logistic regression, the log-link ($g(p_i) = \log(p_i)$) from the relative-risk regression is a typical application. As on the odds ratio scale, the tests can be derived very similarly for the risk ratio scale. For the subgroup-specific treatment effects and the log-link, we obtain

$$\delta_1 = \log[\mathbb{P}(Y_i = 1|X_{iT} = 1, X_{iS} = -\pi_2)] - \log[\mathbb{P}(Y_i = 1|X_{iT} = 0, X_{iS} = -\pi_2)] = \beta_T - \pi_2\beta_{TS}$$

respectively

$$\delta_2 = \log[\mathbb{P}(Y_i = 1|X_{iT} = 1, X_{iS} = \pi_1)] - \log[\mathbb{P}(Y_i = 1|X_{iT} = 0, X_{iS} = \pi_1)] = \beta_T + \pi_1\beta_{TS}.$$

By definition, we then also have

$$\delta_1 = \log\left(\frac{\mathbb{P}(Y_i = 1|X_{iT} = 1, X_{iS} = -\pi_2)}{\mathbb{P}(Y_i = 1|X_{iT} = 0, X_{iS} = -\pi_2)}\right) =: \log(RR_1)$$

and

$$\delta_2 = \log\left(\frac{\mathbb{P}(Y_i = 1|X_{iT} = 1, X_{iS} = \pi_1)}{\mathbb{P}(Y_i = 1|X_{iT} = 0, X_{iS} = \pi_1)}\right) =: \log(RR_2).$$

Hence, δ_1 and δ_2 equal the subgroup-specific relative risks. Again, $\delta_2 - \delta_1 = \beta_{TS}$ holds.

The overall treatment effect Δ is generally defined as the expected difference of the event probabilities in both treatment groups on the linear predictor scale:

$$\Delta = E(\mathrm{g}[\mathbb{P}(Y_i = 1|X_{iT} = 1)] - \mathrm{g}[\mathbb{P}(Y_i = 1|X_{iT} = 0)]) = \beta_T.$$

For the relative risk regression, we obtain:

$$\Delta = E(\log[\mathbb{P}(Y_i = 1|X_{iT} = 1)] - \log[\mathbb{P}(Y_i = 1|X_{iT} = 0)]) = E\left(\log\left(\frac{\mathbb{P}(Y_i = 1|X_{iT} = 1)}{\mathbb{P}(Y_i = 1|X_{iT} = 0)}\right)\right).$$

The difference of the overall treatment effect from the subgroup-specific effects can be expressed on the relative risk scale as follows:

$$RR_1 = exp\{\Delta\}\exp\{-\pi_2\ \beta_{TS}\} \quad \text{and} \quad RR_2 = exp\{\Delta\}\exp\{\pi_1\beta_{TS}\},$$

where $exp\{\Delta\}$ can be seen as an overall effect on the relative risk scale.

Hence, again, consistency of the treatment effect could be claimed, if (10) holds for some consistency margin θ_c. The two one-sided tests, respectively the interval inclusion principle, the considerations for the choice of the margin, and the derivation of the confidence intervals can be applied to the risk ratio in the same way as for the odds ratios in section 2.2. The consistency tests in the relative risk regression have not been examined regarding their characteristics using simulations since the focus is on the logistic regression and further simulations go beyond the scope of this manuscript.

3 | SIMULATION STUDY

3.1 | Setup

We investigate the performance of the two consistency tests depending on the choice of the consistency margin θ_c, the magnitude of interaction ϕ, the event probabilities in the treatment groups, and the proportion of patients in the subgroups $\pi_1,\ \pi_2$ in a simulation study. Similar to Ring et al,[10] the underlying model for the simulation assumes the same event probability p_C in both subgroups for the control treatment. This is equivalent to $\beta_S = 0$ in the model (1). Inserting this into (1), it follows that

$$p_C = P(Y_i = 1|X_{it} = 0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

independent of the subgroup. The event probability in the control group is thus only dependent on the parameter β_0. In the simulations, β_0 is chosen such that values of 0.12, 0.27, and 0.5 are achieved for p_C. We further define the average event probability in the treatment group as

$$p_T = \text{logit}^{-1}\{E(\text{logit}[P(Y_i = 1|X_{it} = 1)])\} = \frac{e^{\beta_0+\beta_T}}{1 + e^{\beta_0+\beta_T}},$$

where the last equation follows from the fact that $E(X_{iS}) = E(X_{iT}X_{iS}) = 0$ by the choice of the coding of the covariates. In the simulations, we fix β_0 as explained above, and then choose β_T such that the χ^2 test comparing the average event probabilities of the treatment groups has a power of 80% with a sample size of $N = 100$. For those parameters β_0 and β_T, the sample size is adjusted to achieve a power of 80% to 95%. The parameter ϕ is varied over a grid from 0 to 1, and the corresponding β_{TS} is calculated according to (8). Furthermore, the proportion of subjects in subgroup S_1, π_1 (and implicitly also π_2) is varied over a grid from 0 to 1. The subgroup affiliation of the study population is determined by drawing N times from a binomial distribution with probability π_1. This sampling strategy can by chance lead to one subgroup with none or very few patients. Therefore, we restrict both subgroups to contain at least 5% and not more than 95% of the study population. If one subgroup amounted to less then 5% of the study population, thus $\hat{\pi}_k < 0.05$, sampling is repeated. Assigning half of each subgroup to each treatment complies with a stratification based on the subgroup characteristic. For each combination of parameters, 10 000 simulation runs are performed in R (version 3.6.2,[16] using packages snow and snowfall for parallel computation[17,18]) to determine the power of the homogeneity and equivalence test depending on different values of the consistency margin θ_c (simulation R code is provided as supplementary material). Ring et al argue that these parameters “reflect treatment effects and associated odds ratios that are observed in a variety of oncological indications.” The event probabilities associated with the parameters chosen for the simulations correspond to objective response rates (ORRs), that is, “the proportion of patients with tumor size reduction of a predefined amount and for a minimal time period,”[19] observed in several oncological trials.

In patients with metastatic pancreatic cancer, ORRs range from 7% to 32% for different treatments in phase II and III studies.[20,21] For different types of non-small cell lung cancer (NSCLC), response rates of 12% and 19%, 65% as well as of 31% and up to 74% were observed in phase III clinical trials.[22-24] These ranges of event probabilities are covered by our simulation study. The assessed odds ratios lie within the range of the quality-of-life response rate of 3.43 (event probabilities of 43% vs. 18%) observed for castration-resistant prostate cancer patients in a phase III trial.[25] The parameter values as well as the resulting event probabilities, overall and subgroup-specific odds ratios are listed in Table B1 in the appendix.

3.2 | Results

The performance of the consistency tests was examined dependent on the margin θ_c, the strength of interaction ϕ, and the proportion of subgroups π_1. In all simulations, the overall treatment effect is fixed based on a sample size of $N = 100$ and an overall power of 80%. Parameters examined can be found in Table B1 in the appendix. The homogeneity test is assessed in comparison to the equivalence test examined by Ring et al.[10] Since Ring et al already described the performance of the equivalence test for balanced subgroups the focus is on the homogeneity test. However, the results of both tests can be found in Figures 1 to 6. For both tests, subgroup-specific event probabilities are calculated to correspond to pre-specified values of subgroup size and ϕ. In all examinations, the consistency test is less powerful for the same margins (see Figures 1 to 6) since confidence intervals around the effect estimate β_{TS} are always wider and further off 0 than those around $\pi_1\beta_{TS}$ and $\pi_2\beta_{TS}$, $\pi_1, \pi_2 < 1$ (see Theorem 1 and Equation (12)). However, one needs to keep in mind that the interpretation of the margins is different for both tests. While the margin for the equivalence test restricts the discrepancy between both subgroup-specific effects in the homogeneity test, the margin constrains the deviation of the subgroup effects to the treatment effect in the whole study population. Thus, both tests are indeed applied to the same population but are not directly comparable.

All parameters investigated, namely, the overall power, the subgroup proportion π_1, consistency margin θ_c, the treatment affect in the control group expressed by β_0 as well as the sample size N, affect the power of the consistency tests at least to some extent. For the sake of clarity, only parameters representing clear differences between parameter values and trends are presented. For example, results for overall power of 90% and 95% are omitted since both values did not change the performance of the consistency test much. Additional results for $\beta_0 = 0$ can be found in the online supplement, results

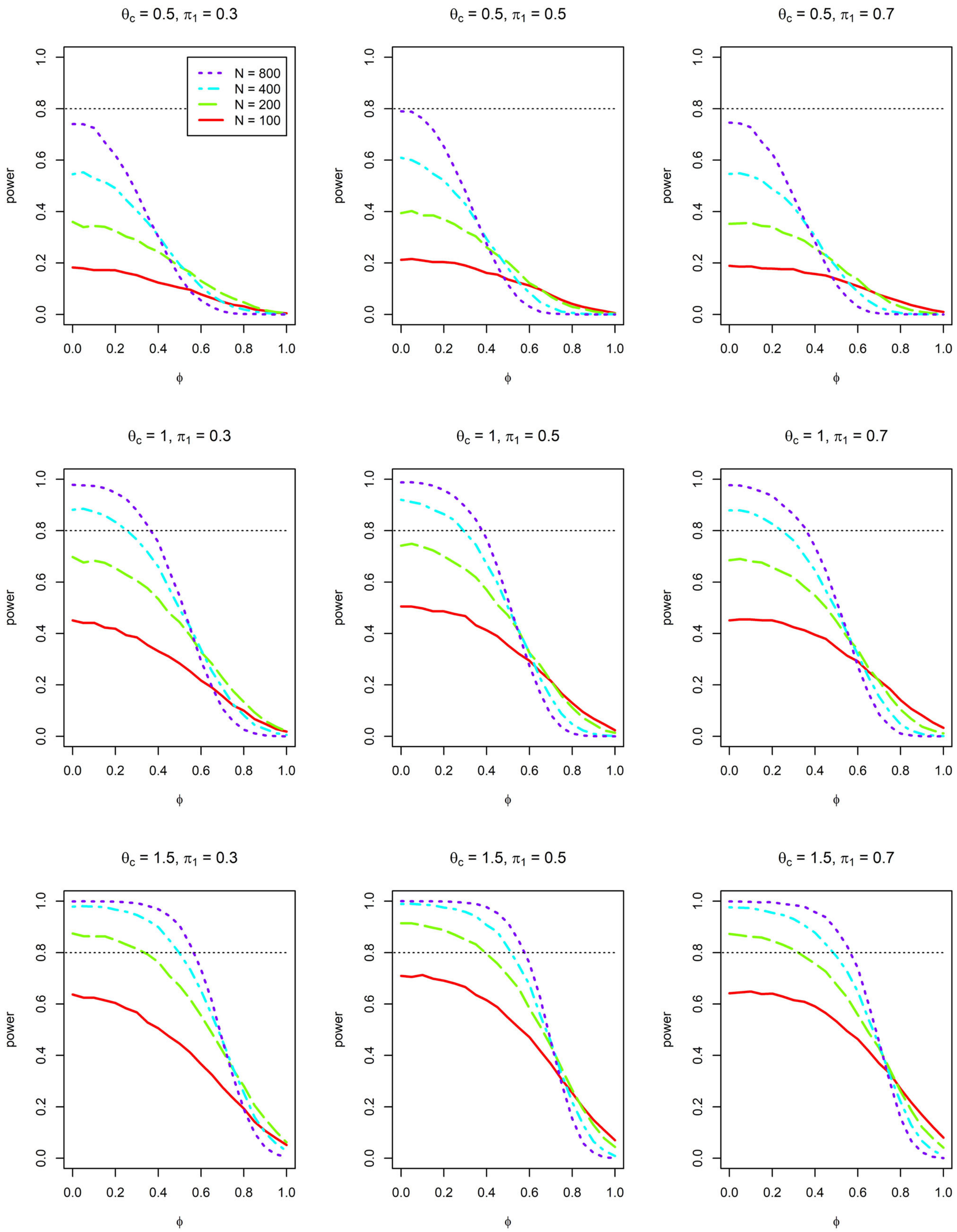
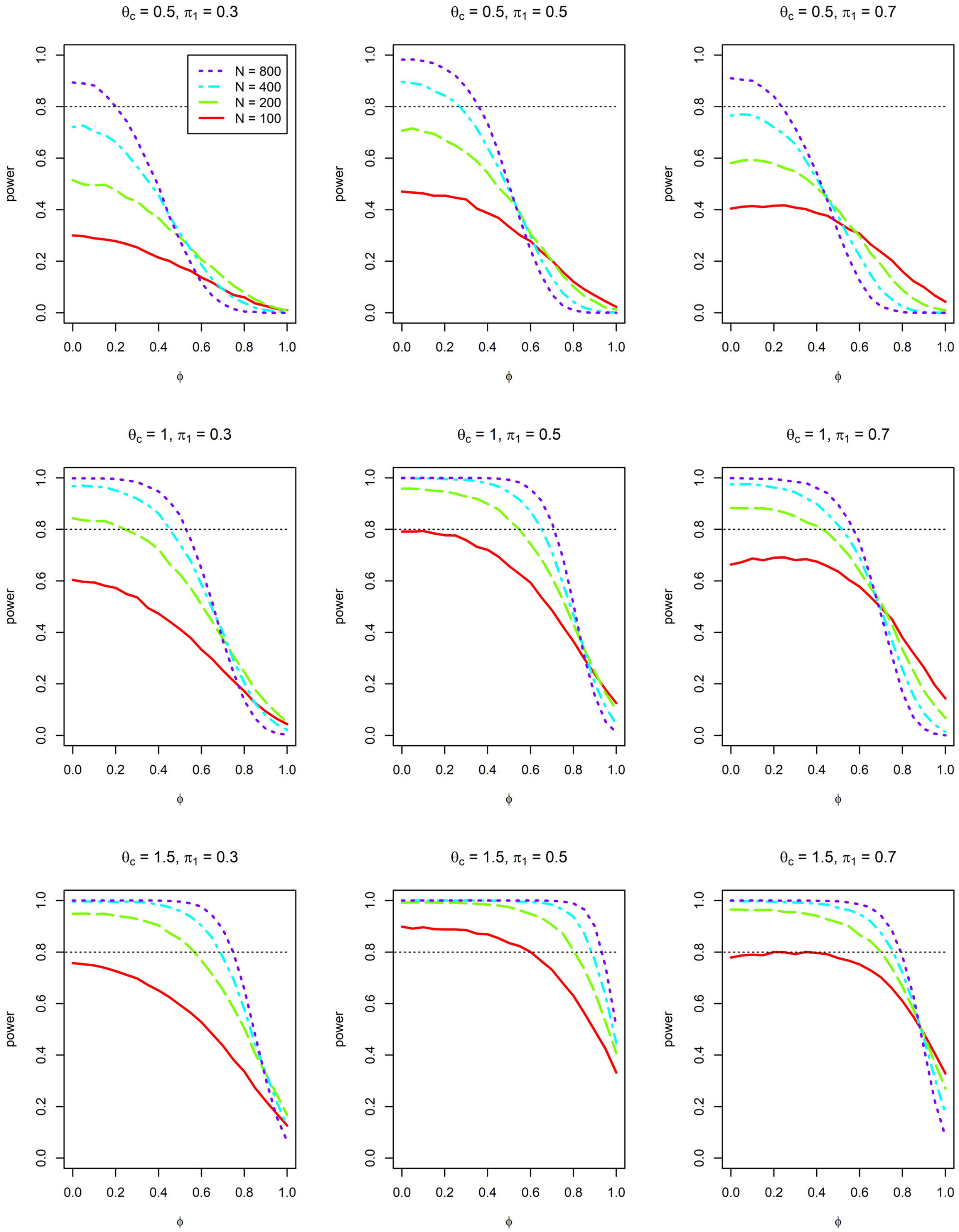


FIGURE 1 Power of the **homogeneity test** as a function of **subgroup differences** ϕ for different sample sizes N, consistency margins θ_c, subgroup proportions π_1 and $\beta_0 = -2$. The event probability in the treatment group is determined such that with a sample size of $N = 100$, an overall power of 80% is reached [Colour figure can be viewed at wileyonlinelibrary.com]



F I G U R E 2 Power of the **equivalence test** as a function of **subgroup differences** ϕ for different sample sizes N, consistency margins θ_c, subgroup proportions π_1 and $\beta_0 = -2$. The event probability in the treatment group is determined such that with a sample size of $N = 100$, an overall power of 80% is reached [Colour figure can be viewed at wileyonlinelibrary.com]

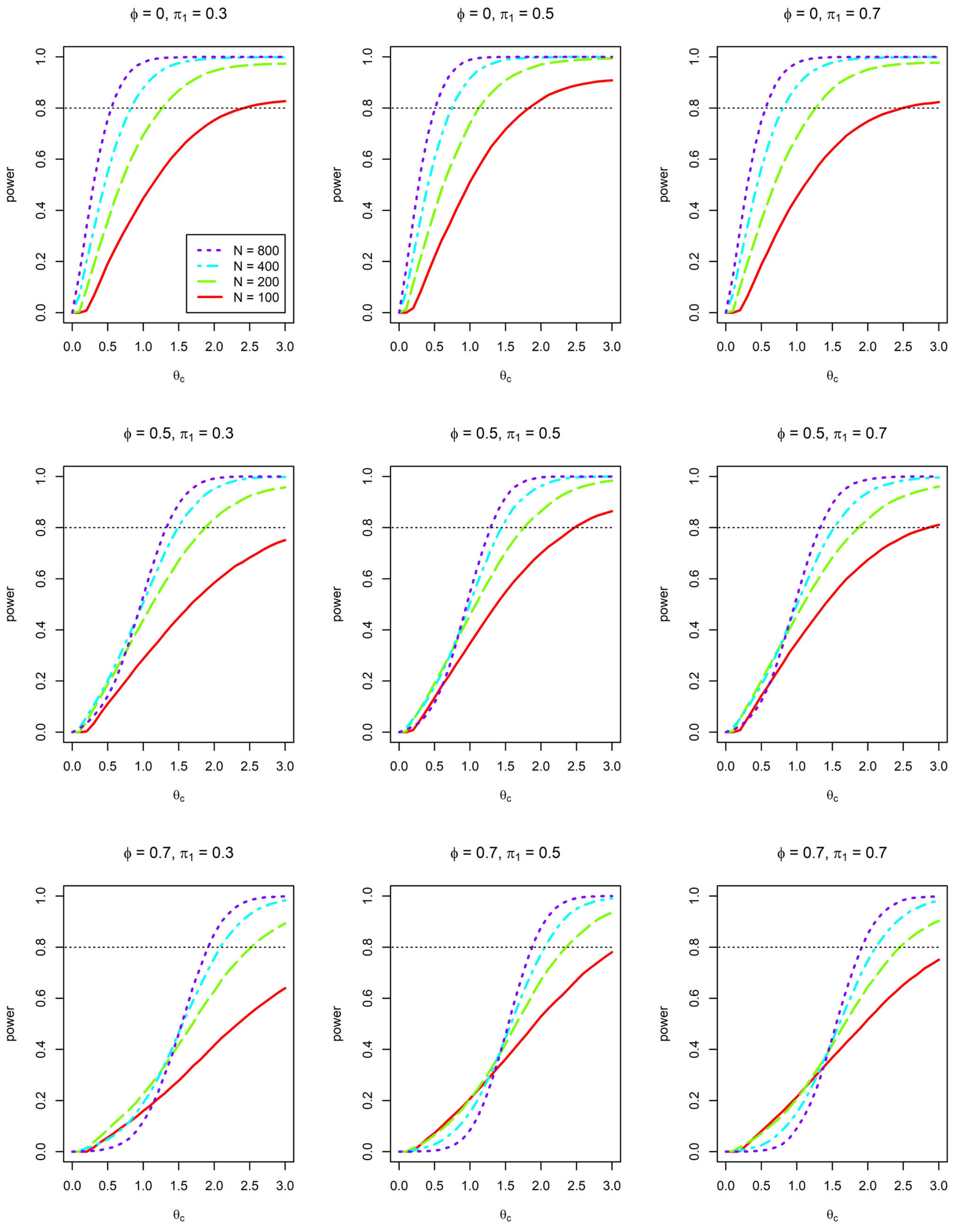
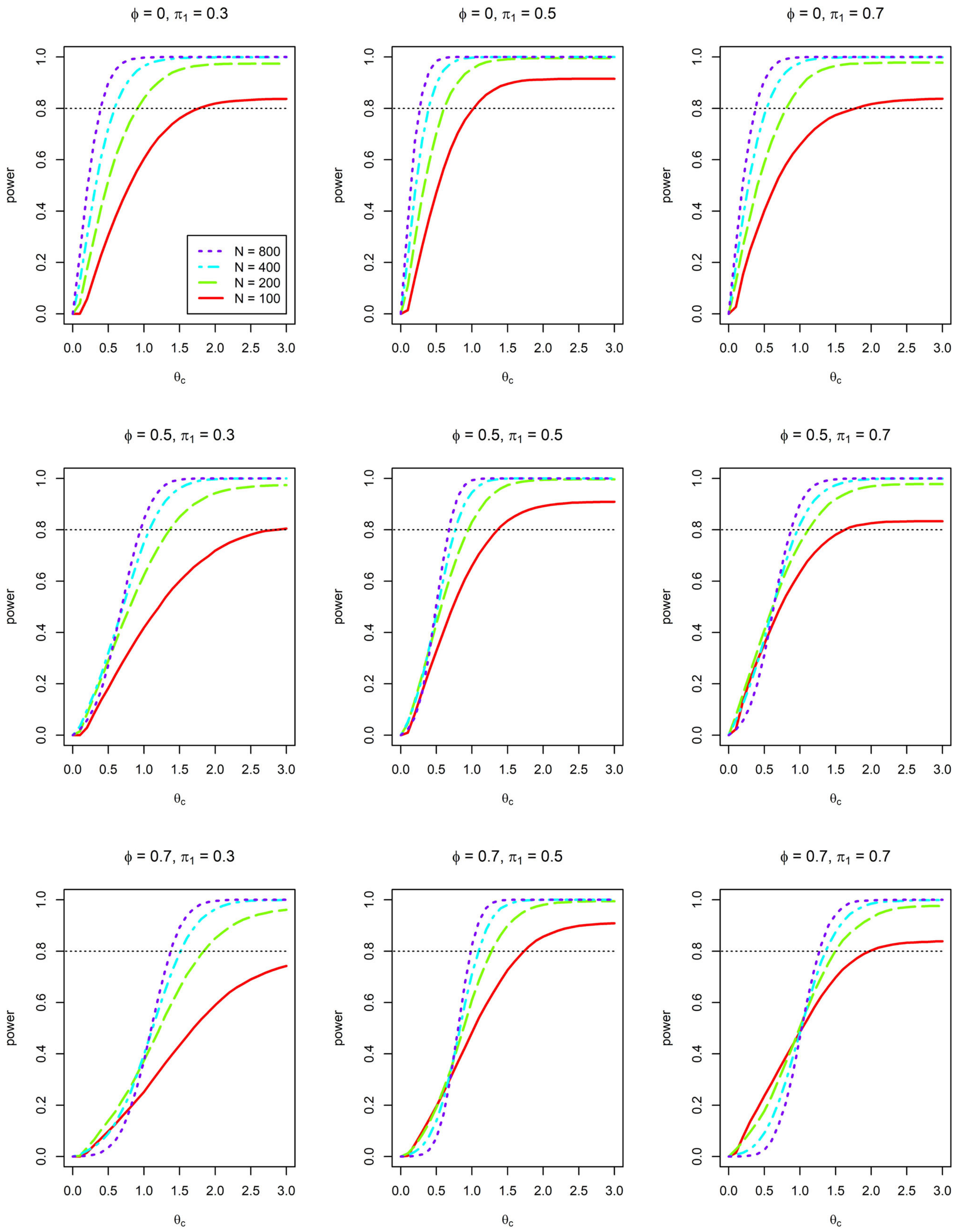


FIGURE 3 Power of the **homogeneity test** as a function of **consistency margin** θ_c for different sample sizes N, subgroup differences ϕ, subgroup proportions π_1 and $\beta_0 = -2$. The event probability in the treatment group is determined such that with a sample size of $N = 100$, an overall power of 80% is reached [Colour figure can be viewed at wileyonlinelibrary.com]



F I G U R E 4 Power of the **equivalence test** as a function of **consistency margin** θ_c for different sample sizes N, subgroup differences ϕ, subgroup proportions π_1 and $\beta_0 = -2$. The event probability in the treatment group is determined such that with a sample size of $N = 100$, an overall power of 80% is reached [Colour figure can be viewed at wileyonlinelibrary.com]

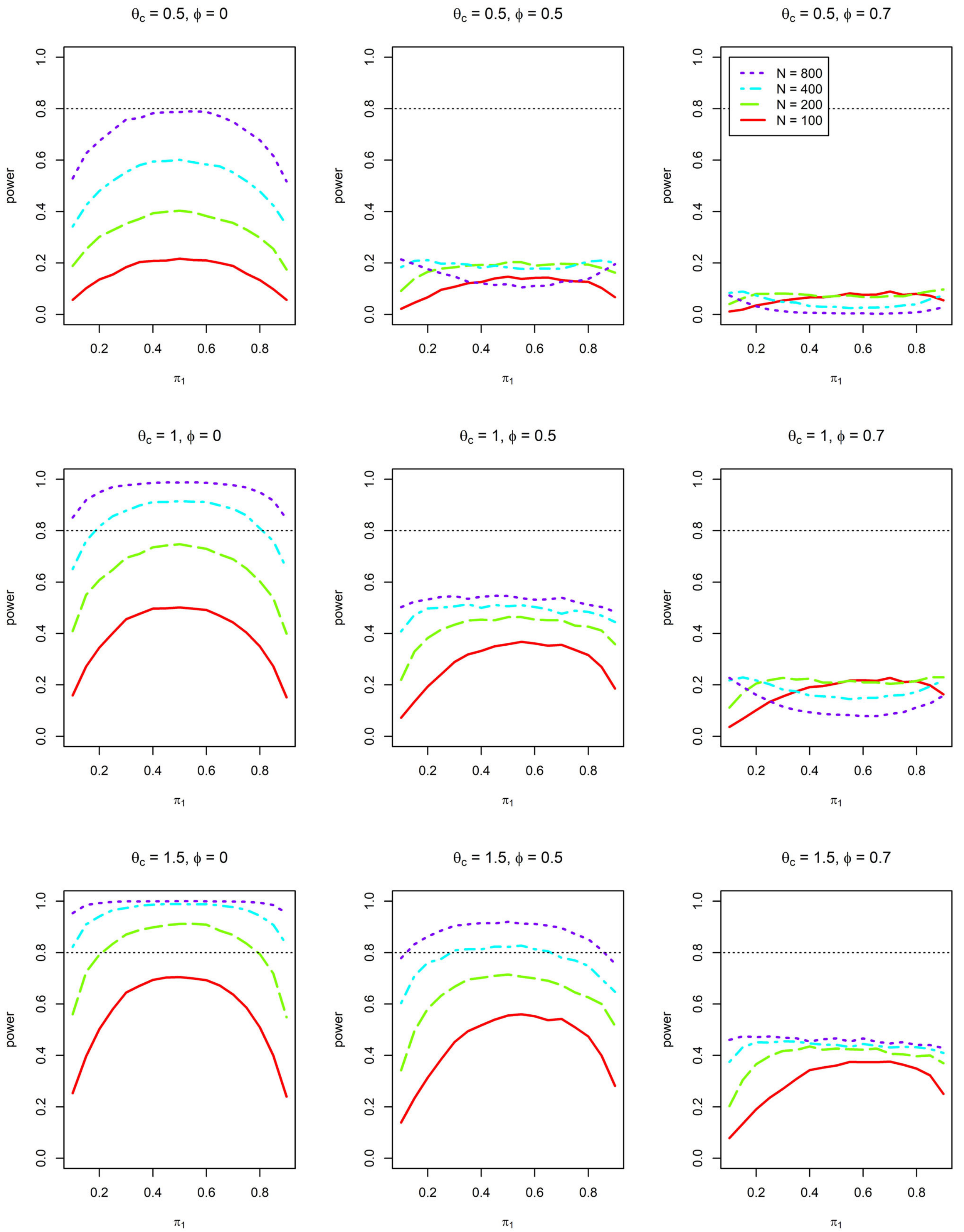


FIGURE 5 Power of the **homogeneity test** as a function of **subgroup proportions** π_1 for different sample sizes N, subgroup differences ϕ, consistency margins θ_c and $\beta_0 = -2$. The event probability in the treatment group is determined such that with a sample size of $N = 100$, an overall power of 80% is reached [Colour figure can be viewed at wileyonlinelibrary.com]

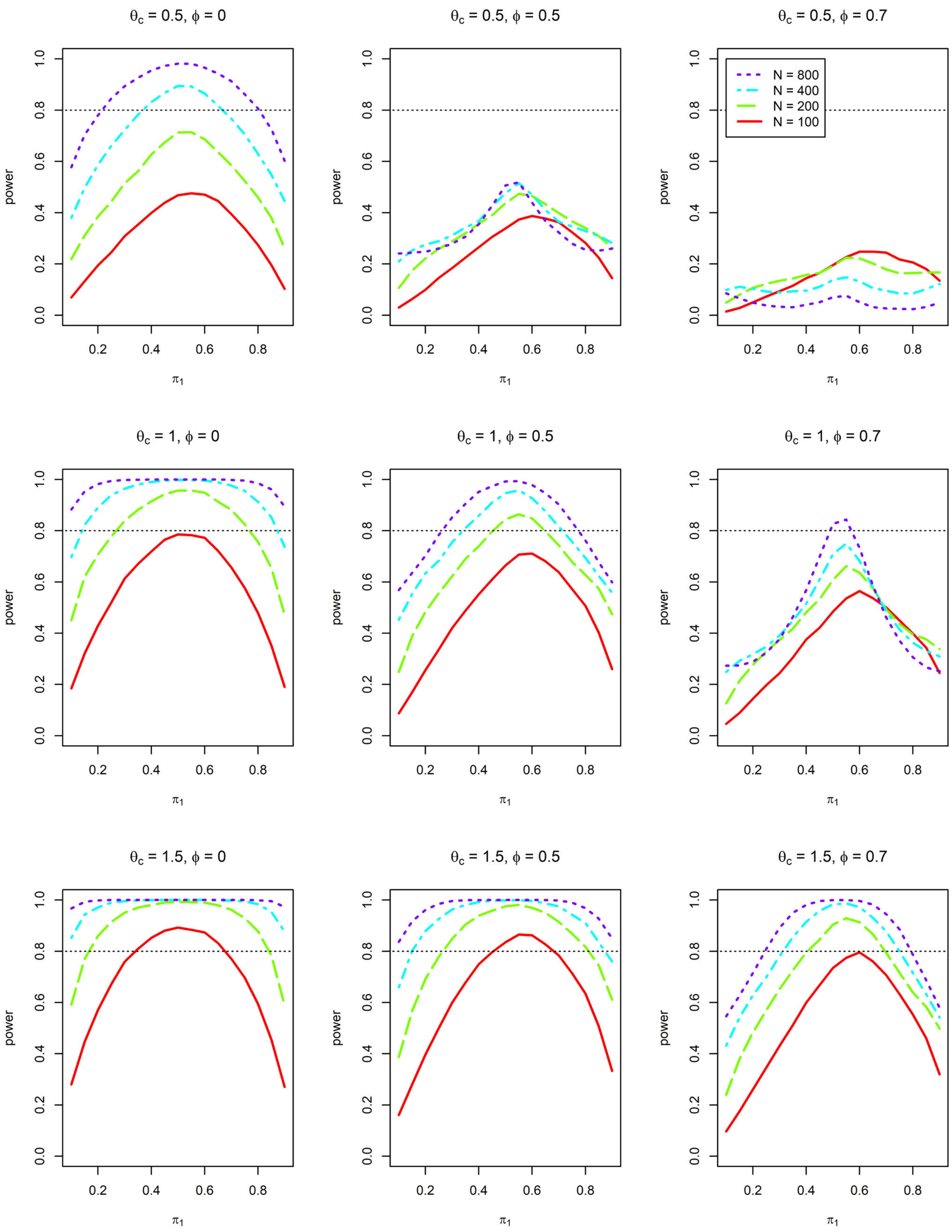


FIGURE 6 Power of the **equivalence test** as a function of **subgroup proportions** π_1 for different sample sizes N, subgroup differences ϕ, consistency margins θ_c and $\beta_0 = -2$. The event probability in the treatment group is determined such that with a sample size of $N = 100$, an overall power of 80% is reached [Colour figure can be viewed at wileyonlinelibrary.com]

for $\beta_0 = -1$ are omitted here as well. Power of both consistency tests is increased slightly for $\beta_0 = 0$ compared to $\beta_0 = -2$, in rare parameter combinations up to 35%. The impact of the event probability in the control group was described further by Ring et al[10] and Grill.[26]

As expected and within the spirit of the test principle, the power of the consistency tests decreases for stronger interactions, thus for higher ϕ and lower margin θ_c (see Figures 1 and 2). Note that power of the consistency tests for high ϕ rather corresponds to the type I error than to the actual power of the test. Depending on the margin, the power of the homogeneity test keeps almost constantly high for small values of ϕ (eg, up to $\phi = 0.6$ for $\theta_c = 1$, $N = 800$ and balanced subgroups and up to $\phi = 0.1$ for $\theta_c = 0.5$, $N = 800$ and unbalanced subgroups) and then drops rapidly forming S-curves. The higher the sample size N the steeper the curve drops for increasing ϕ. These different slopes for different sample sizes lead to intersections of curves at high values of ϕ, thus to higher power of the homogeneity test for lower sample sizes. For a large margin $\theta_c = 1$ and $\theta_c = 1.5$, a power of up to 50% ($N = 800$ in balanced subgroups) remains for declaring homogeneity although one subgroup does not have a treatment effect at all ($\phi = 1$). For unbalanced subgroups and a lower margin, this power—corresponding to the type I error—ranges from 0 for the highest sample size up to 20% for the lowest sample size. The narrower the margins the more the power varies over sample sizes for $\phi = 0$, and vice versa for $\phi = 1$. The power of the overall test (80%) for a sample size of $N = 100$ and balanced subgroups is almost reached for small interactions up to $\phi = 0.3$ for a margin of $\theta_c = 1$ and exceeded for $\theta_c = 1.5$. For unbalanced subgroups, the power is lower.

As expected, the higher the consistency margin θ_c the more likely consistency (with respect to the margin) can be found statistically significant (Figures 3 and 4). For the highest margin in all examined scenarios—including a strong interaction of $\phi = 0.7$—both tests reach power close to 100% for sample sizes $N \geq 200$. This margin of $\theta_c = 3$ allows the subgroup-specific odds ratios to deviate from the overall effect by a factor between 0.05 and 20.1. Even for $N = 100$, a power of nearly or over 80% is reached for the homogeneity test in all scenarios examined. The S-shaped power curves again exhibit lower slopes for smaller sample sizes. Unbalanced subgroups also decrease power: while a sample size of $N = 100$ suffices for a power of ca. 70% for $\phi = 0.5$ and $\theta_c = 1$ in balanced subgroups, only 40% ($\pi_1 = 0.3$) and 65% ($\pi_1 = 0.7$) power are reached for unbalanced subgroups. The strength of the interaction, expressed by larger values of ϕ, shifts the onset of the increase of power towards a higher margin. The power of the overall test for treatment effect (80% for $N = 100$) is reached for margins $\theta_c \geq 1$ depending on subgroup proportions and magnitude of interaction.

The expected impact of the subgroup proportion, π_1, is distorted by the increasing interaction ϕ and narrow margins θ_c (Figures 5 and 6). For an absent interaction, the power reaches its maximum for balanced subgroups at $\pi_1 = 0.5$ over all examined margins. The wider the margin, the higher the sample size and the smaller the interaction the less the power is affected by π_1, for example, the power starts decreasing closer to $\pi_1 = 0$ and 1. For different treatment effects in the subgroups, $\phi = 0.5$ and $\phi = 0.7$, the maximum is shifted towards $\pi_1 = 0.6$—the smaller the sample sizes the more noticeable the shift. In some cases, the equivalence test drops to its minimum power for balanced subgroups.

4 | CLINICAL TRIAL APPLICATION

We applied both consistency tests to the data of a randomized, controlled phase III clinical trial which demonstrated superiority of Metformin over placebo regarding the recurrence rate of polyps one year after polypectomy in patients without diabetes.[27] In post-hoc analyses, the recurrence of polyps was examined in different subgroups identified at baseline, for example, by sex, cholesterol, and blood glucose levels or smoking status. In both sexes, more polyps recurred in the placebo group (see Table 1). Higurashi et al drew no conclusions from the subgroup examinations.

Of 133 patients, 102 were male (subgroup 1, $\hat{\pi}_1 = 0.767$) and 31 were female (subgroup 2, $\hat{\pi}_2 = 0.233$). The effects of the treatment, subgroup and the interaction of both were estimated using a logistic regression model: the treatment was coded 0 for the placebo group and 1 for the Metformin group, the subgroup was coded $\hat{\pi}_1$ for females and $-\hat{\pi}_2$ for males. An interaction test would not have been significant (two-sided P-value of 0.163). The magnitude of interaction can be estimated by the parameter ϕ from the regression coefficients of the overall treatment effect and the interaction (see Equation (8)) and results in $\hat{\phi} = 0.9$ indicating a strong interaction with the smaller subgroup-specific treatment effect 90% below the larger one.

The equivalence test led to an estimate of the ratio of both subgroup-specific treatment effects of $\hat{\beta}_{TS} = -1.25$ (90% CI: -1.38, -1.12, see Equation (12) for calculation of confidence interval). For a predefined margin of $\tilde{\theta}_c = 1.0$, equivalence cannot be shown (see Figure 7 for a graphic illustration). To demonstrate equivalence, the confidence interval must lie between the margins. The chosen margin would allow the ratio to lie between 0.37 and 3.72 on the odds ratio scale which can be calculated as $exp\{\pm\tilde{\theta}_c\}$.

TABLE 1 Recurrence rates of polyps one year after polypectomy in males and females treated with Metformin or placebo[27]

		Treatment			
Subgroup	**Outcome**	**Metformin ($n_T = 62$)**	**Placebo ($n_C = 71$)**	**Subgroup-specific treatment effect[a]**	**Overall treatment effect[b]**
females	non-recurrent polyps	14 (77.8%)	5 (38.5%)	0.18	0.47
	recurrent polyps	4 (22.2%)	8 (61.5%)		
males	non-recurrent polyps	30 (56.6%)	22 (44.9%)	0.62	
	recurrent polyps	23 (43.4%)	27 (51.1%)		

[a] calculated as $exp\{\hat{\beta}_T + \hat{\pi}_1\hat{\beta}_{TS}\}$ and $exp\{\hat{\beta}_T - \hat{\pi}_2\hat{\beta}_{TS}\}$, respectively (see Equations (2) to (5)).
[b] calculated as $exp\{\hat{\beta}_T\}$ (see Equations (6) and (7)).

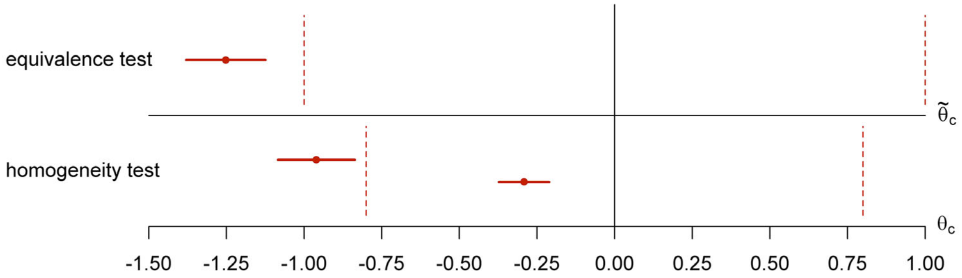


FIGURE 7 Graphic illustration of the results of both consistency tests applied to phase III clinical trial data.[27] The equivalence test is based on the ratio of both subgroup-specific treatment effects and cannot reject the null hypothesis since the confidence interval of the estimated ratio does not fall in between the pre-specified margins (red dashed lines). The homogeneity test is based on each subgroup-specific effect relative to the overall treatment effect and cannot demonstrate homogeneity since only one ratio falls in between the margins [Colour figure can be viewed at wileyonlinelibrary.com]

For the homogeneity test, we calculated $\hat{\pi}_2\hat{\beta}_{TS} = -0.29$ (90% CI: −0.33, −0.21) for subgroup 1 and $\hat{\pi}_1\hat{\beta}_{TS} = -0.96$ (90% CI: −1.08, −0.84) for subgroup 2. Calculation of the respective confidence intervals is described in Theorem 1. For a pre-specified margin of $\theta_c = 0.8$, which would allow a deviation of the subgroup effects from the overall effect on the odds ratio scale by a factor between $exp\{-0.8\} = 0.45$ and $exp\{0.8\} = 2.23$, the null hypotheses cannot be rejected and homogeneity cannot be shown (see Figure 7 for a graphic illustration).

5 | DISCUSSION

The present study extends investigations by Ring et al[10] on the performance of an equivalence test applied for consistency assessment in trials with heterogeneous study populations, binary endpoints, and balanced subgroups by examinations in unbalanced subgroups. In addition, we propose an homogeneity test for the same purpose of consistency assessment but with a different interpretation of consistency. Instead of both subgroup effects the new homogeneity test compares each subgroup effect to the overall treatment effect in the study population, an approach that has not been applied to binary endpoints before. We also outline the derivation of both tests for the relative risk regression. The power of both consistency tests increases for wider margins, higher sample sizes and most important for decreasing differences between subgroup-specific treatment effects. The decrease of power for unbalanced subgroups as observed for the interaction test remains in the consistency tests.[1]

In contrast to the commonly used interaction test[1] applied after a test for overall treatment effect, both consistency tests facilitate test decisions based on the relevance of observed differential treatment effects in subgroups. Nonetheless, an additional consistency test is only practical with sufficient power for a tolerable deviation of subgroup effects to each other and to overall treatment effects. Thus, selection of an appropriate consistency margin remains the critical point during study design.

5.1 | Consistency margins

The consistency margin allows the specification of a maximum acceptable deviation of the subgroup-specific treatment effects from each other or from the overall treatment effect, respectively. In the case of a significant consistency test, the overall interpretation of the study results can be retained for the subgroups. Therefore, besides statistical approaches medical considerations like relevant treatment effects, alternative treatment options, safety aspects, etc. must be taken into account. Since several factors need to be taken into account, no universally acceptable margin can be determined. Ring et al give a numerical example for determination of the consistency margin in a hypothetical clinical trial in which the consistency test is applied to test the secondary hypothesis of non-consistency in the subgroups of both genders.[10] Based on the anticipated overall treatment effect and the sample size calculated for the primary test to reach sufficient power, a clinically acceptable deviation in the subgroups leads to the margin's value. Monte Carlo simulations (or for the values examined here, the diagrams in the results section 3.2) can be utilized to estimate the power of the consistency test. Sample size and/or margin can be tweaked if power is not sufficient.

In our simulations, if the overall test for treatment effect, given a sample size of $N = 100$, reaches sufficient power (80%) the homogeneity test only reaches sufficient power ≥80% for wide enough margins ≥1. However, a margin of $\theta_c = 1$ allows the subgroup-specific treatment effects already to deviate from the overall treatment effect by a factor between 0.37 to 2.7. A margin of, for example, $\theta_c = 2$ allows a deviation factor between 0.13 and 7.4, $\theta_c = 3$ corresponds to a 0.05 to 20.1-fold deviation. In bioequivalence studies, treatment effects are generally only allowed to deviate from the standard drug by a factor between 0.80 and 1.25[28] which corresponds to a margin of only $\theta_c = 0.22$, thus basically obtaining no power. In the equivalence test, these factors restrict the ratio between both subgroup-specific treatment effects. The reasonableness of such a deviation must be verified based on medical considerations. Although a margin $1 \leq \theta_c \leq 2$ appears to be a very clear and relevant deviation it might be reasonable to accept this difference if no or only very few other treatment options are available. If the margin appears to be too liberal, an increase of sample size must be considered.

For example, for marketing authorization of drugs with many alternative therapy options or serious adverse events, the acceptable extent of deviation from the overall treatment effect indeed must be considerably lower. If only small interactions are reasonable, narrow margins must be chosen and sample size likely needs to be increased greatly to reach sufficient power. However, this would in turn lead to an increased power of the overall test possibly enabling smaller, more irrelevant overall treatment effects to be found significant if non-consistency represents the secondary hypothesis. During study planning, this connection between power of the consistency and overall test must be considered and balanced carefully. Since no aspects of the test procedure can be data-driven, the margin needs to be pre-specified.

Independent of the margin's value the question remains what can be concluded from a non-significant consistency test. Just like a non-significant interaction test cannot conclude equivalence, a non-significant consistency test cannot serve as prove of heterogeneity. If the alternative hypothesis stating consistency was true although the null hypothesis could not be rejected a type II error is made. Without further investigation, no conclusion can be drawn from a non-significant consistency test. Therefore, we recommend further examination of the treatment effect in each subgroup in comparison to the overall effect. In doing so not only the effect estimates should be taken into account but also the confidence intervals. Location and width of the confidence intervals of each subgroup must be evaluated in relation to the overall effect.

5.2 | Demonstrating consistency

Our consistency tests are suitable in a phase III clinical trial aiming at marketing approval of the experimental drug for a broad patient population. In contrast, tailoring clinical trials allow a broad label, but also a tailored and enhanced label if treatment effect is limited to a subpopulation or treatment is efficacious in the whole study population, but especially beneficial in a subpopulation.[3] Millen et al propose to approve a broad label if the global treatment effect is significant and a pre-planned test for treatment effect in a subpopulation is not. Instead of a subgroup-specific test which might not reach enough power, a consistency test could answer the actual question of interest: are the subgroups regarding the treatment effect similar enough for a broad label? Due to the predefined margin, the consistency tests allow the consideration of clinically relevant differences in the treatment effect of different subgroups. A test for treatment effect in the subpopulation, on the contrary, can only find a significant treatment effect, independent of clinical relevance.

Millen et al suggest another pathway to come to a broad label in tailored trials: a satisfied influence condition and a non-satisfied interaction condition after finding significant effects in the whole trial population and the subgroup of interest. Examination of the influence condition is only of interest if the overall treatment effect could be shown to be significant. Satisfaction of the influence condition ensures absence of a qualitative interaction, in which one subgroup is even harmed by the treatment. By choice of clinically reasonable margins this can be proven by our homogeneity or equivalence tests. The suggested Gail-Simon test for a qualitative interaction could be applied if consistency is to be rejected in favor of an interaction.[3,29]

The interaction condition is satisfied if the treatment would be considerably more efficacious in one subgroup compared to the complementary group, which could be tested by an interaction test. But again, a non-significant test is not sufficient to claim consistency.[1] If consistency is the aim, one of our consistency tests might be the better choice. Millen et al suggest to use the ratio of the two subgroup-specific effects to test the interaction condition, which we also use in our equivalence test. If the ratio exceeds a pre-specified constant satisfaction of the condition can be claimed. As in our consistency tests, the margin must be chosen based on clinical relevance. Instead of defining one lower bound and testing one-sided once, the pre-specified margin in our equivalence test results in a lower and upper bound, that are tested one-sided twice.

Both proposed consistency tests can be used to examine the influence and interaction condition proposed by Millen et al if consistency is of interest. In a tailored trial planned to approve either label, a consistency test could support decision making.

Alosh and Huque suggest a flexible approach to test for treatment effects in subgroups and whole trial populations in a predefined test sequence. Subgroup-specific tests are intended “once a pre-specified degree of consistency in the efficacy findings between the subgroup and the overall study population is met.”[7] They suggest to evaluate consistency either between subgroups or between subgroup and overall population which our equivalence and homogeneity test, respectively, do as well. The requirement of consistency can be met by, for example, ensuring the absence of a qualitative interaction or the existence of a minimum level of efficacy in the whole population or the complementary subgroup in case of a significant treatment effect in one subgroup. For clinical trials aiming for marketing authorization, these conditions seem rather low.

5.3 | Limitations and assets

In this investigation, we propose and apply a new method for consistency assessment of subgroup effects for binary outcomes—as a potential alternative to the use of interaction tests. Both proposed tests are based on the interval inclusion principle, that is, the estimate and the respective confidence interval are calculated and compared to two margins. The width of the confidence interval depends on the variance and partly determines the power of the test.

We assume that a slight underestimation of the variance of the estimator for the factor between subgroup-specific and overall treatment effect may have occurred. This is due to the fact that both subgroups in each simulated trial were not allowed to contain less than 5% or more than 95% of the study population. This underestimated variance probably led to a small, artificial increase in power of our homogeneity test.

A serious limitation of our and other consistency tests is the difficulty of determining consistency margins which is discussed in detail in Section 5.1.

In the logistic regression model, treatment effects are not easily comparable over different studies using our homogeneity test since different values of ϕ lead to different values of the effect coefficient for the interaction dependent on the event probability in the control group (see Table B1). In turn, this leads to the necessity of choosing margins in dependence of the event probability in the control group and prevents comparability of study results.

Our simulations covered a wide range of parameters—the magnitude of interaction ϕ covered the whole range of quantitative interactions, both subgroups accounted for 10% to 90% of the study population, margins expanded widely, and event probabilities varied over values found in several clinical trials. Investigation of unbalanced subgroups added more realistic scenarios to examinations conducted earlier.[10] For example, in a phase III clinical trial in lung cancer patients by Borghaei et al,[22] the study population is split into different proportions by baseline characteristic which all might be of interest in subgroup analyses—7% were older than 74 years, 55% were male, 92% had a certain disease stage, and 40 % had a certain previous therapy.

High power for low interactions, low power for strong interactions, and S-shaped power curves over the magnitude of interaction with decreasing power for increasing interactions are deemed desirable criteria for a good consistency test.[10] We also observed high power for small interactions and vice versa as well as S-shaped power curves whereas the slope is steeper the higher the sample size. For sample sizes four- to eightfold as high as in the overall test, our homogeneity test is highly selective and would be very suitable to distinguish small from strong differential subgroup effects. Cut-off points can be determined by choosing appropriate margins. However, as discussed earlier, a multiplication of sample size to reach sufficient power in a secondary test is unlikely ethically nor financially justifiable.

5.4 | Outlook

The proposed homogeneity test can be applied to trials with binary endpoints and two subgroups. Similar consistency tests have been developed for normally distributed endpoints and two subgroups.[6,10] Future developments could focus on survival data. Covariables like age, comorbidities, and comedications are likely to split the population into more than two subgroups for which suitable consistency tests are to be designed. If several factors are supposed to be investigated for their impact on the treatment effect adjustment for multiplicity must be considered. As in the numerical example outlined above (see Section 5.1), the consistency test will likely mostly test secondary hypothesis, thus needs to be embedded in appropriate multiple-testing strategies to control the type I error. In order to get a better sense of appropriate margins both tests need to be applied to real phase III clinical trial data, preferably in comparison to the interaction test.

Characterization of the magnitude of interaction can also be defined by parameters ϕ or ψ which describe by which percentage the smaller subgroup-specific treatment effect is below the larger one and the ratio of the difference of subgroup-specific treatment effects and the overall effect respectively (see Section 2.1 for details). Both parameters can be used for normally distributed and binary endpoints and could also be developed for survival data facilitating comparison of subgroup effects even between different types of endpoints. In the logistic regression model, differential effects would be comparable independent of event probabilities in the treatment groups.

6 | CONCLUSIONS

For marketing authorization, a new drug must convincingly show efficacy in the whole study population but also in relevant subgroups. Our proposed homogeneity test can facilitate confirmation of absence of clinically relevant differential treatment effects in subgroups by replacing the conventional interaction test. Choice of consistency margins must be carefully balanced prior to start of the trial bearing statistical and medical aspects in mind.

ACKNOWLEDGEMENTS

The internship of Susann Grill, from which the master's thesis emerged, was financially supported by medac GmbH. Arne Ring is employed by medac GmbH.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study were generated in a simulation study. The program code is part of the supplementary material to this manuscript.

ORCID

Werner Brannath https://orcid.org/0000-0002-8622-3904

Martin Scharpenberg https://orcid.org/0000-0003-1584-0056

REFERENCES

1. European Medicines Agency: EMA/CHMP/539146/2013 - Guideline on the investigation of subgroups in confirmatory clinical trials; 2019. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf. Assessed January 28, 2020.
2. Alosh M, Huque MF, Bretz F, D'Agostino RB. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Stat Med*. 2017;36(8):1334-1360.

3. Millen BA, Dmitrienko A, Ruberg S, Shen L. A statistical framework for decision making in confirmatory multipopulation tailoring clinical trials. *Ther Innov Regul Sci*. 2012;46(6):647-656. https://doi.org/10.1177/0092861512454116.
4. Dmitrienko A, Muysers C, Fritsch A, Lipkovich I. General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *J Biopharm Stat*. 2016;26(1):71-98.
5. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey SG. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess*. 2001;5(33):1-56.
6. Ring A, Day S, Schall R. Consistency of subgroup effects in clinical trials. Paper presented at: Annual Meeting Pharmaceutical Staistics Industry (PSI); May 2016; Berlin/Germany. https://www.psiweb.org/docs/default-source/default-document-library/16psiber-programme-of-abstracts-v2.pdf?sfvrsn=f816d3db_0. Accessed March 27, 2020.
7. Alosh M, Huque MF. A flexible strategy for testing subgroups and overall population. *Stat Med*. 2009;28(1):3-23. https://doi.org/10.1002/sim.3461.
8. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokinet Biopharm*. 1987;15(6):657-680.
9. Ocaña J, Sanchez O, Maria P, Carrasco JL. Carryover negligibility and relevance in bioequivalence studies. *Pharm Stat*. 2015;14(5):400-408.
10. Ring A, Scharpenberg M, Grill S, Schall R, Brannath W. Equivalence tests in subgroup analyses. In: Zhao Y, Chen DG, eds. *New Frontiers of Biostatistics and Bioinformatics*. ICSA Book Series in Statistics. Cham: Springer; 2018:201-238.
11. Brookes ST, Whitely E, Egger M, Davey Smith G, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol*. 2004;57(3):229-236.
12. European Medicines Agency: CPMP/ICH/364/96 - Choice of control group in clincial trials 2001. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-10-choice-control-group-clinical-trials-step-5_en.pdf. Assessed August 13, 2020.
13. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc; 2013.
14. Dehbi HM, Hackshaw A. Investigating subgroup effects in randomized clinical trials. *J Clin Oncol*. 2017;35(2):253-254.
15. Lin HM, Xu H, Ding Y, Hsu JC. Correct and logical inference on efficacy in subgroups and their mixture for binary outcomes. *Biom J*. 2019;61(1):8-26. https://doi.org/10.1002/bimj.201800002.
16. R Core Team *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna: 2019. https://www.R-project.org/.
17. Tierney L, Rossini AJ, Li N, Sevcikova H. *Snow: Simple Network of Workstations. R Package Version 0.4-3*. 2018
18. Knaus J. *Snowfall: Easier Cluster Computing (Based on Snow). R Package Version 1.84-6.1*. 2015.
19. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER).Guidance for industry: clinical trial endpoints for the approval of cancer drugs and biologics; 2007. https://www.fda.gov/. Accessed January 24, 2018.
20. Von Hoff DD, Ervin T, Arena FP, et al. Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. *N Engl J Med*. 2013;369(18):1691-1703. https://doi.org/10.1056/NEJMoa1304369.
21. Conroy T, Desseigne F, Ychou M, et al. Folfirinox versus gemcitabine for metastatic pancreatic cancer. *N Engl J Med*. 2011;364(19):1817-1825. https://doi.org/10.1056/NEJMoa1011923.
22. Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus Docetaxel in advanced nonsquamous non-small-cell lung cancer. *N Engl J Med*. 2015;373(17):1627-1639. https://doi.org/10.1056/NEJMoa1507643.
23. Maemondo M, Inoue A, Kobayashi K, et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med*. 2010;362(25):2380-2388. https://doi.org/10.1056/NEJMoa0909530.
24. Khozin S, Blumenthal GM, Jiang X, et al. US Food and Drug Administration approval summary: Erlotinib for the first-line treatment of metastatic non-small cell lung cancer with epidermal growth factor receptor exon 19 deletions or exon 21 (L858R) substitution mutations. *Oncologist*. 2014;19(7):774-779.
25. Scher HI, Fizazi K, Saad F, et al. Increased survival with Enzalutamide in prostate cancer after chemotherapy. *N Engl J Med*. 2012;367(13):1187-1197. https://doi.org/10.1056/NEJMoa1207506.
26. Grill S. Assessing Consistency of Subgroup Specific Treatment Effects in Clinical Trials with Binary Endpoints [Unpublished Master thesis]. University of Bremen, Bremen; 2017.
27. Higurashi T, Hosono K, Takahashi H, et al. Metformin for chemoprevention of metachronous colorectal adenoma or polyps in post-polypectomy patients without diabetes: a multicentre double-blind, placebo-controlled, randomised phase 3 trial. *Lancet Oncol*. 2016;17(4):475-483. https://doi.org/10.1016/S1470-2045(15)00565-3.
28. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research: Guidance for Industry - Statistical approaches to establishing bioequivalence. 2001. https://www.fda.gov/media/70958/download. Assessed February 1, 2017.
29. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*. 1985;41(2):361-372.

How to cite this article: Grill S, Ring A, Brannath W, Scharpenberg M. Assessing consistency in clinical trials with two subgroups and binary endpoints: A new test within the logistic regression model. *Statistics in Medicine*. 2020;39:4551–4573. https://doi.org/10.1002/sim.8719

APPENDIX A. MATHEMATICAL PROOFS

Proof of Theorem 1. By the central limit theorem, we have

$$\sqrt{N}(\hat{\pi}_k - \pi_k) \overset{\mathcal{L}}{\to} N(0, \sigma^2_{\pi_k}), \tag{A1}$$

with $\sigma^2_k = \pi_k(1 - \pi_k)$ for $k = 1,2$. Furthermore, it is well known that the maximum likelihood estimate $\hat{\beta}_{TS}$ is asymptotically normal:

$$\sqrt{N}(\hat{\beta}_{TS} - \beta_{TS}) \overset{\mathcal{L}}{\to} N(0, \sigma^2_{\beta_{TS}}), \tag{A2}$$

for some variance $\sigma^2_{\beta_{TS}}$. Additionally, we can show that $\hat{\beta}_{TS}$ and $\hat{\pi}_k$ are asymptotically independent:

$$\begin{aligned} Cov(\hat{\beta}_{TS}, \hat{\pi}_k) &= E((\hat{\beta}_{TS} - \beta_{TS})(\hat{\pi}_k - \pi_k)) \\ &= E[E((\hat{\beta}_{TS} - \beta_{TS})(\hat{\pi}_k - \pi_k)|\hat{\pi}_k)] \\ &= E[E(\hat{\beta}_{TS} - \beta_{TS}|\hat{\pi}_k)(\hat{\pi}_k - \pi_k)] \to^{N\to\infty} 0, \end{aligned}$$

where the last convergence follows with $E(\hat{\beta}_{TS} - \beta_{TS}|\hat{\pi}_k) \to^{N\to\infty} 0$. From this, together with (A1) and (A2) we can conclude

$$\sqrt{N}\left(\begin{pmatrix} \hat{\beta}_{TS} \\ \hat{\pi}_k \end{pmatrix} - \begin{pmatrix} \beta_{TS} \\ \pi_k \end{pmatrix}\right) \overset{\mathcal{L}}{\to} N_2(0, \Sigma),$$

where N_2 denotes the bivariate normal distribution and

$$\Sigma = \begin{pmatrix} \sigma^2_{\beta_{TS}} & 0 \\ 0 & \sigma^2_{\pi_k} \end{pmatrix}.$$

Application of the delta-method yields

$$\sqrt{N}(\hat{\pi}_k\hat{\beta}_{TS} - \pi_k\beta_{TS}) \overset{\mathcal{L}}{\to} N(0, \pi_k^2\sigma^2_{\beta_{TS}} + \beta^2_{TS}\sigma^2_{\pi_k}).$$

The asymptotical variance can be consistently estimated by

$$\hat{\sigma}^2_{\pi_k\beta_{TS}} = \hat{\pi}_k^2\hat{\sigma}^2_{\beta_{TS}} + \hat{\beta}^2_{TS}\hat{\pi}_k(1 - \hat{\pi}_k),$$

where $\hat{\sigma}^2_{\beta_{TS}}$ is the respective element from the inverse of the Fisher matrix $\mathbf{F}^{-1}(\hat{\beta})$. Hence, a two-sided $1 - \alpha$ confidence interval for $\pi_k\beta_{TS}$ is given by

$$CI^\alpha = \left[\hat{\pi}_k\hat{\beta}_{TS} \mp z_{1-\alpha}\frac{\hat{\sigma}_{\pi_k\beta_{TS}}}{\sqrt{N}}\right].$$

■

APPENDIX B. SIMULATION SETUP

See Table B1.

TABLE B1 Parameter used for simulation study and resulting event probabilities as well as odds ratios

β_0	p_c	β_T	p_T	$exp\{\Delta\}$	ϕ	β_{TS}	π_1	p_{T1}	p_{T2}	OR_1	OR_2
−2	0.12	1.40	0.35	4.06	0	0	0.30	0.35	0.35	4.06	4.06
							0.50	0.35	0.35	4.06	4.06
							0.70	0.35	0.35	4.06	4.06
					0.50	0.93	0.30	0.22	0.42	2.11	5.37
							0.50	0.26	0.47	2.55	6.48
							0.70	0.29	0.51	3.07	7.81
					0.70	1.51	0.30	0.16	0.46	1.41	6.39
							0.50	0.21	0.54	1.91	8.64
							0.70	0.26	0.61	2.58	11.68
−1	0.27	1.16	0.54	3.20	0	0	0.30	0.54	0.54	3.20	3.20
							0.50	0.54	0.54	3.20	3.20
							0.70	0.54	0.54	3.20	3.20
					0.50	0.78	0.30	0.41	0.60	1.86	4.04
							0.50	0.44	0.63	2.17	4.72
							0.70	0.48	0.67	2.54	5.51
					0.70	1.25	0.30	0.33	0.63	1.33	4.66
							0.50	0.39	0.69	1.71	5.99
							0.70	0.45	0.74	2.20	7.69
0	0.50	1.19	0.77	3.29	0	0	0.30	0.77	0.77	3.29	3.29
							0.50	0.77	0.77	3.29	3.29
							0.70	0.77	0.77	3.29	3.29
					0.50	0.79	0.30	0.65	0.81	1.89	4.17
							0.50	0.69	0.83	2.21	4.89
							0.70	0.72	0.85	2.59	5.73
					0.70	1.28	0.30	0.57	0.83	1.34	4.83
							0.50	0.63	0.86	1.73	6.24
							0.70	0.69	0.89	2.24	8.07