Biostatistics (2009), **10**, 1, *pp.* 94–105
doi:10.1093/biostatistics/kxn018
Advance Access publication on June 12, 2008

Sample size for positive and negative predictive value in diagnostic research using case–control designs

DAVID M. STEINBERG*

Department of Statistics and Operations Research, Tel Aviv University,
Tel Aviv 69978, Israel
dms@post.tau.ac.il

JASON FINE, RICK CHAPPELL

Department of Biostatistics, University of Wisconsin, Madison, WI 53706, USA

SUMMARY

Important properties of diagnostic methods are their sensitivity, specificity, and positive and negative predictive values (PPV and NPV). These methods are typically assessed via case–control samples, which include one cohort of cases known to have the disease and a second control cohort of disease-free subjects. Such studies give direct estimates of sensitivity and specificity but only indirect estimates of PPV and NPV, which also depend on the disease prevalence in the tested population. The motivating example arises in assay testing, where usage is contemplated in populations with known prevalences. Further instances include biomarker development, where subjects are selected from a population with known prevalence and assessment of PPV and NPV is crucial, and the assessment of diagnostic imaging procedures for rare diseases, where case–control studies may be the only feasible designs. We develop formulas for optimal allocation of the sample between the case and control cohorts and for computing sample size when the goal of the study is to prove that the test procedure exceeds pre-stated bounds for PPV and/or NPV. Surprisingly, the optimal sampling schemes for many purposes are highly unbalanced, even when information is desired on both PPV and NPV.

Keywords: Biomarkers; Case–control study; Diagnostic testing; Optimal allocation; Sensitivity; Specificity.

1. INTRODUCTION

Diagnostic test procedures are often characterized by their sensitivity—the probability that a diseased individual gives a positive test result—and their specificity—the probability that a healthy individual gives a negative test result. For both clinician and patient, though, it is also important to consider the positive and negative predictive values (PPV and NPV, respectively) of the test procedure. The PPV is the probability that a subject is diseased, given a positive test result. The NPV is the probability that a subject is healthy, given a negative test result.

Our concern here is study design and in particular sample allocation and sample size determination, when the goal of a diagnostic study is to reach conclusions about PPV and/or NPV. The following

*To whom correspondence should be addressed.

application helped to motivate the research. A biotechnology company wished to prove the efficacy of a diagnostic kit for a certain autoimmune disease. The company's market research indicated that the kit would not be economically feasible unless they could clearly show that the NPV was at least 98%, given an assumed disease prevalence of $1/16$. The research and development team for the project wanted to know how to design a study to achieve this goal.

Such designs are playing an increasingly important role, not only in assay testing but also in a wide range of applications where PPV and NPV are crucial to evaluating the practical utility of a testing procedure. This is particularly true at the later stages of development, where the clinical performance of a test in particular populations is potentially of greater interest than the test's underlying sensitivity and specificity. These measures may be better suited to decisions regarding whether a particular test should be employed in a particular population, as opposed to the interpretation of test results on individual patients. Besides assays, such issues are critical in biomarker development for disease screening, prognosis, and risk assessment (Pepe *and others*, 2001; Baker *and others*, 2006) and in the development of imaging technology for rare diseases. Often a case–control study with subjects drawn from a population with known prevalence is the only feasible design. To our knowledge, there are no published methods providing guidance on choosing numbers of cases and controls in such studies.

The PPV and NPV of a diagnostic procedure are functions of the sensitivity, the specificity, and the disease prevalence. If we denote these quantities by sp, se, and ω, respectively, then

$$\text{PPV} = \frac{\omega \cdot se}{\omega \cdot se + (1-\omega)\cdot(1-sp)} \tag{1.1}$$

and

$$\text{NPV} = \frac{(1-\omega)\cdot sp}{\omega\cdot(1-se) + (1-\omega)\cdot sp}. \tag{1.2}$$

While the quantities (1.1) and (1.2) are well-defined conditional probabilities, they may not be directly estimable, depending on the study design.

In a cross-sectional study of the procedure on the relevant test population, both PPV and NPV can be estimated directly from the study results and standard methods for setting sample size in studying proportions can be applied. This problem is treated by Pepe (2003, Section 8.4.2) and Zhou *and others* (2002, Chapter 6). Wang and others (2006) and Moskowitz and Pepe (2006) consider the related problem of designing a paired cross-sectional study to compare 2 different methods with respect to PPV or NPV.

Cross-sectional studies will be problematic whenever disease prevalence is low, as only a small number of diseased individuals will be enrolled. Case–control studies will then be more efficient. Case–control studies provide direct estimates of the sensitivity and the specificity, but the relative proportion of cases and controls will not reflect the prevalence of the disease, so the standard methods for studying proportions will no longer be applicable. Instead, inference for PPV and NPV will require indirect methods that combine the estimated sensitivity and specificity with the prevalence via (1.1) and (1.2). Moreover, even the direct estimates from a cross-sectional study may require modification using (1.1) and (1.2) if the procedure is to be applied to a population with a different disease prevalence.

We present inferential procedures for PPV and NPV in Section 2, following the work of Li *and others* (2007). We then derive formulas for allocating the sample between the diseased and disease-free subjects and for sample size in a case–control study whose goal is to establish minimal bounds on PPV and/or NPV.

2. Inference for PPV and NPV

We consider a case–control study with n_1 diseased patients and n_2 disease-free control patients ($n = n_1 + n_2$). We denote by s_1 and s_2 the number of cases and controls, respectively, with positive test results. Then, the standard estimators of sensitivity and specificity are $\widehat{se} = s_1/n_1$ and $\widehat{sp} = (n_2 - s_2)/n_2$, respectively. Plugging these estimators into (1.1) and (1.2), for known prevalence, gives consistent estimators of the PPV and NPV:

$$\widehat{\mathrm{PPV}} = \frac{\omega \cdot \widehat{se}}{\omega \cdot \widehat{se} + (1-\omega)\cdot(1-\widehat{sp})} \tag{2.1}$$

and

$$\widehat{\mathrm{NPV}} = \frac{(1-\omega)\cdot \widehat{sp}}{\omega \cdot (1-\widehat{se}) + (1-\omega)\cdot \widehat{sp}}. \tag{2.2}$$

We can simplify the statistical inference by rewriting PPV and NPV in the form

$$\mathrm{PPV} = \frac{1}{1 + \frac{(1-\omega)\cdot(1-sp)}{\omega\cdot se}} = \frac{1}{1+\exp(\phi_1)\frac{1-\omega}{\omega}} \tag{2.3}$$

and

$$\mathrm{NPV} = \frac{1}{1 + \frac{\omega\cdot(1-se)}{(1-\omega)\cdot sp}} = \frac{1}{1+\exp(\phi_2)\frac{\omega}{1-\omega}}. \tag{2.4}$$

We then base statistical inference for PPV and NPV on the two log-likelihood ratios:

$$\phi_1 = \log(1-sp) - \log(se) \tag{2.5}$$

and

$$\phi_2 = \log(1-se) - \log(sp). \tag{2.6}$$

The parameters ϕ_1 and ϕ_2 are often used in clinical epidemiology and statistics. Formulas for inference about ϕ_1 and ϕ_2 are readily available (e.g. Pepe, 2003, Chapters 2 and 3). We repeat these results here, as they will be essential to our subsequent work. In one of the earliest papers to study ϕ_1 and ϕ_2, Simel *and others* (1991) also considered questions of sample size determination when the study goals are phrased directly in terms of the log-likelihood ratios. Boyko (1994) and Dujardin *and others* (1994) provide further discussion of the likelihood ratios and contrast these measures with classification probabilities and predictive values.

We estimate ϕ_1 and ϕ_2 by plugging in the estimators $\widehat{se}$ and $\widehat{sp}$ of the sensitivity and specificity and use the delta method to derive confidence intervals. Denote by P and $Q = 1 - P$ the fraction of diseased and disease-free subjects in the study, and assume that these will be the limiting fractions as the total sample size $n = n_1 + n_2 \to \infty$. Then,

$$\sqrt{n}\begin{pmatrix}\sqrt{P}(\widehat{se} - se)\\ \sqrt{Q}(\widehat{sp} - sp)\end{pmatrix} \to N(0, V), \tag{2.7}$$

where V is a diagonal matrix with entries $se(1-se)$ and $sp(1-sp)$. Application of the delta method then gives

$$\sqrt{n}(\hat{\phi}_i - \phi_i) \to N(0, \sigma_i^2), \tag{2.8}$$

where

$$\sigma_1^2 = \frac{1-se}{se\cdot P} + \frac{sp}{(1-sp)\cdot Q} \tag{2.9}$$

and

$$\sigma_2^2 = \frac{se}{(1-se)\cdot P} + \frac{1-sp}{sp\cdot Q}. \tag{2.10}$$

Replacing se and sp in (2.9) and (2.10) by their sample estimators provides consistent estimators $\hat{\sigma_i^2}$ of the variances σ_i^2.

Asymptotic confidence intervals for PPV and NPV can now be computed from associated confidence intervals for ϕ_1 and ϕ_2. For example, a symmetric $1-\alpha$ confidence interval for ϕ_1 will extend from $L(\hat{\phi_1}) = \hat{\phi_1} - Z_{1-\alpha/2}\hat{\sigma_1}/\sqrt{n}$ to $U(\hat{\phi_1}) = \hat{\phi_1} + Z_{1-\alpha/2}\hat{\sigma_1}/\sqrt{n}$. The matching $1-\alpha$ confidence interval for PPV extends from $1/\{1+\exp(U(\hat{\phi_1}))(1-\omega)/\omega\}$ to $1/\{1+\exp(L(\hat{\phi_1}))(1-\omega)/\omega\}$. These confidence bounds hold simultaneously for all possible values of the prevalence. Often only a lower confidence bound will be required. The lower confidence bound for PPV is found from the one-sided upper confidence bound for ϕ_1.

An alternative approach is to apply Fieller's Theorem to the ratios $(1-sp)/se$ and $(1-se)/sp$. See Li *and others* (2007) for details. When the denominators are far from 0, as they are here, Cox (1990) showed that the 2 methods are asymptotically equivalent.

A simultaneous confidence region for PPV and NPV can be computed from a corresponding region for ϕ_1 and ϕ_2; see Pepe (2003, Section 2.2.4).

3. Sample size for PPV or NPV only

We now address the question of choosing sample sizes. In this section, we consider studies in which only PPV is of interest; analogous results hold for NPV. In Section 4, we extend the results to studies in which inference is required about both PPV and NPV. Throughout we assume that the prevalence ω is known.

3.1 *Sample allocation*

First, we derive the allocation that provides the most precise inference, for a given total sample size, about ϕ_1 and hence about PPV. The optimal choice of P can be found by minimizing the asymptotic variance for $\hat{\phi_1}$, which is proportional to

$$\frac{1-se}{se \cdot P} + \frac{sp}{(1-sp)\cdot(1-P)}. \tag{3.1}$$

Differentiating with respect to P and equating to 0 lead to the condition

$$\frac{1-se}{se \cdot P^2} = \frac{sp}{(1-sp)(1-P)^2}. \tag{3.2}$$

Thus, the optimal ratio of diseased to disease-free subjects for estimating PPV is given by

$$P_{\text{PPV}}/(1-P_{\text{PPV}}) = \sqrt{\frac{(1-se)(1-sp)}{se \cdot sp}}, \tag{3.3}$$

where P_{PPV} is the optimal fraction of diseased patients in the study. For any reasonable diagnostic test, $se + sp > 1$, which implies that the right-hand side of (3.3) is less than 1. For example, if both sensitivity and specificity are expected to be about 0.8, the study should include 4 disease-free subjects for each case. If both are equal to 0.9, a 9:1 ratio is optimal. Using equal allocations results in asymptotic variances that are larger by 36% and 64%, respectively, than those obtained with the optimal allocations.

The situation for NPV is symmetric. Now, the optimal ratio of diseased to disease-free subjects is given by

$$P_{\text{NPV}}/(1-P_{\text{NPV}}) = \sqrt{\frac{se \cdot sp}{(1-se)(1-sp)}}. \tag{3.4}$$

Thus, $P_{\mathrm{NPV}} = 1 - P_{\mathrm{PPV}}$ and the optimal allocation for NPV will take a majority of subjects who have the disease.

An interesting property is that the optimal sample allocations are independent of the disease prevalence. Preliminary estimates of the sensitivity and specificity are needed to compute the ratio of diseased to disease-free subjects at the outset of a study.

For some studies, the cost of accruing and testing diseased subjects may exceed the cost for controls. Such differential costs can easily be taken into account in the allocation scheme. The goal now is to minimize the variance of PPV or NPV subject to the constraint $n(cP + Q) = k$, where c is the cost ratio of a diseased to a disease-free subject and k is a constant related to the overall budget. The optimal ratios are then $\sqrt{(1 - se)(1 - sp)/c \cdot se \cdot sp}$ for PPV and $\sqrt{se \cdot sp/c(1 - se)(1 - sp)}$ for NPV. In both cases, the ratio of cases to controls is reduced by a factor of $\sqrt{c}$.

3.2 *Sample sizes*

In this section, we derive formulas for the minimal sample size needed to achieve a lower $1-\alpha$ confidence bound for PPV (or NPV) that exceeds a prescribed limit $1-\gamma$ with a prescribed probability $1-\beta$. Note that this is mathematically equivalent to rejecting a one-sided null hypothesis, at a fixed level of significance, with a prescribed power $1 - \beta$.

With a fixed allocation P, the lower confidence bound for PPV will achieve the goal if and only if the upper $1 - \alpha$ confidence bound for ϕ_1 is sufficiently small. Specifically, we require that

$$U(\hat{\phi}_1) = \hat{\phi}_1 + Z_{1-\alpha}\hat{\sigma}_1(P)/\sqrt{n} \leqslant \log\left(\frac{\omega}{1-\omega}\frac{\gamma}{1-\gamma}\right). \tag{3.5}$$

The above inequality leads to the condition

$$n \geqslant \frac{(Z_{1-\alpha} + Z_{1-\beta})^2\hat{\sigma}_1^2(P)}{\left\{\phi_1 - \log\left(\frac{\omega}{1-\omega}\frac{\gamma}{1-\gamma}\right)\right\}^2}. \tag{3.6}$$

To apply (3.6), we proceed by replacing the true values by current best guesses, $se(0)$, $sp(0)$ and $\phi_1(0) = \log(1 - sp(0)) - \log(se(0))$, and use the guessed values to compute the variance. The optimal allocation can be found from (3.3), using $se(0)$ and $sp(0)$ in the computation.

An essential ingredient in the sample size calculation is careful selection of the value $1 - \gamma$. For example, suppose we are evaluating a "worthless" diagnostic test that does not distinguish at all between the diseased and disease-free subjects. Then $se = 1 - sp$ and PPV $= \omega$. So the bound should certainly be greater than the prevalence, implying that $\gamma < 1 - \omega$. At the other extreme, the bound must be set to a value that can realistically be achieved by the diagnostic test under study. The guessed values of sensitivity and specificity are used in (3.6) to compute the anticipated value $\phi_1(0)$ for ϕ_1. Combining $\phi_1(0)$ with the prevalence ω gives an anticipated value for the PPV, $\mathrm{PPV}(0) = \left[1 + \exp(\phi_1(0))\frac{\omega}{1-\omega}\right]^{-1}$. If $1-\gamma > \mathrm{PPV}(0)$, the goal exceeds expectations and no sample size will give high power of meeting the goal. Instead, the bound must be chosen so that $1 - \gamma \leqslant \mathrm{PPV}(0)$, which implies that

$$\gamma > \frac{(1-\omega)\exp\{\phi_1(0)\}}{\omega + (1-\omega)\exp\{\phi_1(0)\}}. \tag{3.7}$$

The corresponding sample size equation for a $1 - \gamma$ bound on the NPV is

$$n \geqslant \frac{(Z_{1-\alpha} + Z_{1-\beta})^2\hat{\sigma}_2^2(P)}{\left\{\phi_2 - \log\left(\frac{1-\omega}{\omega}\frac{\gamma}{1-\gamma}\right)\right\}^2}, \tag{3.8}$$

with $\sigma_2(P)$ defined in (2.10).

3.3 *Sensitivity to prior assessment*

The study design depends on prior guesses $se(0)$ and $sp(0)$ for the sensitivity and the specificity. We briefly consider here the question of how our recommendations are affected by incorrect guesses.

The sensitivity and specificity affect our study design via $\mathrm{Var}(\hat{\phi}_1) = \sigma_1^2/n$. At the design stage, that variance is approximated using the prior guesses $se(0)$ and $sp(0)$. Consider the ratio of the true variance to the prior guess

$$\mathrm{VR} = \frac{(1-se)/(se \cdot P) + sp/(1-sp)Q}{(1-se(0))/(se(0)\cdot P) + sp(0)/(1-sp(0))Q}. \tag{3.9}$$

The variance ratio exceeds 1 when $se < se(0)$ or when $sp > sp(0)$, that is, if we are overoptimistic regarding the sensitivity or pessimistic about the specificity. The logic is that both of these settings lead to fewer positive responders in the study.

The power for achieving the desired confidence bound is also affected. The power for the actual sensitivity and specificity is given by $1 - \psi(Z_\beta/\sqrt(\mathrm{VR}) + \delta/(\sqrt(n)\sigma_1))$, where $\delta = \phi_1(0) - \phi_1$ is the difference between the value of ϕ_1 for the guessed parameter values and the true parameter values and $\sqrt(n)\sigma_1$ is the true standard deviation. Both the variance ratio and the difference in ϕ are important.

3.4 *Illustration*

We illustrate the ideas on an application for the evaluation of a diagnostic kit with anticipated sensitivity of 0.8 and specificity of 0.95. The disease prevalence in the population of interest is known to be $1/16$. The developers were convinced that the kit would be marketable if the NPV is high. The NPV for a useless test is $1-\omega = 0.9375$ and the NPV for the anticipated sensitivity and specificity is 0.986, so the goal for the NPV must be set between these bounds. We consider the goal of proving, with 80% power at the 5% level, that the NPV is at least $1-\gamma = 0.98$.

The case–control ratio for the optimal sample allocation is given by $\sqrt{(0.8)(0.95)/(0.2)(0.05)} = 8.72$, i.e. $P_{\mathrm{NPV}} = 89.7\%$ of the study subjects should be cases. The anticipated value of ϕ_2 is $\log(0.2/0.95) = -1.56$. The sample size formula then calls for the number of subjects to be at least 220, with the number of cases at least 197 and the number of disease-free subjects at least 23 (rounding up). The expected numbers of negative responders among the diseased and disease-free subjects are 39.4 and 21.9, respectively.

Had we insisted on equal case and control samples, the resulting sample would have required 358 subjects, almost 63% more than for the optimal sample allocation. Simel *and others* (1991) assumed equal sample sizes for most of their examples and did not address the question of optimal allocation. They also modified one example, assuming that 5 disease-free subjects would be available for each diseased subject, due to limited availability of diseased subjects. The resulting sample size was less than that derived for the balanced design. Our results show that such behavior is not surprising.

The sample size is highly sensitive to the assumed parameter values. If the guessed sensitivity is changed to 0.78, the sample size is 355; if it is changed to 0.82, only 151 subjects are needed. The allocation remains close to 10% diseased subjects for all these settings. As the specificity varies from 0.93 to 0.97, the sample size changes from 257 to 188. Here, the allocation does change somewhat, ranging from 12.1% diseased for 0.93 to 8.1% diseased for 0.97.

3.5 *Small samples*

The analysis thus far relies on asymptotic results for the distributions of $\hat{\phi}_1$ and $\hat{\phi}_2$. We summarize here simulations that were carried out to check the validity of the results for smaller samples. More details on the results of the simulations and R code for using our methods and for the simulations can be found in Sections B and C, respectively, of the supplementary material available at *Biostatistics* online (http://www.biostatistics.oxfordjournals. org).

We studied 4 combinations of sensitivity and specificity $(0.75, 0.85)$, $(0.85, 0.75)$, $(0.9, 0.95)$ and $(0.95, 0.9)$. Four sample sizes were included: 30, 50, 70, and 100 subjects. The subjects were divided between diseased and disease-free samples using the optimal allocation for PPV from (3.3), rounding to the nearest integer. We also used allocations ranging from 0.1 to 0.9 in steps of 0.1.

The optimal allocation consistently achieved a lower standard deviation for $\hat{\phi}_1$ than any of the alternative allocations and also had the lowest bias. The standard deviation with equal allocation was typically 15% to 20% higher than that with the optimal allocation. The optimal allocation also had the lowest rate of problematic samples with either $\widehat{sp} = 1$ or $\widehat{se} = 0$.

The coverage probabilities for the 95% confidence bound, with the optimal allocation, ranged from 93.1% to 95.3%. Nominal coverage could be achieved using multipliers ranging from 1.645 (the conventional value) to 1.73 for the lower values of sensitivity and specificity and from 1.645 to 2.05 with the higher values. Simulations were used to determine these multipliers.

4. SAMPLE SIZE FOR BOTH PPV AND NPV

4.1 *Sample allocation*

To establish that both PPV and NPV exceed some minimal level, at given confidence levels and with given powers, the sample size should be the minimal value of n that satisfies

$$n \geqslant \frac{(Z_{\mathrm{PPV},1-\alpha} + Z_{\mathrm{PPV},1-\beta})^2 \sigma_1^2(P)}{\left\{\phi_1 - \log\left(\frac{\omega}{1-\omega}\frac{\gamma_{\mathrm{PPV}}}{1-\gamma_{\mathrm{PPV}}}\right)\right\}^2} = n_{\mathrm{PPV}}(P) \tag{4.1}$$

and

$$n \geqslant \frac{(Z_{\mathrm{NPV},1-\alpha} + Z_{\mathrm{NPV},1-\beta})^2 \sigma_2^2(P)}{\left\{\phi_2 - \log\left(\frac{1-\omega}{\omega}\frac{\gamma_{\mathrm{NPV}}}{1-\gamma_{\mathrm{NPV}}}\right)\right\}^2} = n_{\mathrm{NPV}}(P) \tag{4.2}$$

for some allocation P. We have added subscripts to emphasize that the confidence levels, powers, and desired lower bounds for PPV and NPV need not be equal. The resulting optimization problem for the allocation is to determine P to minimize $\max\{n_{\mathrm{PPV}}(P), n_{\mathrm{NPV}}(P)\}$. The following theorem describes the solution. The proof is given in Section D of the supplementary material available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org).

THEOREM 4.1 Assume that $se + sp > 1$. Then $P_{\mathrm{PPV}} < P_{\mathrm{NPV}}$. Further, we have the following.

1. The equation $n_{\mathrm{PPV}}(P) = n_{\mathrm{NPV}}(P)$ has no roots in the interval $(0, 1)$ or has a unique root $P^* \in (0, 1)$.
2. If no solution exists, if $0 < P^* < P_{\mathrm{PPV}}$, or if $P_{\mathrm{NPV}} < P^* < 1$, then either $n_{\mathrm{PPV}}(P) > n_{\mathrm{NPV}}(P)$ or $n_{\mathrm{PPV}}(P) < n_{\mathrm{NPV}}(P)$ for all $P \in [P_{\mathrm{PPV}}, P_{\mathrm{NPV}}]$. In the former case, the optimal allocation is P_{PPV} and the sample size is given by $n_{\mathrm{PPV}}(P_{\mathrm{PPV}})$. In the latter case, the optimal allocation is P_{NPV} and the sample size is given by $n_{\mathrm{NPV}}(P_{\mathrm{NPV}})$.
3. If $P_{\mathrm{PPV}} \leqslant P^* \leqslant P_{\mathrm{NPV}}$, then the optimal fraction of cases is P^* and the sample size can be computed from either n_{PPV} or n_{NPV}.

4.2 *Illustration*

We return to the application described earlier, now adding requirements on both PPV and NPV. As before, we assume that the goal for NPV is to establish that it exceeds 98%. The anticipated PPV for the test is

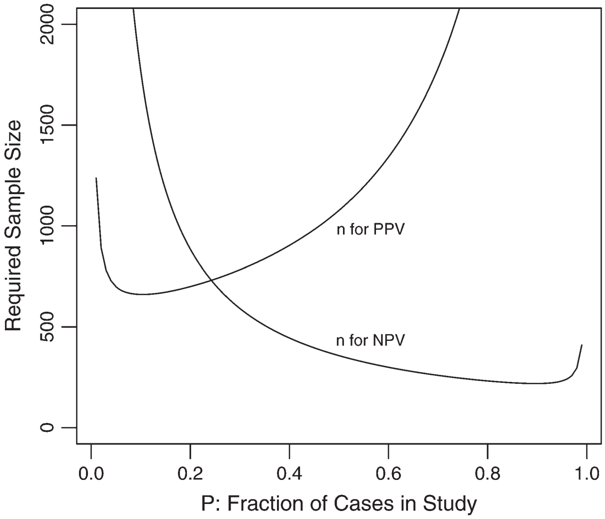

Fig. 1. The required sample sizes n_{PPV} and n_{NPV} versus the sample allocation fraction P for the case study.

51.6%, and we first consider a goal of establishing that the PPV exceeds 40%. For both PPV and NPV, we take 5% as the (one-sided) level of significance and 80% as the desired power.

To find the optimal allocation, we solve the equation $n_{\text{PPV}}(P) = n_{\text{NPV}}(P)$, computing ϕ_1 and ϕ_2 from their guessed values. We use a simple grid search in this step. The solution, $P = 0.242$, is between $P_{\text{PPV}} = 0.103$ and $P_{\text{NPV}} = 0.897$, and thus the optimal allocation calls for 24.2% of the subjects to be cases. Figure 1 plots $n_{\text{PPV}}(P)$ and $n_{\text{NPV}}(P)$ versus P. The minimal sample size (rounding to the nearest integer) is 731, with 177 cases and 554 controls. An even allocation would require 1078 subjects to meet both demands, almost a 50% increase over the optimal sample.

Suppose that we replace the goal of 40% for PPV with a goal of only 25%. That increases the emphasis on the NPV requirement. The optimal allocation now takes 67.5% of the subjects as cases, with 181 cases and 87 controls. Note that the number of cases is almost the same for the 2 requirements on PPV, but the number of controls changes dramatically.

Figure 2 plots the optimal fraction of cases as a function of the bound on PPV. The fraction drops from about 0.9 if the bound is low to about 0.1 if the bound is near the anticipated PPV of 51.6%. For all bounds above approximately 0.45, P_{PPV} is the optimal fraction.

Figures 3 and 4 plot the minimum bounds on the numbers of cases and controls, respectively, in the study against the bound on PPV. The number of cases is not very sensitive to the bound on PPV for most of its range and drops smoothly from near 197 (the number required when considering only the bound on NPV) to about 175 as the focus shifts to the bound on PPV. However, there is a sharp increase in the number of cases when the bound on PPV reaches approximately 0.45. This increase corresponds exactly to the bound at which the optimal fraction of cases no longer decreases. From this point on, the required number of cases increases dramatically, exceeding 700 when the bound on PPV is 0.48.

Figure 4 shows that the required number of controls changes dramatically as the bound on PPV is altered. Only 23 controls are required when the PPV bound is 0.1 or less, but that number increases to 56 for a bound of 0.2, to 143 for a bound of 0.3, and to 554 for a bound of 0.4. When the bound on PPV is fixed at 0.4, the sample depends on the NPV bound only if that bound is very close to the anticipated NPV of 98.6%. For any NPV bound below 97.4%, the optimal sample is the one that concentrates on the PPV bound only, with 68 cases and 593 controls. For tighter bounds on NPV, the required number of cases increases dramatically and there is only a slight decrease in the number of controls. If the bound

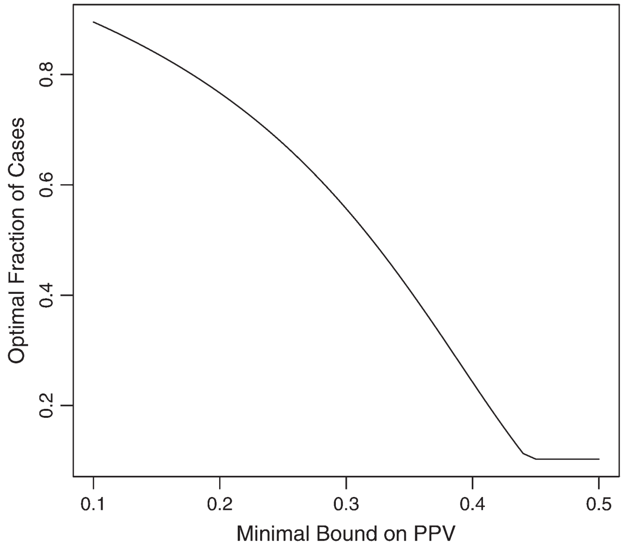

Fig. 2. The optimal fraction of cases as a function of the bound on PPV.

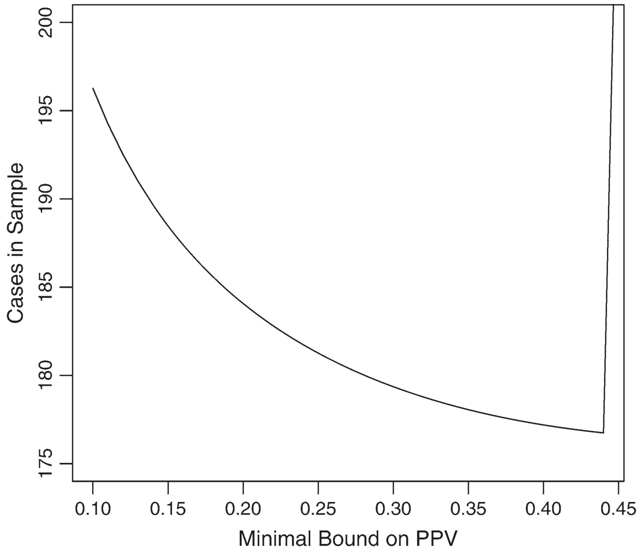

Fig. 3. The number of cases as a function of the bound on PPV.

exceeds 98.5%, the optimal fraction of cases is P_{NPV} and then there is a sharp increase in both the number of required cases and controls.

5. DISCUSSION

We have derived sample size formulas, including optimal allocation to cases and controls, for studies on diagnostic devices whose goal is inference on PPV and NPV. There are 2 surprising features to our

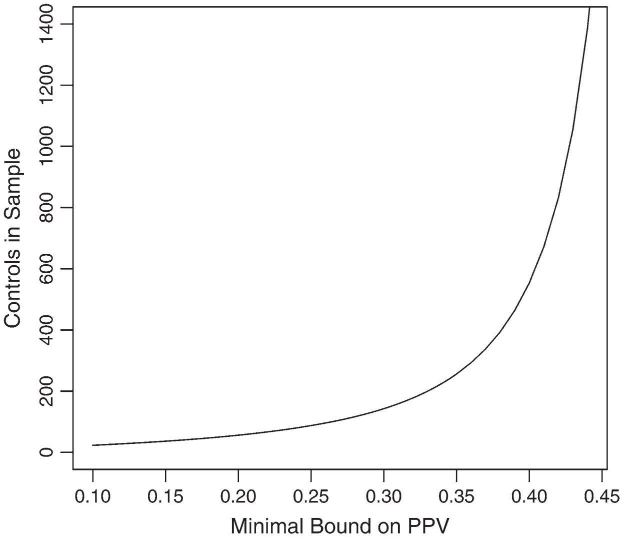

Fig. 4. The number of controls as a function of the bound on PPV.

results. First, balanced allocations can substantially inflate the sample size, even when both NPV and PPV are of interest. Second, when emphasis is on NPV, most of the subjects should be cases, whereas when the emphasis is on PPV, most of the subjects should be disease free. Underlying our sample size results is a general method, applicable to cross-sectional as well as case–control studies, for inference on PPV and NPV in a population whose prevalence differs from that of the original sample.

The transportability of results across populations with differing prevalences relies on the assumption that the characteristics of disease and non-disease, and therefore the sensitivity and specificity of the tests, do not vary across populations. This may be problematic if different populations represent different spectrums of disease. In such settings, separate studies may be required for specific populations. In practice, one needs to consider carefully whether the population in which inferences are desired is suitably similar to the population that provided subjects for a case–control study. This has potential implications for choice of population in a given design.

A main goal in many cohort studies is to compare outcome variables across groups. Balanced or nearly balanced sampling will then provide the most precise comparisons. This would be true, for example, in a clinical study to provide information on secondary end points. In diagnostic testing, our results imply that highly unbalanced samples will typically be much more efficient than balanced samples for inference about PPV and NPV. As we have shown, unbalanced designs will also be much less likely to generate problematic samples in which the estimated sensitivity or specificity equals 1.

Knowledge of the disease prevalence in the population of interest is key to our results. However, in many applications there may not be a single prevalence of interest for defining PPV and NPV. The prevalence may not be known with certainty, or the diagnostic procedure may be desired for use in multiple populations having different prevalences (e.g. for screening large populations versus testing individuals who already have symptoms). In such scenarios, it may be desirable to design a study that deals appropriately with a range of prevalences. We recommend varying prevalences to adjust the bounds on PPV and NPV accordingly to ensure that performance is adequate across the entire prevalence range. For example, in the context of our illustration, had the researchers asserted a prevalence of 3% rather than 6.25%, then the goal for NPV would surely need to be higher than the value of 98% that we adopted. A simple and general solution is to specify the desired bounds for NPV and PPV for each feasible prevalence, solve each of these design problems, and then take the largest resulting sample size. This will be satisfactory for all prevalences, in the sense of achieving adequate power. If the sample sizes differ dramatically across the range, this can serve as a tip-off to the investigators that their prevalence-dependent goals may not be consistent with one another.

If the disease prevalence is unknown, a natural alternative is to conduct a cross-sectional study of the relevant population, which permits estimation of the prevalence and direct estimation of both PPV and NPV. However, much larger sample sizes are required than the ones that we derive here. Consider the settings that we analyzed in our application, with the goal of proving that NPV is greater than 0.98 (with a one-sided 5% test and 80% power). Standard sample size formulas show that this goal requires a sample with at least 2231 negative subjects. Moreover, about $1/16$ of the cross-sectional sample will be positive subjects, so the total sample size must be 2380, more than 10 times as large as the sample based on prior knowledge of the prevalence.

In certain designs, it may be possible to estimate prevalence using a sample from an underlying population and then to estimate sensitivity and specificity using a subsample from the population sample. Such nested case–control studies present similar design issues, in particular when inferences about PPV and NPV are of interest. Estimation of PPV and NPV may now be based on the full sample prevalence estimate and the subsample estimates of sensitivity and specificity (Langholz and Thomas, 1990). The results in the current paper may not be directly applicable because the prevalence is estimated and not fixed. We must acknowledge additional variability caused by the estimated prevalence, as well as the effects of its correlation with estimated sensitivity and specificity. An exception is if one employs an asymptotic regime where the ratio of the size of the subsample to the full sample size converges to zero as the full sample size gets large. In this case, variability in estimating the prevalence is asymptotically negligible. Under such an assumption, the results in the current paper are valid for designing the subsample, treating the prevalence from the finite population as known.

Our results are derived from the asymptotic distributions for the estimated sensitivity and specificity. For small samples, we present simulation results indicating that the optimal allocation continues to be efficient, that the optimal allocation limits the probability of problematic samples, and that the asymptotic 95% confidence limit continues to have close to 95% coverage. Hence, a larger multiplier may be needed to achieve exactly 95% coverage, with a corresponding need to increase the sample size. The supplementary material, available at *Biostatistics* online, includes a set of R functions for computing optimal allocations and sample sizes using the asymptotic formulas and for determining the need for a larger multiplier for the lower confidence bounds for PPV or NPV. Alternative methods might provide improved inference for small samples. For example, one can use the exact distributions of $\hat{se}$ and $\hat{sp}$ to generate profile likelihoods for PPV and NPV. Our design methodology is still appropriate.

ACKNOWLEDGMENTS

We would like to thank Dr Nir Dotan of Glycominds Ltd for posing the problem that we have studied here. We also wish to thank an anonymous referee and the associate editor for comments that helped to improve the presentation. The work of David M. Steinberg was carried out in part while visiting the Department of Biostatistics at the University of Wisconsin-Madison. *Conflict of Interest:* None declared.

REFERENCES

BAKER, S. G., KRAMER, B. S., MCINTOSH, M., PATTERSON, B. J., SHYR, Y. AND SKATES, S. (2006). Evaluating markers for the early detection of cancer: overview of study designs and methods. *Clinical Trials* **3**, 43–56.

BOYKO, E. J. (1994). Ruling out or ruling in disease with the most sensitive or specific diagnostic test: short cut or wrong turn? *Medical Decision Making* **14**, 175–179.

COX, C. (1990). Fieller's theorem, the likelihood and the delta method. *Biometrics* **46**, 709–718.

DUJARDIN, B., VAN DEN ENDE, J., VAN GOMPEL, A., UNGER, J. P. AND VAN DER STUYFT, P. (1994). Likelihood ratios: a real improvement for clinical decision making? *European Journal of Epidemiology* **10**, 29–36.

LANGHOLZ, B. AND THOMAS, D. C. (1990). Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *American Journal of Epidemiology* **131**, 169–176.

LI, J., FINE, J. P. AND SAFDAR, N. (2007). Prevalence dependent diagnostic accuracy measures. *Statistics in Medicine* **26**, 3258–3273.

MOSKOWITZ, C. S. AND PEPE, M. S. (2006). Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. *Clinical Trials* **3**, 272–279.

PEPE, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.

PEPE, M. S., ETZIONI, R., FENG, Z., POTTER, J. D., THOMPSON, M. L., THORNQUIST, M., WINGET, M. AND YASUI, Y. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* **93**, 1054–1061.

SIMEL, D. L., SAMSA, G. P. AND MATCHAR, D. B. (1991). Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *Journal of Clinical Epidemiology* **44**, 763–770.

WANG, W., DAVIS, C. S. AND SOONG, S.-J. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares. *Statistics in Medicine* **25**, 2215–2229.

ZHOU, X.-H., OBUCHOWSKI, N. A. AND MCCLISH, D. K. (2002). *Statistical Methods in Diagnostic Medicine*. New York: John Wiley & Sons.

[*Received April 16, 2007; first revision October 9, 2007; second revision March 2, 2008;*
third revision April 27, 2008; accepted for publication May 12, 2008]