# A Two-Stage Validation Study for Determining Sensitivity and Specificity


## Tor D. Tosteson, Linda Titus-Ernstoff, John A. Baron, and Margaret R. Karagas

Biostatistics and Epidemiology Group, Department of Community and Family Medicine, Dartmouth Medical School, Hanover, New Hampshire



A two-stage procedure for estimating sensitivity and specificity is described. The procedure is developed in the context of a validation study for self-reported atypical nevi, a potentially useful measure in the study of risk factors for malignant melanoma. The first stage consists of a sample of $N$ individuals classified only by the test measure. The second stage is a subsample of size $m$, stratified according the information collected in the first stage, in which the presence of atypical nevi is determined by clinical examination. Using missing data methods for contingency tables, maximum likelihood estimators for the joint distribution of the test measure and the "gold standard" clinical evaluation are presented, along with efficient estimators for the sensitivity and specificity. Asymptotic coefficients of variation are computed to compare alternative sampling strategies for the second stage. — Environ Health Perspect 102(Suppl 8):11–14(1994)



Key words: validation study, sensitivity, specificity, atypical nevi, missing data methods


## Introduction

In epidemiologic research, validation studies often have as a goal the determination of the sensitivity and specificity of a "test" for the presence of a risk factor. This test is usually easier and cheaper to administer than a more accurate "gold standard" method. Knowledge of the specificity and sensitivity of the test can be used in sample size calculations for subsequent studies of the effect of the risk factor and to adjust relative risk estimates for measurement error. This knowledge can also be important in the evaluation of the clinical utility of the test as a diagnostic tool.

An example of the use of a test measure occurs in epidemiologic studies of incidence and prevalence of neoplastic skin disease, where the use of self-reported counts of atypical nevi rather than clinical examination can lead to substantial reduction in study costs. Recent studies of the importance of the presence of nevi as a predictor of melanoma have employed both self-reported counts [e.g., Bain et al. (*1*)] and physical examination by a dermatologist or trained interviewer [e.g., Augustsson et al. (*2,3*)]. However, the sensitivity and specificity of self-reports of the presence of atypical nevi are not well known, although validation studies of self-reported aggregate measures of body nevus density (*4*) suggest a measurable correlation with interviewer determinations, and studies using both types of data are able to show significant relative risks for melanoma using either measure (*5*).

A recent survey in Sweden sucessfully solicited mail questionnaire responses from 50,000 women between the ages of 35 and 55 ("Women's Lifestyles and Health Study," H-O Adami, unpublished). Respondents were asked to examine their skin for the presence of atypical moles. At present, no analysis of the results of the survey has been performed. A number of alternative designs are being considered to use the survey to estimate the sensitivity and specificity of self-reported atypical nevi (SRAN). Specifically, a two-stage procedure is proposed in which the first stage would be a random sample of the original cohort to determine the prevalence of SRAN. In the second stage, a random sample of the individuals identified in the first stage would be examined to obtain physician-diagnosed atypical nevi (PDAN). The second-stage sample could be stratified by level of SRAN as determined in the first stage. In this case, an advantage in estimating sensitivity and specificity could be achieved by manipulating the relative sizes of the test positive and test negative samples.

The purpose of this article is to demonstrate the utility of a two-stage validation study design as applied to self-reported counts of atypical nevi. In the sections that follow, estimators for sensitivity and specificity in the proposed designs are described, along with their approximate standard errors. The relative efficiencies of the designs are used to compare the proposed designs and to assess the design for the SRAN validation study.

## Design Options

Figure 1 depicts two possible designs for a validation study. Under design A, a first-stage, simple random sample of size $N$ is take to obtain SRAN. Then a second-stage subsample of size $m$ is taken to obtain PDAN. As a result, the validation study contains $m$ complete observations.

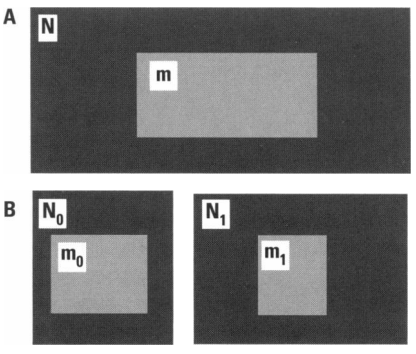


**Figure 1.** Design options for two-stage study. (*A*) Select $N$ to get SRAN and random subsample $m$ to get PDAN. (*B*) Select $N$ to get SRAN. Take random sample of size $m_0$ from negative SRAN and $m_1$ from positive SRAN.


This paper was presented at the 4th Japan-US Biostatistics Conference on the Study of Human Cancer held 9–11 November 1992 in Tokyo, Japan.

This research was supported by Grants CA50597 and CA23108 from the National Cancer Institute, and Grant ACS SIG-17 from the American Cancer Society.



Address correspondence to Tor D. Tosteson, Biostatistics and Epidemiology Group, Dept. of Community and Family Medicine, Dartmouth Medical School, Hanover, NH 03755. Telephone (603) 650-1492. Fax (603) 650-1103.

At the first stage, design B also specifies a sample of size $N$ to obtain SRAN. However, in the second stage the number of individuals selected who have a positive SRAN, $m_1$, and the number of individuals who have a negative SRAN, $m_0$ are controlled, under the constraint that $m_0 + m_1 = m$. For simplicity, it is assumed that the counts of atypical nevi have been reduced to a binary classification (i.e., presence or absence of nevi). The object of the design is to manipulate the ratio $m_1/m_0$ to achieve a higher precision than the other designs at an equivalent cost. Of course, it is not possible for $m_1$ to exceed the number of test positives in the first stage, $N_1$, or for $m_0$ to exceed the number of test negatives in the first stage, $N_0$.

The data from both these designs can be arranged as in Figure 2. The number of complete observations is $m$, the size of the second stage in both designs A and B. The number of incomplete observations is $r = N - m$. These are the individuals not selected for the second stage, and thus only having the SRAN (test) classification. In design A, the margins $m_1$ and $m_0$ are not fixed, whereas in design B, they are chosen to achieve design objectives such as minimizing the variance of estimators of sensitivity and specificity.

## Estimation

In describing the methods of estimation and the properties of the estimators, the following notation will be adopted. Let $\{\pi_{ij}, i = 0,1; j = 0,1\}$ represent the joint distribution of the binary indicators SRAN and PDAN, with $i$ indexing the level of SRAN and $j$ indexing the level of PDAN. Thus $\pi_{10}$ is the probability that SRAN is one or more and PDAN is zero. Sensitivity can be expressed as $s = \pi_{11}/(\pi_{01} + \pi_{11})$, and specificity as $S = \pi_{00}/(\pi_{00} + \pi_{10})$. The prevalence is $\pi = \pi_{01} + \pi_{11}$. To simplify the following developments, let $\theta = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})'$.

The primary objective of a validation study is to estimate $s$ and $S$. However, the prevalence may also be of interest, and

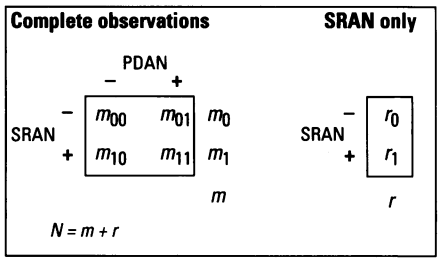


**Figure 2.** Representation of data from two-stage designs.

could be included in the development that follows with little trouble.

A method of estimation can be devised for designs A and B by treating the $r$ individuals not included in the second stage sample as having missing data for the gold standard PDAN. In design A, the $m$ individuals included in the second stage are selected randomly without regard to their SDAN status. Conversely, the $r$ individuals not included in the second stage can also be considered a simple random sample of the $N$ individuals in the first stage.

For the second stage of design B, $m_0$ individuals are selected randomly from among the $N_0$ SDAN negative individuals identified in the first stage, and $m_1$ individuals are selected randomly from the $N_1$ SDAN positive individuals. Thus the probability of having a missing value of PDAN (i.e., not being included in the second stage) depends upon SDAN status.

For both design A and design B, the data for PDAN are what Little and Rubin (*6*) refer to as "missing at random." In fact, they discuss a closely related problem in their chapter on models for partially classified contingency tables. In the section dealing with monotone missing data patterns, they give formulas for the maximum likelihood estimates of the elements of $\theta$, which in this case is the joint distribution of PDAN and SRAN. These formulas may be written as

$$\hat{\pi}_{ij} = \frac{\left[m_{ij} + (m_{ij}/m_i)r_i\right]}{N} \qquad [1]$$

Intuitively, the estimators can be thought of as distributing the individuals not selected for the second stage between the two PDAN classifications. Formulas for the elements of the asymptotic covariance matrix of $\theta$ are given in Little and Rubin (*6*, section 9.2.3). This covariance matrix will be denoted as $\Sigma_\theta$. More detailed expressions, provided in the appendix to the current article, clearly illustrate the dependence of the precision of the distribution estimates on the choice of $N$, $m_1$, and $m_0$.

Consistent and efficient estimates for $s$ and $S$ are easily obtained as

$$\hat{s} = \frac{\hat{\pi}_{11}}{\hat{\pi}_{01} + \hat{\pi}_{11}}, \quad \text{and} \quad \hat{S} = \frac{\hat{\pi}_{00}}{\hat{\pi}_{00} + \hat{\pi}_{10}}. \qquad [2]$$

Using a delta method approximation, the asymptotic variances of $S$ and $s$ are

$$\mathrm{Var}(\hat{s}) = \left(\frac{\partial s}{\partial q}\right)' \Sigma_\theta \left(\frac{\partial s}{\partial q}\right)$$
$$\mathrm{Var}(\hat{S}) = \left(\frac{\partial S}{\partial q}\right)' \Sigma_\theta \left(\frac{\partial S}{\partial q}\right) \qquad [3]$$

where

$$\left(\frac{\partial s}{\partial q}\right) = \frac{1}{(\pi_{01} + \pi_{11})^2} \begin{pmatrix} 0 \\ -\pi_{11} \\ 0 \\ -\pi_{01} \end{pmatrix}$$

and

$$\left(\frac{\partial S}{\partial \theta}\right) = \frac{1}{(\pi_{00} + \pi_{10})^2} \begin{pmatrix} \pi_{10} \\ 0 \\ -\pi_{00} \\ 0 \end{pmatrix}$$

Consistent estimates of the variances of $s$ and $S$ are obtained by replacing $\pi_{ij}$ by the estimates $\hat{\pi}_{ij}$.

## Efficiency Calculations

The variance formulas given in Equation 3 can be used to compare the efficiency of designs A and B for a given sensitivity, specificity, prevalence, and first- and second-stage sample sizes. The coefficients of variations $\mathrm{CV}(\hat{s}) = \sqrt{\mathrm{Var}(\hat{s})}/s$ and $\mathrm{CV}(\hat{S}) = \sqrt{\mathrm{Var}(\hat{S})}/S$ are used as a basis for the comparisons shown in this section.

Figure 3 shows the coefficients of variation for sensitivity and specificity for a first stage sample size of $N = 10{,}000$, a second stage sample size of $m = 400$, and a range of values for $m_1$ in design B. A sensitivity of $s = 0.9$, a specificity of $S = 0.7$, and a prevalence of $\pi = 0.15$ are assumed. For design A, $m_1$, the number of PDAN positives in the second stage, is not controlled, and the coefficients of variation are shown as constant over all values of $m_1$. The plots reveal that, in design B, an increase in $m_1$ tends to increase the imprecision of estimation for sensitivity and decrease the imprecision for specificity. This reflects the nature of the adjustment for the stratified sampling scheme.

Figure 4 shows the sum of the coefficients of variation for sensitivity and specificity for designs A and B. All the assumptions are the same as in Figure 3, except that the lower panel uses a sensitivity of $s = 0.99$. The results show that design B is better than design A for only a small range of values for $m_1$ for a sensitivity of 0.9, whereas a fairly broad range of values

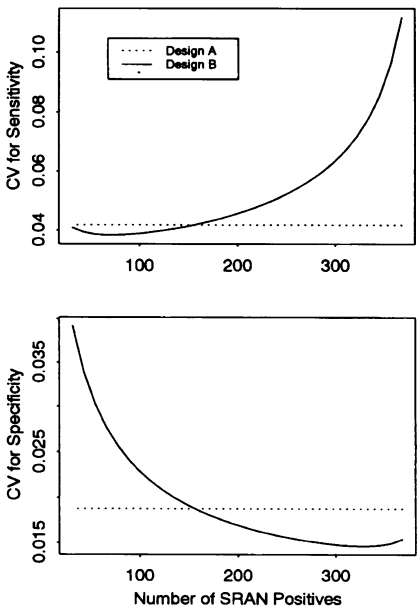


**Figure 3.** Coefficients of variation (CV) for specificity and specificity as a function of relative stratum sizes.

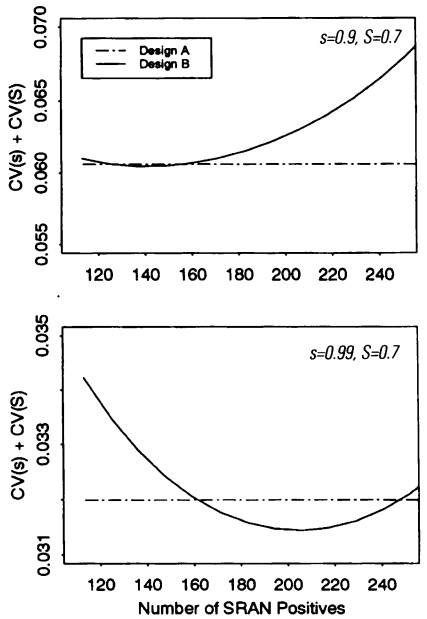


**Figure 4.** Total coefficient of variation, for two different sensitivities.

**Table 1.** Optimal sample sizes and coefficients of variation for estimates of sensitivity and specificity by design type and test characteristics of SRAN.

| Assumed test characteristics for SRAN | | Design B stratified second-stage sample | | | |
|---|---|---|---|---|---|
| | | Optimal second-stage sample sizes | | Coefficient of variation | |
| Sensitivity | Specificity | SRAN, + | SRAN, − | Sensitivity, CV% | Specificity, CV % |
| 0.7 | 0.7 | 129 | 271 | 7.9 | 2.3 |
| 0.7 | 0.9 | 87 | 313 | 7.8 | 1.3 |
| 0.9 | 0.7 | 126 | 274 | 4.0 | 2.4 |
| 0.9 | 0.9 | 97 | 303 | 4.2 | 1.4 |

for $m_1$ gives an improved precision for $s$ = 0.99.

## A Validation Study for SRAN

The ability to self-diagnose atypical nevi has important implications for screening efforts aimed at prevention and early detection of melanoma. If it can be shown that atypical nevi are accurately self-reported, large population surveys aimed at targeting high risk groups for intervention or prevention, could be inexpensively conducted through mail surveys. In addition to screening applications, the validation study may provide a foundation for population-based follow-up studies of melanoma risk associated with atypical nevi, case-control evaluation of risk factors for atypical nevi, and preventive efforts against melanoma.

In the recent survey in Sweden, 50,000 women between the ages of 35 and 55 (above) were asked to examine their lower extremities for the presence of irregular moles resembling the atypical nevi pictured in color photographs. A proposed validation study would evaluate self-screening efforts of survey respondents living in the Uppsala area using the two-stage design proposed in this article (L. Titus-Ernstoff, unpublished grant application).

In the first stage, a random sample of 2000 women living in one of two counties adjacent to the Uppsala medical facilities will be selected from among the respondents to the original survey. Based upon the results of a population-based study of atypical nevi in Gothenburg, Sweden (*2,3*), it is estimated that about 300 of the 2000 women who live in the eligible counties will report an atypical nevus. A total of 400 women will be selected randomly for the second-stage validation study. Study recruits will undergo a physician-conducted skin examination, during which pigmentation characteristics—including mole counts—and atypical mole counts will be recorded.

Table 1 shows optimal strata sizes and coefficients of variation for design B, as applied to the Gothenburg example. Four assumptions are used for sensitivity and specificity. Anticipated coefficients of variation are shown for $\hat{s}$ and $\hat{S}$. These results demonstrate that a two-stage design for measuring sensitivity and specificity can improve on overall precision by controlling the number of test positives and negatives in the second stage.

## Discussion

A new proposal for two-stage designs of validation studies has been presented. These designs may lower the cost of validation studies by reducing the use of the more expensive gold standard measurement. A method of estimation has been derived by using methods for missing data.

It should be noted that similar issues have been been considered in investigations of "verification" or "work-up bias" (*7*) in diagnostic test assessment. An example of such a situation might be a study of "silent" coronary heart disease in which determining the gold standard disease classification would involve giving an invasive test to populations of apparently healthy individuals, thus incurring a high degree of noncompliance. In these situations, the decision to apply the invasive procedure may be influenced by the results of the screening test, with a positively screened individual being more likely to receive the gold standard. This selection can bias the estimate of the operating characteristics of the tests, and has been termed work-up bias.

The two-stage design suggested in this paper can be viewed as a study which deliberately incurs work-up bias. The estimates of sensitivity and specificity presented using corrections for missing data are equivalent to the bias-corrected estimators suggested by Begg and Greenes (*7*).

## Appendix

### Large Sample Covariance Matrix for $\hat{\theta}$

Little and Rubin (6, section 9.2.3) give estimates of the large sample covariance matrix of $\hat{\theta}$, $\Sigma_\theta$, as follows. The diagonal elements are given by

$$\hat{\mathrm{Var}}\left(\hat{\pi}_{ij}\right)=\frac{\hat{\pi}_{ij}\left(1-\hat{\pi}_{ij}\right)}{m}\left[1-\frac{\hat{\pi}_{j\cdot i}}{1-\hat{\pi}_{ij}}\frac{r}{n}+c_i\frac{1-\hat{\pi}_{j\cdot i}}{1-\hat{\pi}_{ij}}\right] \qquad [4]$$

where $\hat{\pi}_{j\cdot i} = \hat{\pi}_{ji}/(\hat{\pi}_{i0} + \hat{\pi}_{i1})$ and $c_i = m\,(\hat{\pi}_{i0} + \hat{\pi}_{i1})/m_i - 1$. The off-diagonal elements are given as

$$\hat{\mathrm{Cov}}\left(\hat{\pi}_{ij},\hat{\pi}_{ij'}\right)=\frac{-\hat{\pi}_{ij}\hat{\pi}_{ij'}}{m}$$
$$\left[1+\frac{\left\{1-\left(\hat{\pi}_{i0}+\hat{\pi}_{i1}\right)\right\}}{\left(\hat{\pi}_{i0}+\hat{\pi}_{i1}\right)}\frac{r}{n}+\frac{c_i}{\left(\hat{\pi}_{i0}+\hat{\pi}_{i1}\right)}\right],\left(j\neq j'\right)$$
$$\hat{\mathrm{Cov}}\left(\hat{\pi}_{ij},\hat{\pi}_{i'j'}\right)=\frac{-\hat{\pi}_{ij}\hat{\pi}_{i'j'}}{m},\left(i\neq i'\right) \qquad [5]$$

To obtain asymptotic variances for the purposes of comparing designs, the estimates in these formulas are replaced by the value of the parameters $\theta$. To compute the asymptotic variances for design A, $c_i$ is set to 0.

### REFERENCES


1. Bain C, Colditz GA, Willet WC, Stampfer MJ, Green A, Bronstein BR, Mihm MC, Rosner B, Hennekens CH, Speizer FE. Self-reports of mole counts and cutaneous malignant melanoma in women: methodological issues and risk of disease. Am J Epidemiol 127:703–12 (1988).
2. Augustsson A, Stierner U, Rosdahl I, Suurkula M. Common and dysplastic naevi as risk factors for cutaneous malignant melanoma in a Swedish population. Acta Derm Venereol (Stockholm) 71:518–24 (1991).
3. Augustsson A, Stierner U, Rosdahl I. Prevalence of common and dysplastic naevi in a Swedish population. Br J Dermatol 124:152–156 (1991).
4. Walter SD, Marrett LD, Hertzman C. Reliability of interviewer and subject assessments of nevus counts in a study of melanoma. J Clin Epidemiol 44:633–40 (1991).
5. Beral V, Evans S, Shaw H, Milton G. Cutaneous factors related to the risk of malignant melanoma. Br J Dermatol 109:165–172 (1983).
6. Little JA, Rubin DB. Statistical Analysis with Missing Data. New York:John Wiley & Sons; 1987.
7. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 39:207–215 (1983).