8

STUDY DESIGN AND HYPOTHESIS TESTING

8.1 The phases of medical test development

8.1.1 *Research as a process*

The development of a medical test is a process. At the beginning of the process there are small exploratory studies that seek to identify how best to apply the test and whether or not it has potential for use in practice. At the end of the process there are studies that seek to determine the value of the test when applied in particular populations. In this chapter we first outline the series of research steps involved in developing a test.

It is important to have this 'big picture' in mind when designing a particular study. Where the study fits into the development process is critical for defining appropriate scientific objectives for a study, and consequently its design and evaluation. We categorize here the development process for a medical test into five distinct phases. Later in this chapter we will discuss sample size calculations. It facilitates our discussion of sample size calculations to consider them separately for studies in each phase.

Those familiar with therapeutic research will recognize that there is already an analogous well-established paradigm for the development of a new therapeutic agent. The research process for therapeutic drugs is categorized into five phases: a preclinical testing phase, three clinical phases prior to regulatory approval, known as phases 1, 2 and 3, and a post-marketing surveillance phase, sometimes called phase 4. The process is so well established that regulatory agencies in Europe, the United States and Japan have outlined a joint document with guidelines for study design and evaluation at each phase (ICH, 1999). Preclinical testing involves *in vitro* and animal studies of toxicity and biologic efficacy. Phase 1 studies typically involve establishing pharmacokinetic profiles, toxicity parameters and preliminary measures of biologic efficacy in humans. Appropriate doses, routes and regimens for administering the drug are also determined in phase 1. Phase 2 studies evaluate biologic efficacy. That is, the effects of the treatment on biologic measures of disease, which are supposedly targeted by the drug, are determined. If the treatment is successful in phase 2 then a comprehensive and usually large phase 3 study is undertaken to determine if the treatment is better than existing therapies in ways that tangibly benefit the patient. Clinical efficacy in phase 3 is often defined by mortality or quality of life parameters. A treatment approved in phase 3 for marketing will need to be monitored for low-frequency adverse effects that occur when the treatment is made available on a large scale. Such effects observed in this so-called post-marketing phase 4 may not be apparent in

the studies at earlier phases conducted with limited sample sizes.

Categorizing the research development process for therapeutic agents has lead to widely accepted standards for study design and evaluation, and a development process that is regarded as reasonably rigorous and efficient. In a similar vein, categorizing the phases of development for a medical test can help clarify study objectives and streamline the development process, potentially making the process more rigorous and efficient.

8.1.2 *Five phases for the development of a medical test*

The phases shown in Table 8.1 were proposed for cancer biomarker research (Pepe *et al.*, 2001). We adapt this structure here for more general tests. Although this paradigm will not apply exactly to all tests, the basic structure is useful to consider for many tests.

Phase 1 is the initial phase. Its purpose is basically exploratory, to see if the test might be worth developing and evaluating rigorously. It is therefore appropriate to study the test in a wide range of circumstances (Guyatt *et al.*, 1986). Subjects with a variety of characteristics should be tested. In particular, subjects with diverse manifestations and severities of disease are tested. Non-diseased subjects with conditions that might be confused with disease should be tested in order to gain some insight into the limitations of the test for distinguishing disease from non-disease. The test should be implemented in a variety

Table 8.1 *Phases of research for the development of a medical test*

Phase	Description	Typical objectives	Typical design
1	Exploratory investigations	Identify promising tests and settings for application	Case-control study with convenience sampling
2	Retrospective validation	Determine if minimally acceptable $(\mathrm{FPF}_0, \mathrm{TPF}_0)$ are achieved	Population-based case-control sampling
3	Retrospective refinement	Define criteria for screen positivity. Determine covariates affecting $S_{\bar{D}}$ and ROC. Compare promising tests. Develop algorithms for combining tests	Large-scale comprehensive population-based case-control study
4	Prospective application	Determine positive predictive values, detection and false referral probabilities when the test is applied in practice	Cohort study
5	Disease impact	Determine effects of testing on cost and mortality associated with disease	Randomized prospective trial comparing new test with standard of practice

of settings and by a variety of test operators. The operating parameters for the test can be varied. For example, the frequency and/or intensity of the auditory stimulus might be varied for a hearing test. The protocol for collecting and storing the clinical specimen might be varied for a laboratory test. In summary, one should determine at this early phase if the test is reasonably robust to the circumstances under which it is performed or if it only operates well in particular settings. The reproducibility of results is an important factor to address in phase 1, and the test should be improved in this regard if necessary. Often, at the exploratory phase, subjects and samples that are conveniently available (e.g. from a clinic or blood bank) are studied. Rigorous evaluation of a test in a well-defined sense begins in the next phase.

Phase 2 is called the validation phase to contrast it with the exploratory (hypothesis generating) phase 1. Selection criteria for cases and controls, the tester (if applicable) and the protocol for performing the test are specified in rigorous detail in phase 2. This allows the results to be interpreted without ambiguity and as pertaining to a relevant well-defined population. Sampling of cases and controls was discussed in the early chapters of this book. The choice of tester (e.g. technician and radiologist in an imaging study) is also an important factor to consider. In phase 2, expert testers might be employed in contrast to later phases where testers might be population based. We discussed common sources of bias of which investigators should be aware in Chapter 1. Care must be taken to avoid these in phase 2, as in all phases of development. A key objective of phase 2 is to ascertain true and false positive fractions of the test in a particular setting. Thus, in order to design a phase 2 study, some minimally acceptable true and false positive fractions should be specified in advance. The study can then be designed with adequate sample sizes, so that conclusions can be drawn from it with regard to the test meeting these minimal operating characteristics or not.

A test that meets these criteria in phase 2 and appears promising for further development should undergo thorough evaluation in more comprehensive case-control studies before it proceeds to be applied as a practical testing tool in prospective studies at phase 4. We call phase 3 this intermediate phase. Often a primary objective of phase 3 is to determine criteria that should be used for defining screen positivity in phase 4. ROC curves can be employed to determine an operating point with desirable trade-offs between the TPF and the FPF, and the corresponding threshold can then be used as the positivity criterion in phase 4. Factors affecting test results from non-diseased subjects should be determined at this stage. If necessary, covariate-specific thresholds can then be defined. In addition, covariates that affect the TPFs or the ROC curve should be identified so that the populations or circumstances in which the test is performed can be optimized in phase 4. Tests are often compared in case-control studies. We consider such studies to be part of phase 3 also, because the purpose is usually to select tests that should undergo prospective phase 4 evaluation. In addition, algorithms for combining tests to define a useful composite for application in

phase 4 may be developed in phase 3.

Phases 1, 2 and 3 are typically retrospective case-control studies. Testing is therefore done only for research purposes (the disease status of the patient is already determined by other means). In contrast, the prospective cohort studies of phase 4 apply the test to subjects whose disease status is generally unknown at the time of testing. The results of the test frequently determine if further diagnostic work-up is even undertaken for a patient. Due to their prospective nature, the care of the patient enters into consideration when designing phase 4 studies. For example, definitive diagnostic testing may not be undertaken for subjects that screen negative in phase 4, or may only be undertaken for a subset of such patients, as discussed in Chapter 7. The objective of phase 4 is to determine the operating characteristics of the test when used as a diagnostic tool in a designated patient population. The extent and characteristics of disease detected with the test are determined. The false referral probability and characteristics of subjects that falsely screen positive must be determined. Promising tests are often compared in phase 4, because this is the phase where their practical performances as diagnostic tools are measured, and hence where the most relevant comparisons can be made.

Although a test accurately diagnoses disease, this does not necessarily mean that there is benefit to the patient. As delineated in Chapter 1 (see Table 1.1), effective treatment must be available for the disease detected with the test. Testing, work-up and treatment must be affordable and acceptable to patients, and so on. Ideally, the test is evaluated for its impact on the patient population before becoming part of routine healthcare. Its impact can be measured in terms of disease outcomes (mortality and quality of life) and cost. Such can be done through a randomized clinical trial, for example. Studies that evaluate the overall impact of the test on the population are called phase 5 studies.

Some general principles of study design were discussed in Chapter 1. These apply to studies at all phases of test development, but particularly starting at phase 2. Rigorous definitions of disease, clear protocols for applying the test, criteria for enrolling subjects and so forth must be undertaken with the same sort of care that is typically required for therapeutic studies. Sources of bias (see Section 1.2.5) must be minimized. Efforts should include blinding, for example. Comparative studies of test accuracy should be undertaken following the principles discussed in Section 3.1. Finally, in phase 5, randomized trials with mortality and cost as outcome measures are ideal and the reader is referred to the large body of literature on the design of randomized trials for study design principles (see Pocock, 1982, for example).

We focus in this chapter on the sample size calculations needed to ensure that conclusions can be drawn from a study. Sample size calculations for phases 2, 3 and 4 are considered in detail in Sections 8.2, 8.3 and 8.4, respectively. The following section describes matching and stratification as two additional issues that can be considered in study design.

8.2 Sample sizes for phase 2 studies

A phase 1 study is exploratory by definition. Its design is not based on a specific well-defined hypothesis. Rather, its purpose is to generate such a hypothesis for testing in phase 2. Thus formal sample size calculations are not considered here for phase 1, and we begin with phase 2. Multiple strategies are possible for sample size calculations. There are at least as many possibilities as there are for data analysis! We first describe a strategy that we find particularly straightforward and conceptually appealing for binary tests. It is extended to continuous tests in Section 8.2.2. Another strategy that has been proposed is described in Section 8.2.3.

8.2.1 *Retrospective validation of a binary test*

A phase 2 study is designed when a well-defined test and its target population are already identified. Assume that random samples of cases and controls will be drawn from the population and estimates of test accuracy will be made. In this section we assume that Y is binary, and therefore estimates of (FPF, TPF) will be made.

One needs to identify values for (FPF, TPF) that are minimally acceptable in order to design the study. Let $(\text{FPF}_0, \text{TPF}_0)$ denote such values. These are specified by the investigators and depend on the trade-off between false positives and true positives that are acceptable within the context of the test, disease, available resources and the population in which it is to be applied (Baker, 2000). Suppose that the goal of the phase 2 study is to determine if the test meets these minimal criteria.

Formally, the study will test the null hypothesis depicted in Fig. 8.1, namely

$$H_0 : \{\text{TPF} \leqslant \text{TPF}_0 \quad \text{or} \quad \text{FPF} \geqslant \text{FPF}_0\}. \tag{8.1}$$

From a study that rejects H_0 it will be concluded that $\text{TPF} > \text{TPF}_0$ and $\text{FPF} < \text{FPF}_0$, i.e. that the test meets minimal criteria.

The hypothesis can be tested by calculating a joint $1-\alpha$ confidence region for (FPF, TPF), as described in Section 2.2.2. When the null hypothesis is one-sided, a rectangular confidence region made up of the cross-product of two one-sided, $1 - \alpha^* = \sqrt{1-\alpha}$ confidence intervals is appropriate, as shown in Fig. 8.2. If the $1 - \alpha$ confidence region for (FPF, TPF) lies entirely within the region of acceptable values (unshaded region in Fig. 8.1), one can reject H_0 and make a positive conclusion about the test. We refer the reader back to Chapter 2 for a discussion of confidence interval construction.

The sample sizes for the phase 2 study, n_D and $n_{\bar{D}}$, should be chosen sufficiently large to ensure that a positive conclusion will be drawn with power $1 - \beta$ if the accuracy of the test is in fact at some specified, desirable levels. We denote these desirable classification probabilities by $(\text{FPF}_1, \text{TPF}_1)$. These reflect a test with levels of performance that the research community would want to undergo further development. In summary, we require that

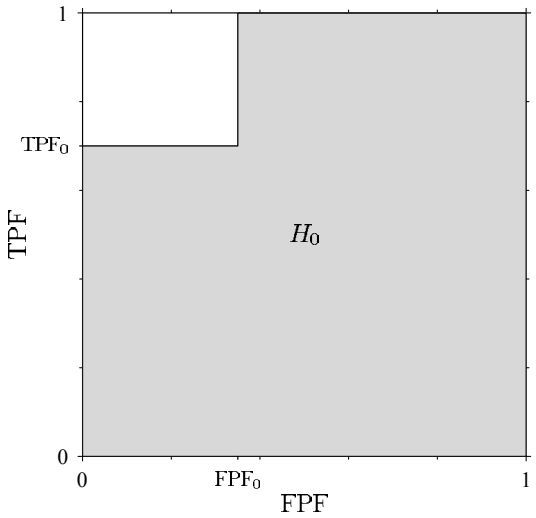

FIG. 8.1. Regions in the (FPF, TPF) space for a binary test that correspond to unacceptable tests (H_0, shaded region) and acceptable tests (unshaded region)

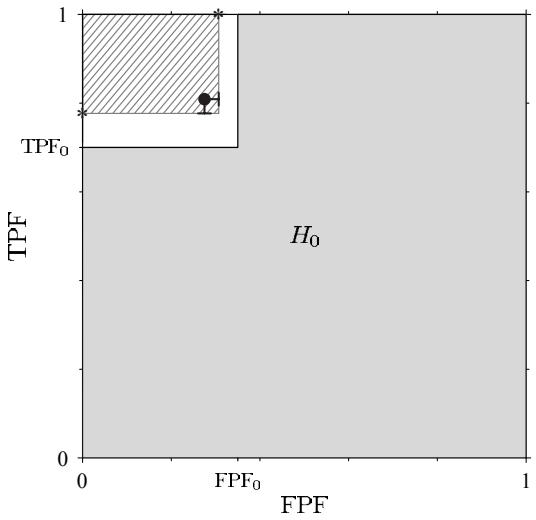

FIG. 8.2. A one-sided rectangular confidence region for (FPF, TPF) of the exercise stress test calculated from the CASS data. The classification probabilities meet the minimal criteria: FPF $\leqslant 0.35$ and TPF $\geqslant 0.70$. The points indicated with asterisks represent $\text{FPF}_U^{\alpha^*}$ and $\text{TPF}_L^{\alpha^*}$, the upper and lower $\alpha^* = 1 - \sqrt{1-\alpha}$ confidence limits for FPF and TPF, respectively

$$1-\beta = \mathrm{P}[\mathrm{TPF}_L^{\alpha^*} > \mathrm{TPF}_0 \quad \text{and} \quad \mathrm{FPF}_U^{\alpha^*} < \mathrm{FPF}_0 | \mathrm{FPF}_1, \mathrm{TPF}_1]$$
$$= \mathrm{P}[\mathrm{TPF}_L^{\alpha^*} > \mathrm{TPF}_0 | \mathrm{TPF}_1]\, \mathrm{P}[\mathrm{FPF}_U^{\alpha^*} < \mathrm{FPF}_0 | \mathrm{FPF}_1]\,,$$

where $\mathrm{TPF}_L^{\alpha^*}$ and $\mathrm{FPF}_U^{\alpha^*}$ denote the lower and upper one-sided limits of the confidence intervals for TPF and FPF, respectively. If n_D is chosen so that $\mathrm{P}[\mathrm{TPF}_L^{\alpha^*} > \mathrm{TPF}_0 | \mathrm{TPF}_1] = \sqrt{1-\beta}$ and $n_{\bar{D}}$ so that $\mathrm{P}[\mathrm{FPF}_U^{\alpha^*} < \mathrm{FPF}_0 | \mathrm{FPF}_1] = \sqrt{1-\beta}$, then this ensures adequate study power $1-\beta$ since the product is $1-\beta$. We define $\beta^* = 1 - \sqrt{1-\beta}$.

If the confidence limits are based on asymptotic normal distribution theory for the estimates, then sample sizes can be based on the asymptotic variance formulae. These yield the following sample size requirements:

$$n_D = \frac{\left(Z^{1-\alpha^*}\sqrt{\mathrm{TPF}_0(1-\mathrm{TPF}_0)} + Z^{1-\beta^*}\sqrt{\mathrm{TPF}_1(1-\mathrm{TPF}_1)}\right)^2}{(\mathrm{TPF}_1 - \mathrm{TPF}_0)^2} \tag{8.2}$$

and

$$n_{\bar{D}} = \frac{\left(Z^{1-\alpha^*}\sqrt{\mathrm{FPF}_0(1-\mathrm{FPF}_0)} + Z^{1-\beta^*}\sqrt{\mathrm{FPF}_1(1-\mathrm{FPF}_1)}\right)^2}{(\mathrm{FPF}_1 - \mathrm{FPF}_0)^2}\,, \tag{8.3}$$

where $Z^{1-\alpha^*} = \Phi^{-1}(1-\alpha^*)$ and $Z^{1-\beta^*} = \Phi^{-1}(1-\beta^*)$.

Since phase 2 studies tend to be small, confidence limits may be better calculated using exact methods. Sample sizes that yield adequate power for such analysis can be calculated with simulation studies. The above asymptotic theory-based formulae provide useful starting points for simulation studies, as illustrated next.

Example 8.1

It is hoped that a urinary test for chlamydia is 95% specific and 90% sensitive. It must be shown to be at least 80% specific and 75% sensitive in order to be considered for further evaluation. Thus $(\mathrm{FPF}_1, \mathrm{TPF}_1) = (0.05, 0.90)$ and $(\mathrm{FPF}_0, \mathrm{TPF}_0) = (0.20, 0.75)$. Conclusions will be based on a 90% rectangular confidence region using one-sided exact confidence limits.

If the study is to have 90% power, the formulae (8.2) and (8.3) based on asymptotic theory indicate that $n_D = 64$ and $n_{\bar{D}} = 46$. A set of 5000 simulation studies generating binary test data with classification probabilities $(\mathrm{FPF}_1, \mathrm{TPF}_1)$ show that these sample sizes yield 88% power. Raising the sample sizes to $n_D = 70$ and $n_{\bar{D}} = 50$ increases the power to 91%. Therefore, about 70 cases and 50 controls should be enrolled in the phase 2 validation study. ∎

8.2.2 *Retrospective validation of a continuous test*

When Y, the result of the test, is on a continuous scale, the question to answer is whether or not, for some threshold c, the dichotomized test, $I[Y > c]$, has

acceptable performance. That is, can the test operate at TPF and FPF values that reach or exceed the minimal criteria? In Fig. 8.3 the shaded region again corresponds to unacceptable test performance while performance parameters in the unshaded region are acceptable. If the ROC curve for Y passes through this unshaded region, then it is an acceptable test, because for some threshold it has acceptable (FPF, TPF) values. On the other hand, if the ROC curve lies entirely in the shaded region, then the test is unacceptable. For the ROC curves in Fig. 8.3, we see that test A meets the minimal criteria but test B does not.

Observe that the ROC curve for a test crosses the unshaded region if and only if $\mathrm{ROC}(\mathrm{FPF}_0) \geqslant \mathrm{TPF}_0$. An equivalent formulation is that $\mathrm{ROC}^{-1}(\mathrm{TPF}_0) \leqslant \mathrm{FPF}_0$. We write the null hypothesis as

$$H_0 : \mathrm{ROC}(\mathrm{FPF}_0) \leqslant \mathrm{TPF}_0 \,. \tag{8.4}$$

Under the null hypothesis, the ROC curve for Y lies fully in the unacceptable region of the (FPF, TPF) space. A hypothesis test can be based on the lower $(1-\alpha)$-level confidence limit for $\mathrm{ROC}(\mathrm{FPF}_0)$. If this lower limit exceeds TPF_0, then (8.4) is rejected and we conclude that the test meets the minimal criterion for further development. In Fig. 8.4 we show the 95% lower confidence limit for the ROC(0.2) of the CA-19-9 marker for pancreatic cancer based on data from Wieand *et al.* (1989). If the minimally acceptable levels for the (FPF, TPF) of a pancreatic cancer biomarker were $(0.2, 0.6)$, say, we would conclude that CA-19-9 meets these criteria. That is, we reject H_0: $\mathrm{ROC}(0.2) \leqslant 0.6$.

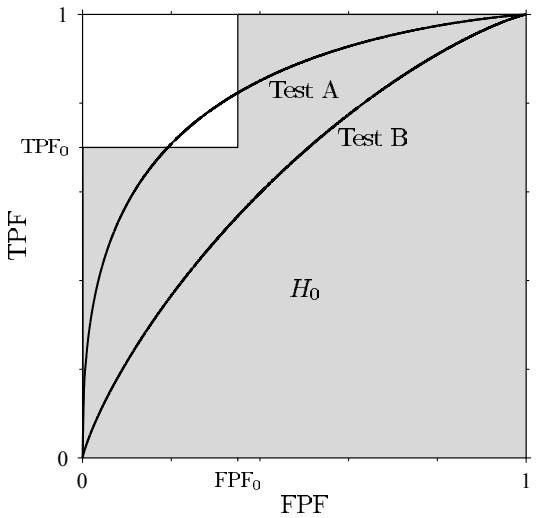

FIG. 8.3. ROC curves for two (hypothetical) tests. The upper one meets the minimally acceptable criterion that it can attain operating points which exceed $(\mathrm{FPF}_0, \mathrm{TPF}_0)$, whereas the lower one does not

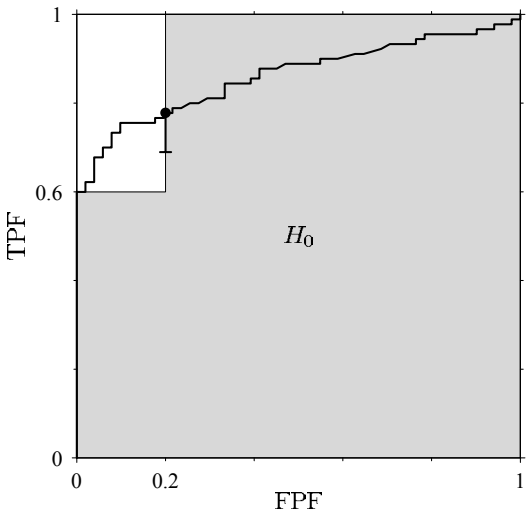

FIG. 8.4. A test of the null hypothesis that, at the threshold corresponding to $\mathrm{FPF}_0 = 0.2$, the TPF does not exceed 0.6 for the CA-19-9 marker of pancreatic cancer. Shown is the lower 95% confidence limit for ROC(0.2) using data from Wieand *et al.* (1989)

Turning now to sample size calculations, we need to consider how the lower confidence limit for $\mathrm{ROC}(\mathrm{FPF}_0)$ will be constructed. A confidence limit based on asymptotic distribution theory derived in Chapter 5 is

$$\mathrm{ROC}(\mathrm{FPF}_0)_L^\alpha = \hat{\mathrm{ROC}}(\mathrm{FPF}_0) - \Phi^{-1}(1-\alpha)\sqrt{\mathrm{var}\{\hat{\mathrm{ROC}}(\mathrm{FPF}_0)\}}\,. \tag{8.5}$$

In practice, we find that the confidence limits based on the logit transform, $\mathrm{logit}\,\hat{\mathrm{ROC}}(\mathrm{FPF}_0)$, that were described in Section 5.2.3, have better coverage in small samples and we implement analyses with these limits. However, asymptotic theory-based sample size calculations are similar for both the untransformed and transformed approaches. Thus, for simplicity, we proceed here with the expression (8.5) for the untransformed lower limit. A positive conclusion is drawn from the study if we find that $\mathrm{ROC}(\mathrm{FPF}_0)_L^\alpha > \mathrm{TPF}_0$.

Suppose that the diagnostic test, in fact, has a TPF value of TPF_1 when the threshold corresponding to FPF_0 is used. That is, suppose that $\mathrm{ROC}(\mathrm{FPF}_0) = \mathrm{TPF}_1$. Then, the power of the study to draw a positive conclusion is

$$\mathrm{P}\left[\hat{\mathrm{ROC}}(\mathrm{FPF}_0) - \Phi^{-1}(1-\alpha)\sqrt{\mathrm{var}\{\hat{\mathrm{ROC}}(\mathrm{FPF}_0)\}} > \mathrm{TPF}_0\right], \tag{8.6}$$

where the probability is calculated assuming that $\mathrm{TPF}_1 = \mathrm{ROC}(\mathrm{FPF}_0)$.

With the power (8.6) specified at some desired level $1-\beta$, this implies that the sample sizes n_D and $n_{\bar{D}}$ should be chosen to satisfy

$$\frac{V_1}{n_D}\left(\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)\right)^2=(\mathrm{TPF}_0-\mathrm{TPF}_1)^2\,,$$

where V_1/n_D is the asymptotic theory-based expression for $\mathrm{var}\{\hat{\mathrm{ROC}}(\mathrm{FPF}_0)\}$ calculated under the alternative hypothesis $\mathrm{ROC}(\mathrm{FPF}_0)=\mathrm{TPF}_1$. Using the empirical ROC curve estimator, $\hat{\mathrm{ROC}}_e(\mathrm{FPF}_0)$, and the analytic form for its asymptotic variance derived in Result 5.1, we write

$$V_1=\mathrm{TPF}_1(1-\mathrm{TPF}_1)+\kappa r_1^2\mathrm{FPF}_0(1-\mathrm{FPF}_0)\,, \tag{8.7}$$

where r_1 denotes the slope of the ROC curve at FPF_0 and κ denotes the ratio of cases to controls, $n_D/n_{\bar{D}}$. The next result summarizes the discussion.

Result 8.1

If study conclusions are based on $\mathrm{ROC}(\mathrm{FPF}_0)_L^\alpha$ exceeding TPF_0, where $\mathrm{ROC}(\mathrm{FPF}_0)_L^\alpha$ is the lower $1-\alpha$ confidence interval calculated with the empirical ROC curve, then in order to achieve power $1-\beta$ when $\mathrm{TPF}_1=\mathrm{ROC}(\mathrm{FPF}_0)$ we require that

$$n_D=\frac{\left\{\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)\right\}^2}{(\mathrm{TPF}_0-\mathrm{TPF}_1)^2}V_1\,, \tag{8.8}$$

where V_1 is defined in (8.7). ∎

Example 8.2

The largest acceptable false positive fraction for a test based on a new biomarker is $\mathrm{FPF}_0=0.10$. It is anticipated that it will be 95% sensitive ($\mathrm{TPF}_1=0.95$) at the threshold corresponding to $\mathrm{FPF}=0.10$, but it must be shown to be at least 75% sensitive there ($\mathrm{TPF}_0=0.75$) in order to proceed with further development. The study will enrol equal numbers of cases and controls ($\kappa=1$) and we choose $\alpha=0.05$ and $\beta=0.10$. All components of the sample size formula (8.8) are now defined except for r_1, the slope of the ROC curve at FPF_0. We have

$$n_D=\frac{(1.64+1.28)^2\left\{(0.95)(0.05)+(r_1)^2(0.10)(0.90)\right\}}{(0.95-0.75)^2}\,.$$

Suppose that the biomarker is anticipated to have a binormal ROC curve, $\mathrm{ROC}(t)=\Phi(a+b\Phi^{-1}(t))$. Then by the chain rule we have

$$r_1=\frac{\delta}{\delta t}\mathrm{ROC}(t)=b\frac{\phi(a+b\Phi^{-1}(t))}{\phi(\Phi^{-1}(t))}\,.$$

Suppose further that we anticipate the slope parameter $b=1$ under both the null and alternative hypotheses. Under these assumptions we can determine the values of a that correspond to each of the hypotheses. We do this by noting that $\mathrm{ROC}(\mathrm{FPF}_0)=\Phi(a+\Phi^{-1}(\mathrm{FPF}_0))=0.75$ under the null, which implies that

$a = 1.96$. Similarly, under the alternative $\mathrm{ROC}(\mathrm{FPF}_0) = 0.95$, which implies that $a = 2.93$. Substituting $b = 1$ and these values for a into the expression for r_1 yields corresponding values of 1.81 and 0.58 for r_1. To be conservative, we use the larger value, having found from experience that this generally provides a better sample size calculation. Substituting $r_1 = 1.81$ into the expression for n_D we find $n_D = 73$. The choice $\kappa = 1$ implies that $n_{\bar{D}} = 73$ also.

This sample size calculation is based on asymptotic distribution theory. Simulation studies are used to assess the adequacy of the calculations. In particular, we generate data under H_1 and calculate the empirical power as the proportion of simulated studies in which a positive conclusion is drawn. Data for $n_D = 73$ cases and $n_{\bar{D}} = 73$ controls are generated from the binormal model with ROC intercept $a = 2.93$ and slope $b = 1$, as we assumed in the sample size calculations above. Normal distributions are used, with $Y_{\bar{D}} \sim \mathrm{N}(0, 1)$ and $Y_D \sim \mathrm{N}(a/b, 1/b^2)$. Recall that, since the analysis uses only the ranks of the data, the Gaussian distributional forms used in the simulations are irrelevant. The study power, calculated as 89%, appears to be adequate.

Additional simulations are performed under the null hypothesis, simply to check if inference using the confidence limits is valid with sample sizes of 73. That is, we wish to confirm that the size of the test procedure using the 95% lower confidence limit is the nominal 5%. Data are generated from the binormal model under the null. Hence we choose $a = 1.96$ and $b = 1.0$ so that $\mathrm{ROC}(\mathrm{FPF}_0) = \mathrm{TPF}_0$. The null hypothesis is rejected in 6% of simulations, close enough to the nominal level. We reiterate that these confidence limits, based on the logit transform, are better behaved in small samples than are those based on the untransformed estimate of $\mathrm{ROC}(\mathrm{FPF}_0)$. The null is rejected in 10% of simulations that used the latter confidence limits, which is unacceptably large compared with the nominal 5% level. ∎

8.2.3 *Sample size based on the AUC*

A non-binary diagnostic test can also be evaluated by comparing its AUC or other ROC summary index with a value that is considered to be minimally acceptable. Specifying a minimally acceptable AUC may be more difficult, in my opinion, than specifying a minimally acceptable (FPF, TPF) combination. However, the strategy does have the advantage of incorporating information across multiple operating points of the test, rather than being limited to only one point, as in the previous subsection.

Suppose that we estimate the AUC either with nonparametric or other methods and compare the estimate with the minimally acceptable value, which we denote by AUC_0. To test

$$H_0 : \mathrm{AUC} \leqslant \mathrm{AUC}_0$$

the $1-\alpha$ lower confidence limit is calculated and H_0 is rejected if it exceeds AUC_0. In practice, we use confidence limits based on logit $\hat{\mathrm{AUC}}$, but again, to simplify the sample size calculations, we suppose that the lower limit of the confidence

interval is based on asymptotic normality of the untransformed estimate, $\hat{\text{AUC}}$. That is, we reject H_0 if

$$\hat{\text{AUC}} - \Phi^{-1}(1-\alpha)\sqrt{\hat{\text{var}}(\hat{\text{AUC}})} > \text{AUC}_0\,,$$

where $\hat{\text{var}}(\hat{\text{AUC}})$ denotes the estimated variance.

To calculate sample sizes using standard calculations so that a power of $1-\beta$ is achieved when the AUC is at some desirable value, AUC_1, the following equation must be satisfied:

$$\frac{\text{var}(\hat{\text{AUC}})\left(\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)\right)^2}{(\text{AUC}_1-\text{AUC}_0)^2} = 1\,. \tag{8.9}$$

Once α, β, AUC_0 and AUC_1 are specified, the remaining task is to postulate a value for the variance. This is not a simple task because the variance depends, not only on the sample sizes and postulated AUCs, but on the underlying probability distributions. The next result shows that for the empirical AUC estimate, $\hat{\text{AUC}}_e$, the asymptotic variance expression can be written in terms of the ROC curve. Therefore, if one postulates an ROC curve under the null and alternative, the asymptotic variance can be calculated and so approximate sample sizes can be derived.

Result 8.2

In large samples,

$$\text{var}(\hat{\text{AUC}}_e) = \text{var}_D\,/n_{\bar{D}} + \text{var}_{\bar{D}}\,/n_D\,,$$

where

$$\begin{aligned}\text{var}_D &= \int_0^1 (\text{ROC}(t))^2\,\mathrm{d}t - \text{AUC}^2\,,\\ \text{var}_{\bar{D}} &= \int_0^1 (\text{ROC}^{-1}(t))^2\,\mathrm{d}t - (1-\text{AUC})^2\,.\end{aligned}$$

Proof Consider the expression for $\text{var}(\hat{\text{AUC}}_e)$ given in Result 5.5:

$$\text{var}(\hat{\text{AUC}}_e) = \frac{\text{var}\left(S_{\bar{D}}(Y_D)\right)}{n_D} + \frac{\text{var}\left(S_D(Y_{\bar{D}})\right)}{n_{\bar{D}}}\,.$$

Observe that

$$\begin{aligned}\text{var}\left(S_D(Y_{\bar{D}})\right) &= \text{E}\left\{S_D(Y_{\bar{D}})\right\}^2 - \left[\text{E}\left\{S_D(Y_{\bar{D}})\right\}\right]^2\\ &= -\int_{-\infty}^{\infty} S_D^2(y)\,\mathrm{d}S_{\bar{D}}(y) - \text{AUC}^2\\ &= \int_0^1 S_D^2\left(S_{\bar{D}}^{-1}(t)\right)\mathrm{d}t - \text{AUC}^2\\ &= \int_0^1 \left(\text{ROC}(t)\right)^2\,\mathrm{d}t - \text{AUC}^2\,.\end{aligned}$$

Similarly

$$\begin{aligned}\text{var}\left(S_{\bar{D}}(Y_D)\right) &= \text{E}\left\{S_{\bar{D}}(Y_D)\right\}^2 - \left[\text{E}\left\{S_{\bar{D}}(Y_D)\right\}\right]^2 \\ &= \int_0^1 \left\{\text{ROC}^{-1}(t)\right\}^2 \,\text{d}t - \left\{\int_0^1 \text{ROC}^{-1}(t)\,\text{d}t\right\}^2\end{aligned}$$

and the second term is $(1 - \text{AUC})^2$. ∎

As a consequence of (8.9) and Result 8.2, we have the following expression for sample sizes:

$$n_D = (\kappa\,\text{var}_D + \text{var}_{\bar{D}}) \left\{\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}{\text{AUC}_1 - \text{AUC}_0}\right\}^2, \tag{8.10}$$

where $\kappa = n_D/n_{\bar{D}}$. Approximations to $\text{var}(\hat{\text{AUC}}_e)$ have been reported previously. Hanley and McNeil (1982) assume exponential probability distributions for Y_D and $Y_{\bar{D}}$ to derive an expression. Obuchowski (1994) found that the Hanley and McNeil approximation performed poorly for binormal ROC curves and proposed an alternative expression suitable for the binormal setting. Our Result 8.2 is much more general than either of theirs.

Given an assumed ROC curve, one can calculate $\text{var}(\hat{\text{AUC}}_e)$ in at least two ways. One can use the expressions in Result 8.2 directly and numerically integrate the squared terms. Another tactic is to simulate large amounts of data from the ROC model, transform the raw data to placement values and calculate the variance terms $\text{var}(S_{\bar{D}}(Y_D))$ and $\text{var}(S_D(Y_{\bar{D}}))$ empirically. We take the latter approach in the next example.

Example 8.3

Suppose that the standard biomarker has a binormal ROC curve with $a = 0.545$ and $b = 1.0$. That is, its AUC is $\Phi\left(a/\sqrt{1+b^2}\right) = 0.65$. A new biomarker will be considered for further development if its AUC is shown to be greater than this value. Thus, the minimally acceptable AUC is $\text{AUC}_0 = 0.65$. The new biomarker is anticipated to have a binormal ROC curve with $a = 1.19$ and $b = 1.0$. Thus, we identify AUC_1 as 0.80.

Let us calculate the variance components in Result 8.2, assuming that the ROC curve for the biomarker is what we anticipate it will be: $\text{ROC}(t) = \Phi(1.19 + \Phi^{-1}(t))$. We generated 10 000 disease and non-disease observations from the binormal ROC curve, with $Y_{\bar{D}} \sim \text{N}(0,1)$ and $Y_D \sim \text{N}(1.19, 1)$. Recall again that, because the ROC curve is a function of only the ranks of the data, the actual distributional forms chosen are irrelevant. All that matters is that the ROC curve is of the stipulated form. The placement values calculated have variances

$$\text{var}(S_{\bar{D}}(Y_D)) = 0.048 \quad \text{and} \quad \text{var}(S_D(Y_{\bar{D}})) = 0.045\,.$$

Therefore, if the biomarker is as good as we hope, we will have

$$\text{var}(\hat{\text{AUC}}_e) = \frac{0.048}{n_D} + \frac{0.046}{n_{\bar{D}}}.$$

Suppose that equal numbers of cases and controls will be enrolled to the study. Substituting into eqn (8.8), with $\alpha = 0.05$ and $\beta = 0.10$, yields

$$n_D = \frac{(1.64 + 1.28)^2}{(0.80 - 0.65)^2}(0.048 + 0.046) = 36.$$

Asymptotic theory therefore suggests that 36 cases and 36 controls be enrolled, in order to be 90% sure that the lower 95% confidence limit for the AUC based on the empirical estimator will exceed AUC_0.

The calculations are based on asymptotic theory that may not hold exactly in small samples. They provide starting points for simulation studies that fine tune the sample size calculations. Data are generated from the binormal model as before, but now using sample sizes $n_D = n_{\bar{D}} = 36$ and repeated 500 times. The lower 95% confidence limit for the AUC exceeds 0.65 with a rate of only 81%. This power is not as large as desired. We therefore increase the sample sizes to $n_D = n_{\bar{D}} = 50$, repeat the simulation study and find that the power is 90%. Thus sample sizes of $n_D = n_{\bar{D}} = 50$ are recommended. Finally, data are generated under the null hypothesis and confirm that the rejection rate is adequate for the study. It is 5.4%, a value that is close enough to the nominal level. ∎

Our development of sample size calculations is based on the nonparametric estimate of the AUC. If a parametric estimator of the AUC is to be used for inference, then sample size calculations could acknowledge the smaller variance that such an estimator is likely to have relative to that of $\hat{\text{AUC}}_e$. Obuchowski (1994) provides a variance formula for the fully parametric AUC estimator that assumes normal distributions for test results, namely $Y_D \sim \text{N}(\mu_D, \sigma_D^2)$ and $Y_{\bar{D}} \sim \text{N}(\mu_{\bar{D}}, \sigma_{\bar{D}}^2)$. As described in Example 5.3, this AUC estimator has the form

$$\hat{\text{AUC}}_N = \Phi\left(\frac{\hat{\mu}_D - \hat{\mu}_{\bar{D}}}{\sqrt{\hat{\sigma}_D^2 + \hat{\sigma}_{\bar{D}}^2}}\right),$$

where parameters are estimated with sample means and variances. Obuchowski's expression for $\text{var}(\hat{\text{AUC}}_N)$ relies on the asymptotic joint normal distribution for $\{\hat{\mu}_D, \hat{\sigma}_D^2, \hat{\mu}_{\bar{D}}, \hat{\sigma}_{\bar{D}}^2\}$ and the delta method. Variance formulae for parametric distribution-free estimates of the AUC (see Section 5.5) have not been derived and may be complicated. We suggest that $\hat{\text{AUC}}_e$ be used for sample size calculations, even if a fully parametric or a parametric distribution-free estimator will be used for analysis. The rationale is based on the observation that $\hat{\text{AUC}}_e$ is very efficient, at least under the normal theory model (Dodd, 2001). Therefore, calculated sample sizes will not be that much larger than required for those AUC estimators that make parametric assumptions.

Summary measures other than the AUC can be used to quantify the accuracy of a test and may also provide the basis for power calculations. The principles for power calculations are the same. However, since we often do not have explicit analytic expressions for variances, the calculations must be done entirely using simulation studies. For instance, when using nonparametric estimates of the partial AUC, a simulation-based sample size calculation seems to be the only currently available option. Obuchowski and McClish (1997) and Obuchowski (1998) present analytic expressions for variances of fully parametric estimates of the partial AUC from which sample sizes can then be calculated. These might provide useful starting points for simulation-based calculations of sample sizes needed for studies that will use semiparametric or nonparametric inference about the partial AUC.

Use of an ROC summary index such as the AUC or pAUC is likely to be more efficient than the use of a single ROC point for inference. Therefore the strategy of the previous section is likely to require larger sample sizes than the calculations in the current section. To illustrate, consider the assumed binormal curves of Example 8.3. The ROC point at $\text{FPF}_0 = 0.1$ corresponds to TPFs of $\text{TPF}_0 = 0.23$ under H_0 and $\text{TPF}_1 = 0.46$ under the alternative. The asymptotic theory-based calculations for testing $\text{ROC}(\text{FPF}_0) = \text{TPF}_0$ with 90% power at $\alpha = 0.05$ yield $n_D = n_{\bar{D}} = 115$. These are substantially larger than the AUC-based sample sizes calculated in Example 8.3.

The main disadvantage of basing inference on the AUC or pAUC is that these measures are less clinically relevant than the measure considered in the previous section, namely the TPF corresponding to the minimally acceptable FPF. In my opinion, specification of a practically meaningful improvement in the AUC (i.e. AUC_1) is likely to be more difficult than specification of an improvement in the TPF (i.e. TPF_1) corresponding to the minimally acceptable false positive fraction, $\text{ROC}(\text{FPF}_0)$. However, others may disagree with me on this point. Perhaps experience with the AUC index in a specific context may lead one to an intuition for the magnitudes of improvement that correspond to clinically meaningful improvements in test performance in that context.

8.2.4 *Ordinal tests*

Sample size calculations geared specifically towards ordinal tests have not received sufficient attention in the literature. Obuchowski has been the main contributor to this area. She employs ROC summary indices as the basis for inference. Her calculations, however, use variances that apply to continuous data and are based on the assumption that test results have normal distributions. She implicitly assumes that these apply to ROC summary indices estimated from ordinal data with the Dorfman and Alf (1968) binormal distribution-free method. Simulation studies (Obuchowski and McClish, 1997) suggest that these variances for continuous data underestimate the actual variances of summary indices calculated from ordinal study data. Nevertheless, the continuous data sample size formulae may provide reasonable starting points for simulation-based calculations

of sample size for ordinal tests. Obuchowski and McClish refer to the simulation program ROCPWR, which is part of the ROC software developed primarily for ordinal data by Metz and colleagues at the University of Chicago.

Another approach is to dichotomize the ordinal test for the purposes of sample size calculations. The methods described in Section 8.2.1 can then be employed. This requires one to specify an appropriate category as the threshold for defining the binary test. Further work to allow some flexibility in this regard would be worthwhile.

8.3 Sample sizes for phase 3 studies

The types of objectives for a phase 3 study are more varied than they are for phase 2. We consider here sample size calculations for studies that address three different types of objectives. First, there are studies that seek to compare two different tests. We assume that a paired case-control study design is employed. The two different tests may actually be the same test, but done at different time points or under different circumstances on the same subject. The key is that within-subject comparisons are to be made. Next, case-control studies that seek to compare tests in different subpopulations will be considered in order to determine if subject characteristics affect test performance. The key statistical aspect here is that comparisons are made between subjects, not within subjects. The sample size calculations would also apply to the comparison of two tests in an unpaired design. Lastly, we consider estimation of the threshold value corresponding to a pre-specified false positive fraction, one important component of the effort in phase 3 to define a screen positive criterion that can be employed in phase 4.

8.3.1 *Comparing two binary tests—paired data*

In this book we have emphasized the multiplicative scale for quantifying the relative performance of two tests. Thus, for the two tests under consideration, test A and test B, we base inference on $\mathrm{rTPF}(A,B) = \mathrm{TPF}_A/\mathrm{TPF}_B$ and $\mathrm{rFPF}(A,B) = \mathrm{FPF}_A/\mathrm{FPF}_B$. Other scales can be used. In particular, absolute differences, $(\mathrm{FPF}(A) - \mathrm{FPF}(B), \mathrm{TPF}(A) - \mathrm{TPF}(B))$, have been used (Obuchowski, 1998; Obuchowski and Zhou, 2002). Their large sample theory calculations yield sample sizes that are similar to ours because the test procedures are asymptotically equivalent under the null hypothesis.

Some comparative studies seek to determine if one test is superior to the other. However, in some instances a test might be preferable to another even if its accuracy parameters are not superior. For example, if an existing test is costly or invasive, then a new inexpensive noninvasive test may be preferable to the existing test as long as its accuracy is not substantially less. In this case the scientific objective is to determine if the accuracy of the new test is substantially inferior to the standard or not.

Therapeutic clinical trials are often classified as superiority studies or as non-inferiority (equivalence) studies (ICH, 1999). The same idea is pertinent to