Brinton et al. Trials (2015) 16:458
DOI 10.1186/s13063-015-0951-3

TRIALS

METHODOLOGY Open Access

An internal pilot design for prospective cancer screening trials with unknown disease prevalence

John T. Brinton[1*], Brandy M. Ringham[2] and Deborah H. Glueck[2]

Abstract

Background: For studies that compare the diagnostic accuracy of two screening tests, the sample size depends on the prevalence of disease in the study population, and on the variance of the outcome. Both parameters may be unknown during the design stage, which makes finding an accurate sample size difficult.

Methods: To solve this problem, we propose adapting an internal pilot design. In this adapted design, researchers will accrue some percentage of the planned sample size, then estimate both the disease prevalence and the variances of the screening tests. The updated estimates of the disease prevalence and variance are used to conduct a more accurate power and sample size calculation.

Results: We demonstrate that in large samples, the adapted internal pilot design produces no Type I inflation. For small samples (*N* less than 50), we introduce a novel adjustment of the critical value to control the Type I error rate. We apply the method to two proposed prospective cancer screening studies: 1) a small oral cancer screening study in individuals with Fanconi anemia and 2) a large oral cancer screening trial.

Conclusion: Conducting an internal pilot study without adjusting the critical value can cause Type I error rate inflation in small samples, but not in large samples. An internal pilot approach usually achieves goal power and, for most studies with sample size greater than 50, requires no Type I error correction. Further, we have provided a flexible and accurate approach to bound Type I error below a goal level for studies with small sample size.

Keywords: Cancer screening, Internal pilot area under the curve, Type I error, Power, Receiver operating characteristic analysis

Background

Lingen et al. [1] proposed a study to compare the diagnostic accuracy of two screening modalities for the detection of oral pre-malignant and malignant lesions. During the planning phase of the trial, Lingen et al. considered a paired design with the full area under the receiver operating characteristic curve (AUC) as the outcome.

In a paired cancer screening trial, each participant is given two screening tests [1–4]. The participants are typically volunteers drawn from a standard screening population. Thus, the trial includes both participants with disease and participants without disease. At entry, the disease status of the participants is unknown. Presumably, the disease status of the participants in the trial mirrors the prevalence in the population.

The sample size for the trial proposed by Lingen et al. depended on the prevalence of disease in the population. The reported prevalence of oral malignant and pre-malignant lesions varied by as much as 16.5 % [5], even in published reports, depending on the population studied. If the prevalence of lesions was 12.1 %, as observed by [5], 2,450 participants would have been required to achieve 95 % power for the trial. However, if the prevalence of lesions was 0.2 % [6], Lingen and his colleagues would have needed to recruit 116,100 participants, a 47-fold increase.

All researchers have an ethical responsibility to choose an accurate sample size. Participants in cancer screening trials may face emotional and physical harm from needless biopsy, false positive diagnoses, and

* Correspondence: john.brinton@dhha.org
[1]Denver Health Medical Center, 777 Bannock St., MC 6551, Denver, Colorado 80204, USA
Full list of author information is available at the end of the article

over-diagnosis of non-fatal disease. A study that overestimates the sample size required for a cancer screening trial exposes study participants to needless harm. A study that underestimates the sample size lacks the power to answer the research question, while still exposing study participants to potential harm.

One possible solution to the ethical dilemma is an internal pilot study. In an internal pilot design, investigators use information from the first fraction of study participants accrued to estimate unknown parameters [7–10]. The estimates can then be used to calculate an updated sample size.

Previous work on internal pilot designs for screening studies has assumed that the ratio of cases is known prior to the start of the study and that the ratio is fixed throughout the course of the study. Wu et al. [11] proposed an internal pilot approach for the comparison of the diagnostic accuracy of screening tests, but, like Coffey and Muller [12], assumed that the ratio of cases to non-cases was known before the study, and fixed by design during the study. In addition, the method of Wu et al. [11] does not control for possible Type I error inflation. While Gurka et al. [13] considered the use of internal pilot designs for observational studies, they did not suggest any Type I error correction techniques. In general, in small samples, internal pilot designs can inflate Type I error [14]. There are multiple approaches for controlling Type I error inflation in internal pilots, when the inflation occurs due to variance re-estimation [12, 15–18].

We broaden the definition of the internal pilot design to match the sampling scheme in cancer screening trials. We adapt internal pilot methodology to the cancer screening setting by: 1) allowing the ratio of cases to non-cases to vary randomly throughout the study, 2) re-estimating the sample size with internal pilot sample estimates of both the disease prevalence and the variance of the outcome, and 3) adjusting the critical value to control for possible Type I error rate inflation caused by sample size re-estimation. The critical value correction depends on the unconditional distribution of the test statistic. We show that the approach allows investigators to attain a targeted power level, and control Type I error rate inflation in small samples. We demonstrate, via simulation, that no correction is needed for large samples. The internal pilot approach is applied to two oral cancer screening examples: one small one, where the correction is needed, and one large one, where no correction is needed. We conclude the manuscript with a discussion of the results.

Methods

Study design, hypothesis test, and sample size re-estimation

A novel internal pilot study design for screening trials

The novel internal pilot design includes the following steps:

1. Initial planning stage: Initial estimation of the sample size needed.
2. Pilot stage: Collection of paired screening test scores from a fraction of the planned sample size.
3. Re-estimation: Sample size re-estimation using pilot-sample based variance and prevalence estimates.
4. Additional data collection: Collection of additional data based on the sample size re-estimation.
5. Analysis: Hypothesis testing, using an adjusted critical value to prevent Type I error inflation.

We expand the notation of Coffey and Muller [9, 12] and Coffey et al. [19] to accommodate our modifications in the internal pilot study design. Throughout the manuscript lower case letters represent fixed variables and upper case letters represent random variables. Matrices are written in bold text.

Data for the internal pilot study can be organized into four sets according to the stage of the study that is of interest (Fig. 1). Let $k \in \{0, 1, 2, +\}$ index the stage of interest. Variables indexed by $k = 0$ describe the

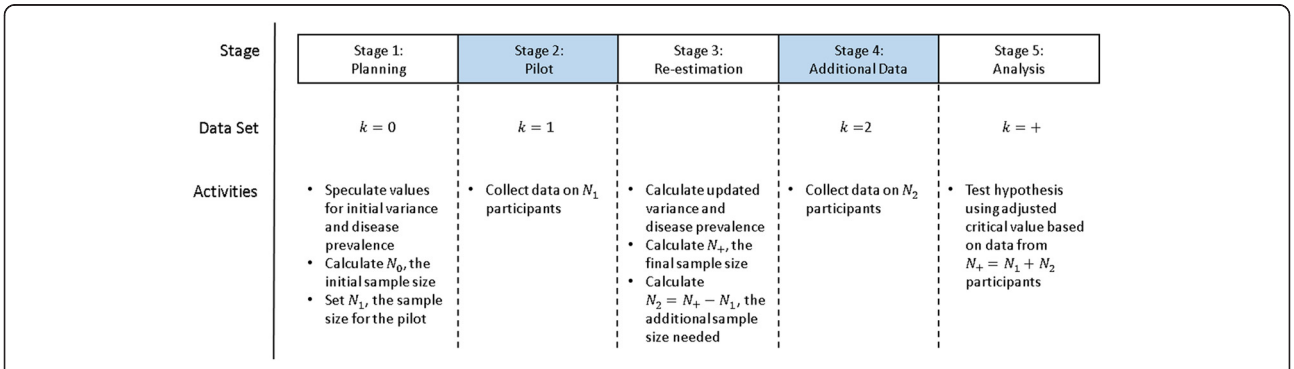

Fig. 1 Adapted internal pilot design for a cancer screening study

initial planning stage. Since no data has been collected, planning stage variables take on planned or speculated values. Variables indexed by $k = 1$ and $k = 2$ identify data observed in the pilot stage and the additional data collection stage, respectively. Variables indexed by $k = +$ describe the entire sample, which includes data from all participants.

Let the random variable N_{dk} be the number of study participants in stage k with disease status $d \in \{n, c\}$, with n indicating no disease, and c disease. For example, N_{c1} is the number of individuals with disease in the pilot sample. When the subscript d is dropped, the random variable N_k denotes the number of people both with and without disease in the kth stage of the study. For example, N_1 is the total number of individuals in the pilot sample, and N_+ is the final sample size.

Let $N_{\min}$ and $N_{\max}$ be the minimum and maximum sample sizes allowed by the study investigator, and assume that $N_+ \in [\min(N_1, N_{\min}), N_{\max}]$. Let n_0 be the initial sample size estimate, and define $\lambda = n_1/n_0$. Let $\gamma_\pi = \pi/\hat{\pi}_0$, where $\pi \in (0, 1)$ is the true prevalence of disease, and $\hat{\pi}_0 \in (0, 1)$ is the initial estimate of prevalence of disease. Let $\hat{\pi}_1 = n_{c1}/n_1$ be the estimate of prevalence of disease from the pilot data. With σ^2 the true variance of the difference in the two screening test scores, and $\hat{\sigma}_0^2 > 0$ the variance estimate used for the initial sample size calculation, define $\gamma = \sigma^2/\hat{\sigma}_0^2$. Let $\mathrm{SSE}_1 = \hat{\sigma}_1^2 \times (n_1 - 2)$ where $\hat{\sigma}_1^2$ represents the variance of the difference in the two screening test scores estimated after the internal pilot study. Let P_t and α_t be the target power and Type I error level for the study.

A paired comparison of the diagnostic accuracy of two screening tests

Let y_{idj} be the screening test score for individual $i \in \{1, 2, \ldots, N_+\}$, with disease status d, on screening test $j \in \{A, B\}$. Assume that the two screening test scores $[y_{idAk} y_{idBk}]'$ have a bivariate normal distribution with mean $\boldsymbol{\mu}_d = [\mu_{dA} \mu_{dB}]'$, $V(y_{idjk}) = \sigma_{dj}^2$, and $Cov(y_{idAk}, y_{idBk}) = \rho_d \sigma_{dA} \sigma_{dB}$. We assume that differences between the screening test scores for both the cases and non-cases are distributed with equal variance, $V(y_{inA} - y_{inB}) = V(y_{icA} - y_{icB}) = \sigma^2$. Under the bivariate normal assumption, the AUC for screening test j is given by $\Phi[(\mu_{cA} - \mu_{nA})/\sigma]$ ([20], p. 83, Result 4.8) where Φ is the cumulative distribution function of the standard normal. The difference between the AUCs is given by $\Phi[(\mu_{cA} - \mu_{nA})/\sigma] - \Phi[(\mu_{cB} - \mu_{nB})/\sigma]$.

For a paired comparison of the AUCs of the two screening tests, we test the hypothesis $H_0 : (\mu_{cA} - \mu_{nA}) - (\mu_{cB} - \mu_{nB}) = (\mu_{cA} - \mu_{cB}) - (\mu_{nA} - \mu_{nB}) = 0$ against $H_A : \neg H_0$. If H_0 holds, the AUCs, and hence the diagnostic accuracies of the two screening tests, are equal. To test H_0, we fit a general linear univariate model with the difference in the screening test scores as the outcome. The approach was inspired by the work of Demler et al. [21]. We assume that the difference between screening test scores is Gaussian and that the observations on different participants are independent.

The general linear univariate model for the final data set can be written as $\boldsymbol{Y}_+ = X_+ \boldsymbol{\beta} + \epsilon_+$, where $\boldsymbol{Y}_+$ is an $N_+ \times 1$ matrix containing the difference in the screening test scores for each individual, $[y_{idA} - y_{idB}]'$, X_+ is an $N_+ \times 2$ design matrix that identifies disease status, $\boldsymbol{\beta}$ is a 2×1 matrix of mean differences $[\mu_{cA} - \mu_{cB} \mu_{nA} - \mu_{nB}]'$, and ϵ_+ is the $N_+ \times 1$ matrix of errors. We test H_0 by writing the contrast matrix $\boldsymbol{C} = [1 - 1]$, forming $\boldsymbol{\theta} = \boldsymbol{C\beta}$, and using an F statistic ([22], p. 51, Equation 2.32). The final F statistic used in our adapted internal pilot design is written as F_+.

Sample size re-estimation for an internal pilot with unknown disease prevalence

The initial sample size is calculated as in Muller et al. [23]. For that calculation, the study investigator will specify σ_0^2 and β_0. Ideally, speculated values will be based on data from previous studies, closely related published results, or clinical experience.

After the internal pilot, the final sample size can be recalculated using the following iterative algorithm. The goal of the algorithm is to find N_+, where the power of the study is equal to P_t, the target power. First, check to see if the pilot data includes either all cases or all non-cases. If so, set $N_+ = n_0$. Otherwise, calculate the final sample size as follows. With n_{c1} and n_{n1} as observed in the initial pilot, define κ to be the greatest common factor of n_{c1} and n_{n1}. Let $D = n_{c1}/\kappa$, $E = n_{n1}/\kappa$, and $R = (D + E)$.

Speculate that $\boldsymbol{X}_+$ will take the form $\boldsymbol{X}_+ = \mathrm{Es}(\boldsymbol{X}) \otimes 1_m$, where $\mathrm{Es}(\boldsymbol{X})$ is an $(R \times 2)$ matrix such that

$$\mathrm{Es}(\boldsymbol{X}) = \begin{bmatrix} 1_D & 0_D \\ 0_E & 1_E \end{bmatrix}, \quad (1)$$

and m is a positive integer chosen so that $N_+ = mR \geq n_1$.

Calculate the power as $1 - \Pr[F_+ \leq f_{\mathrm{crit}}]$ [23], where $f_{\mathrm{crit}} = F_F^{-1}[(1 - \alpha_t); 1, N_+ - 2]$ and F_+ has a non-central F distribution with 1 numerator degrees of freedom, denominator degrees of freedom $N_+ - 2$, and non-centrality parameter $\omega_+ = \delta_+/\hat{\sigma}_1^2$, where $\delta_+ = (\boldsymbol{\theta} - \theta_0)' \left[\boldsymbol{C}\left(\boldsymbol{X}_+' \boldsymbol{X}_+\right)^{-} \boldsymbol{C}'\right]^{-1} (\boldsymbol{\theta} - \theta_0)$.

Sequentially increment or decrement m until the power of the experiment meets or exceeds P_t, at $m = m_t$. Set the final sample size to be $N_+ = m_t R$, unless $N_+ \geq N_{max}$ or $N_+ \leq N_{min}$. If $N_+ \geq N_{max}$ then set $N_+ = N_{max}$. If $N_+ \leq N_{min}$ then set $N_+ = N_{min}$. Finally, calculate N_2 as $N_2 = N_+ - n_1$.

Simulation studies

Verification of unconditional power

We conducted a simulation study designed to verify the result of Equation (10) below. Simulation study parameters came from modifying an example presented in Kairalla et al. [24]. Kairalla et al. [24] modified a balanced example in Wittes and Brittain [8] so that the numbers of cases to non-cases were unequal. Kairalla et al. then assumed a fixed case mixture throughout the study. We, in turn, modified the example in [24] by allowing the ratio of cases to non-cases to vary randomly.

Initial parameters were set at: $\boldsymbol{C} = [1{-}1]$, $P_t = 0.90$, $\alpha_t = 0.05$, $\beta = [1\ \ 0]'$, and $\sigma_0^2 = 2$. The resulting initial sample size was $n_0 = 96$ participants. With $\lambda = 0.5$, the pilot sample was fixed at $n_1 = 48$. The true rate of disease was set at $\pi = 1/3$. The parameter γ ranged between 0.5 to 2 by 0.25 while γ_π was fixed at 1. Under the alternative hypothesis, the bivariate normal parameters were set at $\mu_{c1} = 3$, $\mu_{c2} = 4$, $\rho_{c12} = 0$, $\mu_{n1} = 0$, $\mu_{n2} = 0$, $\rho_{n12} = 0$, and $\sigma_c^2 = \sigma_n^2 = 1$. To calculate Type I error under H_0, the bivariate normal parameters were set at $\mu_{c1} = 3$, $\mu_{c2} = 3$, $\rho_{c12} = 0$, $\mu_{n1} = 0$, $\mu_{n2} = 0$, $\rho_{n12} = 0$, and $\sigma_c^2 = \sigma_n^2 = 1$. The distributional parameters under the null correspond to an AUC of 0.983 and a difference in AUC of 0.015 under the alternative. All programs were written in version 9.3 of SAS/IML® software [25] and are available upon request. The empirical power was calculated as the proportion of times the null hypothesis was rejected. The experiment was repeated 10,000 times. The maximum absolute deviation (MAD) was calculated as the maximum absolute difference between the empirical estimates and the theoretical value. Using a normal approximation to the distribution of a proportion, the half-width of the 95 % CI for a target power of 0.90 is 0.0053.

Assessment of Type I error rate inflation

We conducted a simulation study to assess the magnitude of the Type I error rate inflation for a variety of experimental conditions. The Type I error rate was simulated for a prospective cancer screening trial with an internal pilot design. The disease prevalence and variance were either correctly or incorrectly specified and then re-estimated using pilot data. The hypothesis test was conducted using either an adjusted or unadjusted critical value.

The empirical Type I error was calculated for 648 different scenarios. The null hypothesis was that there was no difference in the diagnostic accuracy of the screening tests. For each scenario, we simulated 10,000 replicate data sets, conducted the hypothesis test, formed the *P*-value, and decided whether to accept or reject the null hypothesis at the $\alpha_t = 0.05$ level. The number of replicates was chosen so that the 95 % confidence interval of the proportion was no more than 0.005. The empirical Type I error was calculated as the proportion of replicates where the null hypothesis was rejected. For some scenarios, the study population was composed of either all cases or all non-cases. For all such scenarios, we considered there to be insufficient evidence to reject the null hypothesis.

The 648 different scenarios came from a range of parameter values. Parameters of the bivariate normal distributions for the cases and non-cases were fixed at $\boldsymbol{\mu}_n \in \{[0\ 0]'\}$, $\boldsymbol{\mu}_c \in \{[0.2\ 0.2]', [0.5\ 0.5]'\}$, $\sigma_0^2 \in \{0.34\}$, and $\rho_n = \rho_c = 0.5$. This corresponded to a difference in the AUCs of test *A* and test *B* of 0.05 or 0.1, respectively. The proportion of the initial sample size used for the internal pilot was in the range of $\lambda \in \{0.25, 0.5, 0.75\}$. We varied target power, $P_t \in \{0.80, 0.90\}$, the ratio of the true variance to the initial variance estimate, $\gamma \in \{0.5, 1, 1.5\}$, and the ratio of the true population disease prevalence to the initial prevalence estimate, $\gamma_\pi \in \{0.1, 1, 1.9\}$. The initial prevalence estimate was fixed at $\pi_0 = 0.5$, corresponding to a balanced study design.

Validation of Type I error control

We compared our adjusted method to an unadjusted internal pilot approach for a scenario where significant Type I error inflation occurred. The parameters that defined the scenario were $\boldsymbol{\mu}_n = \{[0\ 0]'\}$, $\boldsymbol{\mu}_c = \{[0.3\ 0.92]'\}$, $\sigma_0^2 \in \{0.34\}$, $\rho_n = \rho_c = 0.5$, $\pi = 0.5$, and $\lambda \in \{0.5\}$. The parameters correspond to an AUC of 0.64 for test *A* and an AUC of 0.87 for test *B*. We varied γ between 0.25 and 4. With $P_t = 0.90$ and $\alpha = 0.05$, the initial sample size was 42. The adjusted method was applied to each of three possible prevalence misspecification scenarios with $\gamma_\pi \in \{0.1, 1, 1.9\}$.

Results

Type I error rate control

Overview

In general, internal pilot studies can inflate Type I error rate [14]. Here, we describe a method to bound Type I error rate in internal pilot studies where both the variance of the outcome and the disease prevalence are re-estimated in the internal pilot step. First, we give the unconditional power and hence the Type I error for the *F* test statistic. We uncondition over all possible realizations of N_1, N_{c1}, N_{c2}, and N_2. After demonstrating that the Type I error rate takes on a maximum value across a specified range of γ and γ_π, we describe a method for identifying the values of γ and γ_π at which the maximum occurs. We choose a critical value for the final hypothesis test so that the maximum Type I error rate is bounded.

Unconditional Type I error

We derive the distribution of the F_+ statistic under H_0 and H_A. Under H_0, the formulae give an unconditional Type I error. Under H_A, the formulae give unconditional power. Because both the variance and the disease prevalence are re-estimated, the test statistic is a function of the pilot sample size and the final sample size. Derivation of the distribution of the test statistic requires obtaining three results:

1. The distributions of N_1, N_{c1}, N_{c2}, N_2, and N_+.
2. The distribution of F_+ conditional on N_1, N_{c1}, N_{c2}, N_2, and N_+.
3. The unconditional Type I error and power of the F_+ test statistic.

Under the Type I error rate control subsection each of the three afore mentioned results are presented. Throughout the this subsection we find it useful to use functional notation to emphasize the dependence of variables on N_1, N_{c1}, N_{c2}, N_2, and σ_1^2. For example, we write $N_2(\sigma_1^2, N_{c1}, N_1)$ to indicate that the additional sample size is a function of the pilot variance and the pilot case mixture.

Distributions of , N_1, N_{c1}, N_{c2}, and N_2

The number of participants in the pilot sample is fixed by study design: $n_1 = \lambda n_0$. Assuming a true disease prevalence of π, $N_{c1} \sim \text{Binomial}(n_1, \pi)$ and $N_{c2} \sim \text{Binomial}(N_2, \pi)$. The random variables $\hat{\sigma}_1^2$ and N_{c1} are distributed independently. Summing over all possible values of n_{c1}, the unconditional probability mass distribution of the additional sample is:

$$\begin{aligned}\Pr\{N_+ = n_+\} &= \sum_{n_{c1i}=0}^{n_1} \Pr\{N_+ = n_+ | N_{c1i} = n_{c1i}\} \times \Pr\{N_{c1i} = n_{c1i}\} \\ &= \sum_{n_{c1i}=0}^{n_1} \left(\Pr\{N_+ \le n_+ + 1 | n_{c1i}\} - \Pr\{N_+ \le n_+ | n_{c1i}\}\right) \\ &\qquad \times \Pr\{N_{c1i} = n_{c1i}\}, \end{aligned} \tag{2}$$

where the first line extends Equation 18 of [9], and the second line follows from the law of total probability. The conditional probability mass function of N_+ is calculated by extending Equation 17 of [9] as follows:

$$\Pr\{N_+ \le n_+ | n_{c1i}\} = \Pr\left\{\chi^2(n_1 - 2) \le \frac{(n_1 - 2)}{\sigma^2} \frac{\delta_+}{\omega_+} | N_{c1i} = n_{c1i}\right\}. \tag{3}$$

Note that since $N_2 = N_+ - n_1$,

$$\Pr\{N_+ \le n_+ | n_{c1i}\} = \Pr\{N_2 \le n_2 | n_{c1i}\}. \tag{4}$$

Power of the final hypothesis test conditional on N_1, N_{c1}, N_{c2}, N_2, and N_+

We show the dependence of the power on N_1, N_{c1}, N_{c2}, and N_2.

The additional sample size N_2 is a function of $\hat{\sigma}_1^2$ and N_{c1}. Since the power function is strictly monotone increasing, for fixed values of $\hat{\sigma}_1^2$, n_1, and n_{c1}, there exists one and only one $N_2 = n_2$. However, for a fixed n_1 and n_{c1}, there exist infinitely many $\hat{\sigma}_1^2$, all of which would yield the same final sample size.

Let $q_1(n_2, n_{c1})$ and $q_2(n_2, n_{c1})$ represent the smallest and the largest value of $\hat{\sigma}_1^2$ that would lead to the additional sample size n_2 for a fixed n_1 and n_{c1}. Let $q(n_2, n_{c1})$ be the value of $\hat{\sigma}_1^2$ that falls in the interval $(q_1(n_2, n_{c1}), q_2(n_2, n_{c1})]$.

We can express the approximate power of the F_+ test statistic for a value $f(n_2, n_{c2}, n_{c1})$ as a function of n_2, n_{c2}, and n_{c1}. Let $I(n_2, n_{c2}, n_{c1})$ represent the probability of rejecting H_0 when the alternative is true, conditional on n_{c2}, n_{c1} and the value $q(n_2, n_{c1})$. Then

$$\begin{aligned} I(n_2, n_{c2}, n_{c1}) &= 1 - \Pr\{F_+ \le f(n_2, n_{c2}, n_{c1}) | q(n_2, n_{c1}), n_{c2}, n_{c1}\} \\ &= 1 - \Pr\{c(n_2, n_{c2}, n_{c1}) \cdot \chi^2[a, \omega_+(n_1 + n_2, n_{c2} + n_{c1})] \\ &\qquad - \chi^2(n_2) \le q(n_2, n_{c1}) | q(n_2, n_{c1}), n_{c2}, n_{c1}\}, \end{aligned} \tag{5}$$

where $\nu_+ = N_+ - 2$, $c(n_2, n_{c2}, n_{c1}) = \nu_+ / [2f(n_2, n_{c2}, n_{c1})]$ with $\chi^2[a, \omega_+(n_1 + n_2, n_{c2} + n_{c1})]$ denoting a non-central χ^2 with a degrees of freedom and a non-centrality parameter of $\omega_+(n_1 + n_2, n_{c2} + n_{c1})$. Equation (5) follows from the proof in the Appendix of Coffey and Muller [9].

Expected power of the F test statistic unconditioned from N_1, N_{c1}, N_{c2}, N_2, and N_+

We uncondition Equation (5) from N_{c1}, $q(n_2, n_{c1})$, N_{c2}, and N_2. Using the law of total probability, the unconditional power is

$$I(n_2, n_{c2}) = 1 - \sum_{n_{c1i}=0}^{n_1} I(n_2, n_{c2}, n_{c1i}) \times Pr[N_2 = n_2 | n_{c1i}] \times Pr[N_{c1} = n_{c1i}]. \tag{6}$$

Substituting Equation (6) into Equation (5) gives

$$\begin{aligned} I(n_2, n_{c2}) &= 1 - \sum_{n_{c1i}=0}^{n_1} \Pr\{Q(n_2, n_{c2}, n_{c1i}) \le q(n_2 | n_{c1i}) | q(n_2 | n_{c1i}), n_{c2}, n_{c1i}\} \\ &\qquad \times Pr[N_2 = n_2 | n_{c1i}] \times Pr[N_{c1} = n_{c1i}]. \end{aligned} \tag{7}$$

Unconditioning the power from N_{c2}, we obtain

$$I(n_2) = 1 - \sum_{n_{c2i}=0}^{n_2} I(n_2,\ n_{c2i}) \times Pr[N_{c2} = n_{c2i}|n_2], \qquad (8)$$

leading to

$$\begin{aligned} I(n_2) &= 1 - \sum_{n_{c2i}=0}^{n_2} \Big(\sum_{n_{c1i}=0}^{n_1} Pr\{Q(n_2, n_{c2}, n_{c1i}) \le q(n_2|n_{c1i}) | q(n_2|n_{c1i}),\ n_{c2}, n_{c1i}\} \\ &\quad \times Pr[N_2 = n_2|n_{c1i}] \times Pr[N_{c1} = n_{c1i}]) \times Pr[N_{c2} = n_{c2i}|n_2] \\ &= 1 - \sum_{n_{c2i}=0}^{n_2} \Big(\sum_{n_{c1i}=0}^{n_1} \int_{q_1(n_2,\ n_{c1})}^{q_2(n_2,\ n_{c1})} Pr\{Q\Big(n_2, c(n_2,\ n_{c2i}, n_{c1i}), \delta(n_{c2i}, n_{c1i})\Big) \le t\} \\ &\quad \times \frac{f_{\chi^2}(t, \nu_1)}{Pr\{N_2 = n_2|n_{c1i}\}} dt \times Pr[N_2 = n_2|n_{c1i}] \times Pr[N_{c1} = n_{c1i}]) \\ &\quad \times Pr[N_{c2} = n_{c2i}|n_2], \end{aligned} \qquad (9)$$

with $f_{\chi^2}(t, \nu_1)$ defined in Johnson et al. [26]. The distributional results of Coffey et al. [19] hold, conditional on fixed values of N_1, N_{c1}, and N_{c2}. The expected power is given by

$$\begin{aligned} &Pr\{F_+(N_+, N_{c+}, N_{c1}) \le f(N_+, N_{c+}, N_{c1})\} \\ &= 1 - \sum_{n_{+i}=n_1}^{n_1+n_2} \sum_{n_{+ci}=n_{c1}}^{n_{c1}+n_{c2}} \sum_{n_{c1i}=0}^{n_1} \int_{q_{1(n_{2i},\ n_{c1i})+}}^{\infty} F_{\chi^2}\left[\frac{z}{c(n_{c+i})_+}; 2, \frac{\delta(n_{c+i})_+}{\gamma\sigma_0^2}\right] f_{\chi^2}(z; \nu_+) \\ &\quad \times F_\beta\Big(\frac{q_{2(n_{2i},\ n_{c1i})}}{z}; \frac{\nu_1}{2}, \frac{n_{2i}}{2}\Big) - F_\beta\Big(\frac{q_{1(n_{2i},\ n_{c1i})+}}{z}; \frac{\nu_1}{2}, \frac{n_{2i}}{2}\Big) dz \\ &\quad \times Pr[N_{c1} = n_{c1i}] \times Pr[N_{c2} = n_{c2i}|n_{2i}], \end{aligned} \qquad (10)$$

where $F_+(N_+, N_{c\ +}, N_{c1})$ is the final test statistic, $f(N_+, N_{c\ +}, N_{c1})$ is an observed value, F_{χ^2} is the cumulative distribution function of a non-central χ^2 [27], F_β is the cumulative distribution function of a beta (one) distributed random variable [27], $\nu_1 = n_1 - 2$, and the bounds of the integration depend on n_{c1} and n_2. The Type I error can be calculated from Equation (10) when the null hypothesis is true. Notice that when the null hypothesis is true, the χ^2 distribution in Equation (10) becomes a central χ^2.

Bounding Type I error

There exists a maximum Type I error across a specified range of γ and γ_π. Let α^{max} be the global maximum Type I error. Power for a study design is maximized when the ratio of the number of study participants with disease to the number of study participants without disease is one-to-one. Thus, α^{max} must occur for $\gamma_\pi = 1$. The problem of showing that there is a maximum then reduces to showing that there exists a maximum with respect to γ for $\gamma_\pi = 1$. Coffey and Muller [12] provide evidence to support this assertion.

We propose the following method to find the $\gamma = \gamma^*$ and $\gamma_\pi = \gamma_\pi^b$ for which the maximum Type I error occurs:

9. First, fix a range for $\gamma \in [a, b]$ and $\gamma_\pi \in [c, d]$ a priori, based on the previous literature.
10. Find the value of $\gamma_\pi = \gamma_\pi^b$ that results in a study design with a permissible prevalence value that is closest to a one-to-one ratio of cases to non-cases (that is, the value closest to $1 \in [c, d]$).
11. Finally, for a fixed γ_π^b, find the value of $\gamma = \gamma^*$ that yields the maximum Type I error inflation, using Equation (10) and a golden section search algorithm [28].

The maximum Type I error is bounded by identifying an adjusted critical value for the final test statistic. For $\gamma = \gamma^*$ and $\gamma_\pi = \gamma_\pi^b$ we use a bisection search algorithm to find α^* so that under H_0, $\Pr\{F_+(N_+, N_{c\ +}, N_{c1}) \le f_{\text{adj}}\} = \alpha_t$, where $f_{\text{adj}} = F_F^{-1}[(1 - \alpha^*); 1, N_+ - 2]$.

Simulation studies results

Verification of unconditional power

The simulation study suggested that for the parameters chosen, Equation (10) provides a good estimate of unconditional power. The MAD between predicted empirical power and theoretical power always fell within the 95 % confidence interval (Tables 1 and 2). The half-width of the 95 % CI for a target Type I error of 0.05 is 0.0043.

Assessment of Type I error rate inflation

Results from the simulation are presented in Figs. 2, 3, and 4. Overall, the Type I error rate was inflated when the initial sample size was smaller than 50 and the initial prevalence estimate was correct. As the fraction of the initial sample size estimate used in the pilot study increased, the inflation grew smaller. The initial sample sizes for all 648 scenarios ranged from 12 to 2,028 participants, with an interquartile range of 61 to 635 participants. The median observed Type I error was 0.0495, with a minimum of 0.0244, a maximum of 0.0839, and an interquartile range of 0.0479 to 0.0521.

The figures suggest that no Type I error adjustment is needed when the sample size is large. This observation

Table 1 Empirical versus theoretical power by variance misspecification

γ	Empirical power	Theoretical power	Absolute deviation
0.5	0.9945	0.995	0.0005
0.75	0.9646	0.964	0.0006
1	0.9369	0.934	0.0029
1.25	0.9168	0.922	0.0052
1.5	0.9144	0.916	0.0016
1.75	0.9137	0.913	0.0007
2	0.9136	0.910	0.0036

Table 2 Empirical versus theoretical Type I error by variance misspecification

γ	Empirical type I error	Theoretical type I error	Absolute deviation
0.5	0.0505	0.053	0.0025
0.75	0.0487	0.053	0.0043
1	0.0496	0.052	0.0024
1.25	0.0505	0.052	0.0015
1.5	0.0514	0.052	0.0006
1.75	0.0478	0.051	0.0032
2	0.0507	0.051	0.0003

is consistent with the results from Wu et al. [11]. The results from the simulation study by Wu et al. [11] correspond to the subset of results in Figs. 2, 3, and 4 with $\gamma = 1$ and $\gamma_\pi = 1$. However, Wu et al. [11] did not consider cases with small initial sample sizes, and thus did not observe the Type I error rate inflation shown in our results. In our first example, we present an application with a large sample size where no adjustment is needed to bound the Type I error rate.

Validation of Type I error control

Results from the Type I error control simulation appear in Fig. 5, which shows a comparison of the Type I error inflation for the adjusted and unadjusted methods. The figure plots Type I error rate as a function of γ, cross-classified by γ_π for the two methods. Figure 5 shows that the adjusted method controls the Type I error rate in small samples. The maximum possible Type I error occurred with $\gamma_\pi^b = 1$, $\gamma^* = 0.8541$ for a Type I error of 0.0564. The adjusted Type I error rate was $\alpha^* = 0.0438$. Note that f_{adj} is only assigned a value after the pilot sample is collected and $N_+ = n_+$ is re-estimated.

Applications

Example 1: A large oral cancer screening trial where no adjustment is needed

One implication of this study is that internal pilot designs often require no penalty for re-estimating both outcome variance and disease prevalence. In addition, the internal pilot design ensures that researchers will have sufficient power.

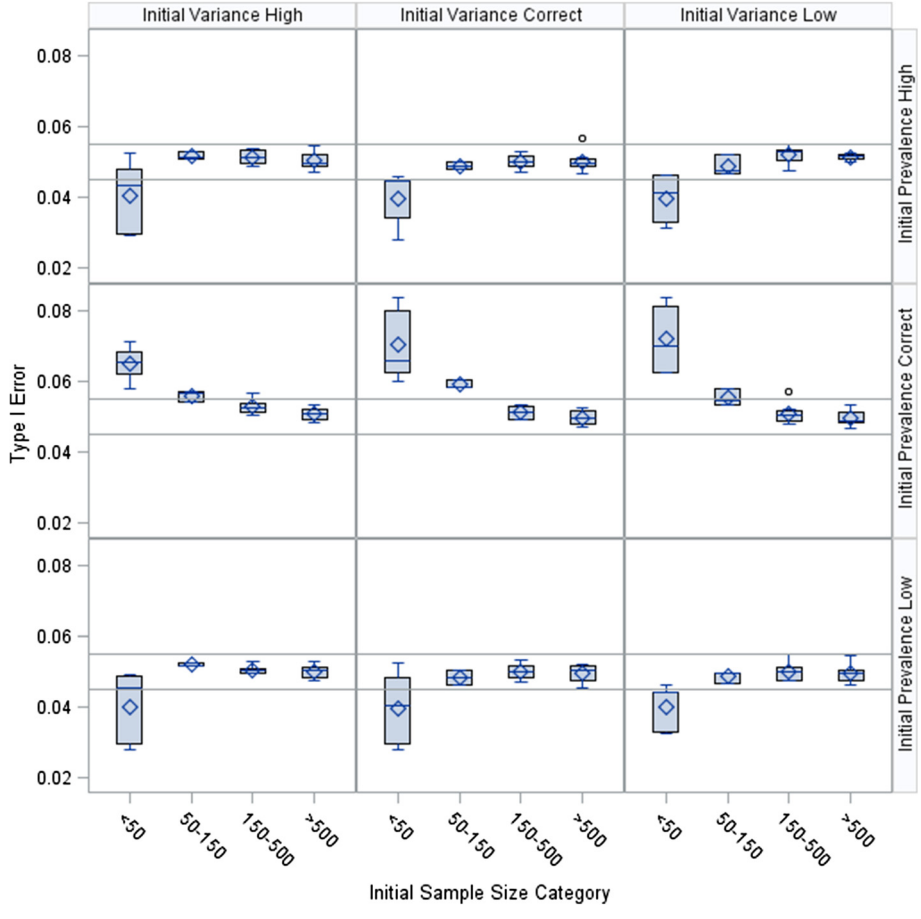

Fig. 2 Type I error rate by scenario with the pilot study size at 25 % of initial sample size estimate

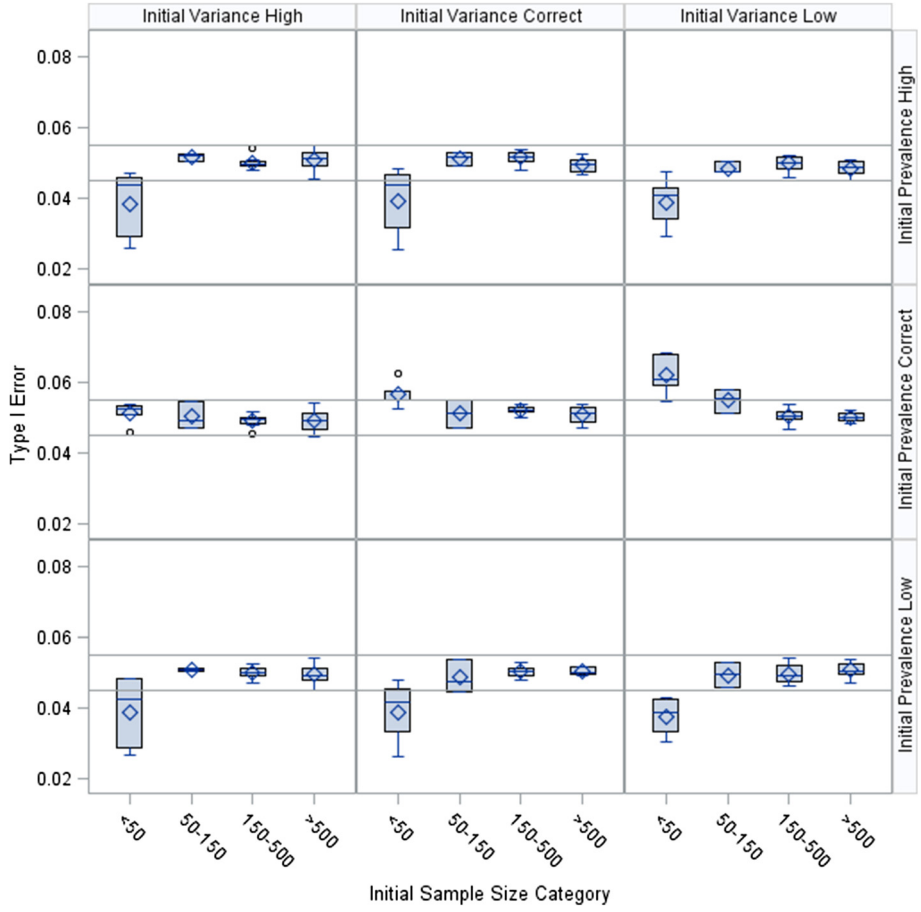

Fig. 3 Type I error rate by scenario with the pilot study size at 50 % of initial sample size estimate

Recall the study by Lingen et al. discussed in the Background section. One aim of the study was to compare the diagnostic accuracy of a combined modality involving both visual and tactile oral exam with VELscope® [29]. The investigators wished to detect oral pre-malignancy and malignancy. There was substantial uncertainty about the rate of oral pre-malignancy and malignancy in the target population. The rate of suspicious lesions varies widely in Western populations, ranging from 0.2 % to 16.7 % [5]. Further, the variance of scores for visual and tactile oral exam and for examinations with VELscope was largely unknown. The uncertainty made an internal pilot design attractive.

One critical step for designing an internal pilot study is choosing $N_{\min}$ and $N_{\max}$. The investigators wished to estimate a confidence interval for the percentage of oral lesions that were benign. To ensure that the confidence interval had a half-width of no more than 0.1 %, the investigators had to make sure that the entire study enrolled at least 96 people with lesions. If the rate of suspicious lesions was about 12.1 %, the minimum sample size could be no less than 800. The upper bound on sample size was fixed by monetary constraints. Previous experience had shown that a sample size of more than 30,000 was fiscally unfeasible. This set $N_{\max}$ at 30,000.

The initial power calculation was based on plausible values from the literature. A conservative estimate for the AUC for visual and tactile oral exam is 0.60. A clinically interesting difference between AUCs is 0.06. This corresponds to $\boldsymbol{\mu}_n \in \{[0\ 0]'\}$, $\boldsymbol{\mu}_c \in \{[0.359\ 0.584]'\}$, $\sigma = 1$, and $\rho_n = \rho_c = 0$. Assuming that the rate of suspicious lesions in the population is 12.1 %, the initial sample size needed for 95 % power is 2,156 non-cases and 294 cases for a total sample size of 2,450.

The final sample size that would be needed for the study would depend on results from the internal pilot. The results presented in the Type I error control validation indicate that Type I error inflation would not be a

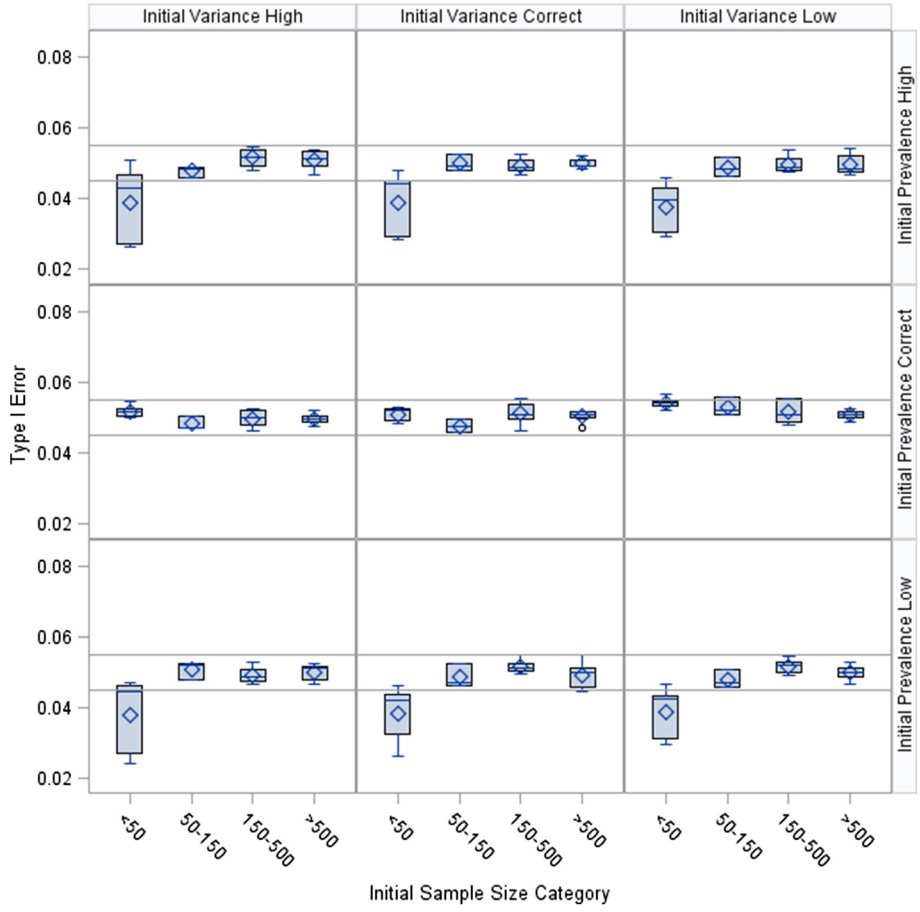

Fig. 4 Type I error rate by scenario with the pilot study size at 75 % of initial sample size estimate

problem for a study designed with an initial sample size of 2,450. Thus, the final hypothesis test could be carried out with α set to 0.05.

Example 2: A small oral cancer screening trial where adjustment prevents Type I error inflation

A second implication of this manuscript is that internal pilot designs with small sample size require an adjustment to prevent Type I error inflation. Small sample sizes often occur because of biological constraints. For example, Wong et al. [30] are currently recruiting for an oral cancer screening trial in people with Fanconi anemia. Fanconi anemia is a rare genetic disease that occurs in roughly 1 in 131,000 people in the United States. People with Fanconi anemia are at increased risk for oral cancer, although the magnitude of the risk is unknown. The prevalence of oral squamous cell carcinoma could be as high as 100 % or as low as 3 % [31, 32].

Because the study is still in progress, the design has not yet been published. To illustrate the results of our manuscript, we show how an internal pilot trial might be used to compare the diagnostic accuracy of two assays for IL-8 for the prediction of oral cancer. In people with Fanconi anemia, IL-8 is a useful biomarker for screening for oral cancer [33, 34].

Consider a trial in which people with Fanconi anemia are given two salivary assays: a salivary bead-based assay for IL-8, and an enzyme-linked immunosorbent assay (ELISA). The diagnostic accuracy (AUC) of the ELISA and the salivary bead-based assay is 0.85 and 0.94, respectively [34, 35]. The target power is set to 0.80. A clinically interesting difference in diagnostic accuracy is a difference between AUCs of 0.09. The target Type I error rate is 0.05. Means and variances of both ELISA and a salivary bead-based assay are available in the literature [34, 35], with $\boldsymbol{\mu}_n \in \{[759.4\ 759.4]'\}$, $\boldsymbol{\mu}_c \in \{[3347.7\ 4700.0]'\}$, and $\sigma_{nA} = \sigma_{nB} = \sigma_{cA} = \sigma_{cB} = 3328174.5$. Modest correlation is set at $\rho_n = \rho_c = 0.5$.

If half the people in the study have oral cancer, the initial sample size required is 84 participants. Thus, the study could be subject to Type I error inflation. If we re-estimate the sample size after the first 42 participants have been collected, the study could have a Type I error rate inflated to 0.054. This inflation occurs at $\gamma_\pi^b = 1$ and

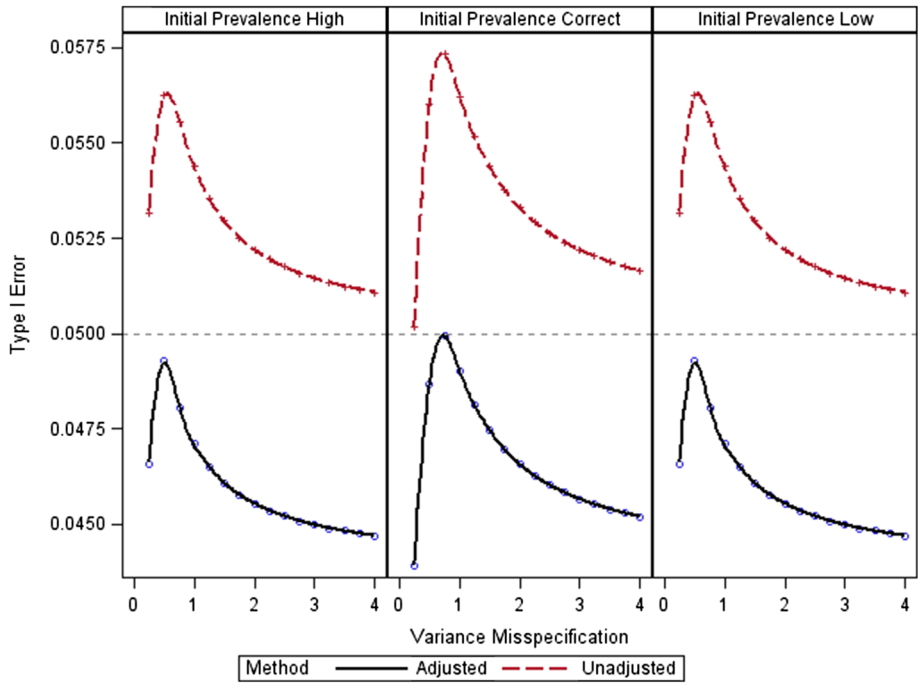

Fig. 5 Bounded Type I error

$\gamma^* = 0.7254$. This is an 8 % inflation from the target Type I error rate of 0.05. Adjusting gives an adjusted alpha level of $\alpha_{adj} = 0.0463$. The adjusted critical value can be calculated as $f_{adj} = F_F^{-1}[(1 - \alpha^*); 1, N_+ - 2]$. Recall that the actual adjusted critical value will depend on the final sample size calculated after the internal pilot is observed. For example, if $n_+ = 100$, then $f_{adj} = 4.07$. Thus with $n_+ = 100$, any observed test statistic larger than 4.07 should be rejected.

Discussion

In this manuscript, we describe an internal pilot approach for cancer screening trials when the disease prevalence is unknown. We demonstrated that conducting an internal pilot study without adjusting the critical value caused Type I error rate inflation in small (N <50) samples, but not in large samples. We also demonstrated that our adjusted method controlled Type I error rate in small samples.

The approach has both strengths and limitations. A strength is that the method allows investigators to obtain expected power at least as high as needed, for all but the most rampant variance and prevalence misspecifications. One limitation is the assumption that the screening test scores have a bivariate normal distribution of the test scores and that the assumptions of the general linear univariate model [22] are met. Secondly, the method may be overly conservative, and result in a Type I error rate lower than nominal. However, for prospective cancer screening trials, being conservative is reasonable. Cancer screening methods may be adopted in large populations, and replicable research is vital for maintaining public trust. Finally, the computing time is somewhat lengthy, because the integration and sums from Equation (10) have high complexity. For any one study design, the amount of time is reasonable. For example, it took less than eight hours to run all programs used in Example 2. In addition, our simulation study demonstrated that the method is not necessary in screening studies with large sample sizes.

Conclusion

We have shown that an internal pilot approach usually achieves goal power, and, for most studies with sample size greater than 50, requires no Type I error correction. Further, we have provided a flexible and accurate approach to bound Type I error below a goal level for studies with small sample size ($N < 50$). Both investigators and statisticians should use the new methods for the design of cancer screening trials.

Abbreviations

AUC: area under the receiver operating characteristic curve; MAD: maximum absolute deviation; ELISA: enzyme-linked immunosorbent assay.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JTB conducted the literature review, derived the mathematical results, designed and programmed the simulation studies, interpreted the results, and prepared the manuscript. DHG assisted with the literature review,

assisted with the mathematical derivations, provided guidance for the design and programming of the simulation studies, and provided expertise on the context of the topic in relation to other work in the field. BMR reviewed the intellectual content of the work and gave important editorial suggestions. DHG conceived of the topic and guided the development of the work. All authors read and approved the final manuscript.

Acknowledgements
The research presented in this paper was supported in part by NIDCR RC2DE020779 and by NIDCR 1 R01 DE020832-01A1. The content of this paper is solely the responsibility of the authors, and does not necessarily represent the official views of the National Institute of Dental and Craniofacial Research, nor the National Institutes of Health. This manuscript was submitted to the Department of Biostatistics and Informatics in the Colorado School of Public Health, University of Colorado Denver, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biostatistics for J. T. Brinton.

Author details
[1]Denver Health Medical Center, 777 Bannock St., MC 6551, Denver, Colorado 80204, USA. [2]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, 13001 E. 17th Place, Aurora, Colorado 80045, USA.

Received: 27 October 2014 Accepted: 3 August 2015
Published online: 13 October 2015

References

1. Lingen MW. Efficacy of oral cancer screening adjunctive techniques. Bethesda (MD): National Institute of Dental and Craniofacial Research, National Institutes of Health, US Department of Health and Human Services (NIH Project Number: 1RC2DE020779-01); 2009.
2. Berg W, Zhang Z, Lehrer D, Jong R, Pisano E, Barr R, et al. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. JAMA. 2012;307(13):1394–404.
3. Lewin JM, Hendrick RE, D'Orsi CJ, Isaacs PK, Moss LJ, Karellas A, et al. Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. Radiology. 2001;218(3):873–80.
4. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med. 2005;353(17):1773–83.
5. Lim K, Moles DR, Downer MC, Speight PM. Opportunistic screening for oral cancer and precancer in general dental practice: results of a demonstration study. Br Dent J. 2003;194(9):497–502. discussion 493.
6. Field EA, Morrison T, Darling AE, Parr TA, Zakrzewska JM. Oral mucosal screening as an integral part of routine dental care. Br Dent J. 1995;179(7):262–6.
7. Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance. Ann Math Stat. 1945;16(3):243–58.
8. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. Stat Med. 1990;9(1–2):65–71. discussion –2.
9. Coffey CS, Muller KE. Exact test size and power of a Gaussian error linear model for an internal pilot study. Stat Med. 1999;18(10):1199–214.
10. Friede T, Kieser M. Sample size recalculation in internal pilot study designs: a review. Biom J. 2006;48(4):537–55.
11. Wu C, Liu A, Yu KF. An adaptive approach to designing comparative diagnostic accuracy studies. J Biopharm Stat. 2008;18(1):116–25.
12. Coffey CS, Muller KE. Controlling test size while gaining the benefits of an internal pilot design. Biometrics. 2001;57(2):625–31.
13. Gurka MJ, Coffey CS, Gurka KK. Internal pilots for observational studies. Biom J. 2010;52(5):590–603. doi:10.1002/bimj.201000050.
14. Wittes J, Schabenberger O, Zucker D, Brittain E, Proschan M. Internal pilot studies I: Type I error rate of the naive t-test. Stat Med. 1999;18(24):3481–91.
15. Zucker DM, Wittes JT, Schabenberger O, Brittain E. Internal pilot studies II: comparison of various procedures. Stat Med. 1999;18(24):3493–509.
16. Miller F. Variance estimation in clinical studies with interim sample size reestimation. Biometrics. 2005;61(2):355–61.
17. Denne JS, Jennison C. Estimating the sample size for a t-test using an internal pilot. Stat Med. 1999;18(13):1575–85. doi:10.1002/(SICI)1097-0258(19990715)18:13<1575::AID-SIM153>3.0.CO;2-Z.
18. Kieser M, Friede T. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. Stat Med. 2000;19(7):901–11. doi:10.1002/(SICI)1097-0258(20000415)19:7<901::AID-SIM405>3.0.CO;2-L.
19. Coffey CS, Kairalla JA, Muller KE. Practical methods for bounding Type I error rate with an internal pilot design. Commun Stat Theory Methods. 2007;36(11):2143–57.
20. Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York, NY: Oxford University Press; 2003.
21. Demler OV, Pencina MJ, D'Agostino RB. Equivalence of improvement in area under ROC curve and linear discriminant analysis coefficient under assumption of normality. Stat Med. 2011;30(12):1410–8.
22. Muller KE, Stewart PW. Linear model theory: univariate, multivariate, and mixed models. New York: Wiley-Interscience; 2006.
23. Muller KE, LaVange LM, Ramey SL, Ramey CT. Power calculations for general linear multivariate models including repeated measures applications. J Am Stat Assoc. 1992;87(420):1209–26.
24. Kairalla JA, Coffey CS, Muller KE. GLUMIP 2.0: SAS/IML software for planning internal pilots. J Stat Softw. 2008;28(7):1–32.
25. Inc. SI. SAS/STAT® 9.3 User's Guide. SAS Institute Inc., Cary, NC. 2011.
26. Johnson NL, Kotz S, Balakrishnan N. Continuous univariate distributions, vol. 1. New York: Wiley-Interscience; 1994.
27. Johnson NL, Kotz S, Balakrishnan N. Continuous univariate distributions, vol. 2. New York: Wiley-Interscience; 1995.
28. Thisted RA. Elements of statistical computing: NUMERICAL COMPUTATION. Ipswich, Suffolk: Chapman and Hall/CRC; 1988.
29. Poh CF, MacAulay CE, Zhang L, Rosin MP. Tracing the "at-risk" oral mucosa field with autofluorescence: steps toward clinical impact. Cancer Prev Res. 2009;2(5):401–4.
30. Wong DT. Oral cancer biomarker study. 2012.
31. Scheckenbach K, Wagenmann M, Freund M, Schipper J, Hanenberg H. Squamous cell carcinomas of the head and neck in Fanconi anemia: risk, prevention, therapy, and the need for guidelines. Klin Padiatr. 2012;224(3):132–8.
32. Rosenberg PS, Socie G, Alter BP, Gluckman E. Risk of head and neck squamous cell cancer and death in patients with Fanconi anemia who did and did not receive transplants. Blood. 2005;105(1):67–73.
33. Elashoff D, Zhou H, Reiss J, Wang J, Xiao H, Henson B, et al. Prevalidation of salivary biomarkers for oral cancer detection. Cancer Epidemiol Biomarkers Prev. 2012;21(4):664–72.
34. Hu S, Arellano M, Boontheung P, Wang J, Zhou H, Jiang J, et al. Salivary proteomics for oral cancer biomarker discovery. Clin Cancer Res. 2008;14(19):6246–52.
35. Arellano-Garcia M, Hu S, Wang J, Henson B, Zhou H, Chia D, et al. Multiplexed immunobead-based assay for detection of oral cancer protein biomarkers in saliva. Oral Dis. 2008;14(8):705–12.