Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2016, Vol. 25(4) 1313–1329

DOI: 10.1177/0962280213480420
smm.sagepub.com
SAGE

Two-stage sampling designs for external validation of personal risk models

Alice S Whittemore and Jerry Halpern

Abstract

We propose a cost-effective sampling design and estimating procedure for validating personal risk models using right-censored cohort data. Validation involves using each subject's covariates, as ascertained at cohort entry, in a risk model (specified independently of the data) to assign him/her a probability of an adverse outcome within a future time period. Subjects are then grouped according to the magnitudes of their assigned risks, and within each group, the mean assigned risk is compared with the probability of outcome occurrence as estimated using the follow-up data. Such validation presents two complications. First, in the presence of right-censoring, estimating the probability of developing the outcomes before death requires competing risk analysis. Second, for rare outcomes, validation using the full cohort requires assembling covariates and assigning risks to thousands of subjects. This can be costly when some covariates involve analyzing biological specimens. A two-stage sampling design addresses this problem by assembling covariates and assigning risks only to those subjects most informative for estimating key parameters. We use this design to estimate the outcome probabilities needed to evaluate model performance and we provide theoretical and bootstrap estimates of their variances. We also describe how to choose two-stage designs with minimal efficiency loss for a parameter of interest when the quantities determining optimality are unknown at the time of design. We illustrate these methods by using subjects in the California Teachers Study to validate ovarian cancer risk models. We find that a design with optimal efficiency for one performance parameter need not be so for others, and trade-offs will be required. A two-stage design that samples all outcome-positive subjects and more outcome-negative than censored subjects will perform well in most circumstances. The methods are implemented in Risk Model Assessment Program, an R program freely available at http://med.stanford.edu/epidemiology/two-stage.html.

Keywords

Bootstrap, calibration, competing risks, discrimination, personal risk models, two-stage sampling

Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA

Corresponding author:
Alice S Whittemore, Department of Health Research and Policy, Redwood Building, Stanford University School of Medicine, Stanford, CA 94305, USA.
Email: alicesw@stanford.edu

1 Introduction

A personal risk model assigns to an individual a probability of developing an adverse outcome in a clinically relevant time period, using his or her covariates. Examples include models for breast cancer development within 5 years,[1] and prostate cancer recurrence within 10 years of diagnosis.[2] The outcome probabilities of interest have been called absolute risks by Benichou and Gail,[3] who stress the importance of using absolute risk as the metric when evaluating disease prevention strategies. Discussion of risk model development, including choice of covariates, allowance for competing mortality and choice of risk period length, can be found in Gail and Pfeiffer,[4] Janes et al.,[5] and Whittemore.[6]

We address the problem of how to use cohort data to assess the performance of a risk model that has been completely specified using data external to the cohort. Such assessment involves applying the model to subjects in a cohort who have been followed for occurrence of the outcome or death from other causes during the defined risk period. Model performance is assessed by estimating outcome probabilities within subgroups of subjects determined by assigned risk, and then examining the behavior of assigned risks in relation to these outcome probabilities. Consider, for example, the problem of validating a prespecified ovarian cancer risk model with data from the California Teachers Study (CTS). This cohort study consists of 133,479 female California public school teachers and administrators who completed a mailed questionnaire in 1995–1996, and who have been followed for subsequent cancer incidence and mortality through 31 December 2007 (see references[7,8] for more information). We used the baseline covariates of each eligible subject in an ovarian cancer risk model to assign her a probability of developing ovarian cancer within 12 years of cohort entry. To check the model's accuracy (also called calibration), we compare how well its assigned risks agree with the estimated 12-year ovarian cancer probabilities in subgroups of women. For example, Figure 1A is a plot of such probabilities versus mean assigned risks in quintiles of risk assigned by a model described in section 4 and in the Supplement. Such plots are called attribute diagrams.[9] Model inaccuracy can be summarized by the bias statistic B,[6] which averages the squared vertical distances of the points from the diagonal line in this figure. A test of the null hypothesis that the model is well calibrated, i.e. that $B = 0$, is provided by the Hosmer–Lemeshow test statistic.[10]

In addition to a model's accuracy, we also need to know how well it discriminates individuals with substantially different actual risks. A common measure of such discrimination is the model's concordance, which is the probability that it assigns a higher risk to a woman who develops ovarian cancer than to one who does not.[3] This measure (which equals the AUC, defined as the area under the (AUC) model's receiver operating characteristic curve),[11,12] ranges from 0.5 for a model with no discrimination to 1.0 for a model that assigns higher risks to all who develop ovarian cancer than to all who do not develop the disease during the risk period.

Assessing a model's accuracy and discrimination presents two complications. First, some subjects have unknown outcome status. For the CTS cohort, for example, we assigned each subject a follow-up time, defined as the number of days between her cohort entry and the first occurrence of ovarian cancer, death, last observation and 12 years of follow-up. Thus, subjects are of three types: (i) outcome-positive subjects, who develop ovarian cancer within the risk period; (ii) outcome-negative subjects, who die from other causes within the period or who survive it without the outcome; and (iii) outcome-unknown subjects (also called censored subjects), who were last observed alive and outcome-free before the full 12 years of follow-up. For the CTS data, 32% of eligible subjects were censored.

In the absence of censored subjects, an unbiased estimate of the outcome probability within a risk group is just the binomial proportion of those who develop the outcome within the risk period.

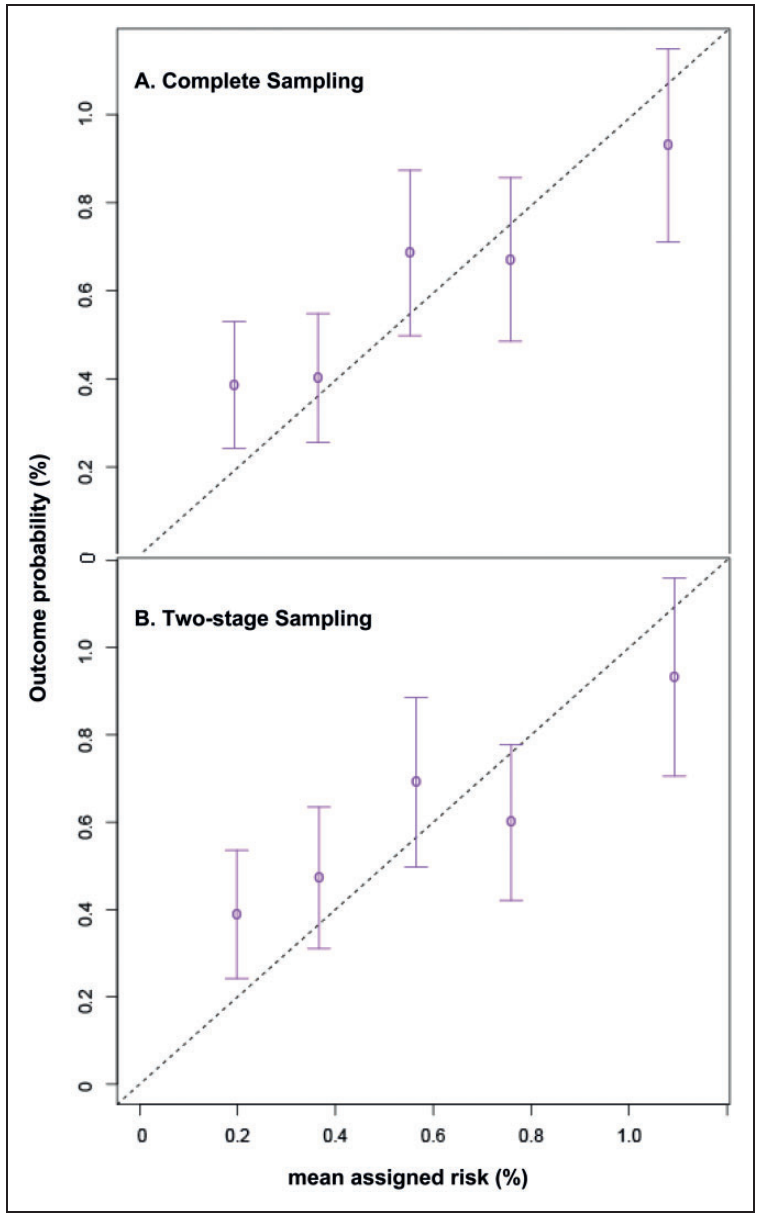

Figure 1. Estimated outcome probabilities (and 95% confidence intervals) versus mean assigned risk in quintiles of risk assigned by Model 1 (see text). (A) complete sampling and (B) two-stage sampling with two sampling categories: outcome positive (sampled with a probability of 1%) and all other subjects (sampled with a probability of 10%).

This binomial estimate has been used in applications for which the fraction of censored subjects was small.[13] However, when censored subjects are numerous, their exclusion can cause substantial upward bias in outcome probability estimates. The magnitude of this bias depends on the ratio of censoring probability to event probability, where an event is outcome or death during the risk

period. For example, if the censoring probability is one-tenth the event probability, excluding censored subjects produces an outcome probability estimate roughly 1.1 times its actual value. In contrast, if the censoring probability is three times the event probability (as observed in the CTS data), the outcome probability estimate is roughly four times the actual value. Unbiased estimates require an approach that accommodates the survival times of censored subjects, while treating death from other causes as a competing risk.[14–16] This approach specifies the group-specific outcome probability (i.e. absolute outcome risk) as the cumulative incidence of the outcome at 12 years, where the latter is obtained by integrating a function of the event-specific hazards over the risk period.[14]

Rare outcomes such as ovarian cancer present a second complication, because validation requires assembling covariates and assigning model risks to large numbers of cohort members, which can be costly when some of the covariates involve biospecimen analysis. For example, the ovarian cancer risk model validation involves 40,139 CTS subjects who met the eligibility criteria, only 227 of whom developed ovarian cancer. Here, we describe a cost-efficient two-stage sampling design that oversamples those subjects most informative for estimating the outcome probabilities. In stage 1, we obtain the information needed to classify all subjects into a few broad sampling categories as determined by, say, their easily available covariates or their outcomes (positive, negative and unknown). In stage 2 we randomly sample each category with a category-specific probability chosen to yield good precision for a performance parameter of interest. We assemble the full set of covariates (both easily available and costly), assign model risks, and analyze survival data only for subjects sampled at stage 2. We then use these data to estimate the distribution of model-assigned risks in the population from which the entire cohort was sampled, and the probability of developing the outcome conditional on the assigned risks.

An important design question concerns the choice of sampling categories and sampling probabilities when evaluating one or more risk models. In a seminal 1938 paper, Neyman defined an optimal two-stage design as one that minimizes, for given total cost, the variance of a key parameter of interest.[17] This strategy can be illustrated with the following example. Suppose we wish to compare a risk model based on readily available covariates (Model 1) to an expanded model that includes costly additional biomarkers, such as genetic mutations conferring elevated ovarian cancer risk (Model 2). Interest might focus on the additional sensitivity and specificity of Model 2 to identify women at high risk, for screening purposes. If the budget allows biomarker assessment for $k < n$ of the n subjects, which subjects do we choose?

In some situations we can address this question by using easily available data on subjects' covariates and outcomes to stratify them into sampling categories. For rare outcomes, for example, choosing all of the outcome-positive subjects helps assess a model's ability to discriminate those who do and do not develop the outcome. We also can assign Model 1 risks to all subjects and use these risks to classify the subjects into sampling categories for assessing the costly covariates. In general however, selecting an optimal sampling partition and set of sampling probabilities may require the very information we want to infer—namely the distribution of risks as assigned using the costly covariates of Model 2. This catch-22, which is similar to that involved in power calculations, can be addressed by using subjects' available data to estimate their values for the costly covariates, then assigning provisional Model 2 risks, classifying subjects into risk groups and performing the calculations needed to approximate the optimal design. We illustrate this strategy by applying it to the CTS data and we compare the performance of designs using covariate-based and outcome-based sampling categories.

Section 2 begins with a brief review of nonparametric maximum-likelihood estimates (NPMLEs) for event-specific absolute risks as applied to data from the complete cohort.[14–16,18] Dinse and

Larson[19] have noted the advantages of expressing these estimates in terms of simple NPMLEs of the event-specific hazards. We show that this approach has the additional advantage of easy extension to yield simple weighted absolute risk estimates using a two-stage sampling design. We also provide closed-form and bootstrap estimates of the variances of these estimates. In section 3, we use simulations to examine the performance of two-stage designs chosen to minimize the variance of either the estimated model bias $\hat{B}$ or its AUC estimate. In section 4, we use data from the CTS to illustrate the inferences and the choice of two-stage design. Section 5 concludes with a brief discussion.

2 Methods

We wish to assess the accuracy and discrimination of a risk model using cohort data that is independent of the data used for model development. When the covariates x needed by the model are available for all subjects, we can use the model to assign a risk $r = f(x)$ to each subject, and then partition the cohort into L subgroups having similar risks. (In section 5 we describe alternatives to such grouping of subjects.) Our goal is to estimate the parameter $\theta = (\gamma, \pi)$ and the covariance matrix of the estimate, $\hat{\theta} = (\hat{\gamma}, \hat{\pi})$. Here $\gamma = (\gamma_1, \ldots, \gamma_{L-1})$ specifies the subject's multinomial group membership probabilities with $\gamma_L = 1 - \sum_{\ell=1}^{L-1} \gamma_\ell$, and $\pi = (\pi_1, \ldots, \pi_L)$ specifies the risk-group-specific outcome probabilities. The estimate $\hat{\theta}$ then allows assessment of model accuracy and discrimination. For example, accuracy can be assessed with the bias statistic

$$\hat{B} = B\left(\hat{\theta}\right) = \left[\sum_{\ell=1}^{L} \hat{\gamma}_\ell (\hat{\pi}_\ell - r_\ell)^2\right]^{1/2} \tag{1}$$

where r_ℓ is the mean assigned risk in group $\ell, \ell = 1, \ldots, L$. The null hypothesis $B(\theta) = 0$ (i.e. that the model is well calibrated to the population risks) can be tested by referring the Hosmer–Lemeshow statistic $n(\hat{\boldsymbol{\pi}} - \mathbf{r})^T \hat{\Sigma}^{-1} (\hat{\boldsymbol{\pi}} - \mathbf{r})$ to a chi-square distribution on L degrees of freedom.[10] Here, $\hat{\Sigma}$ is an estimate of the covariance matrix of $\hat{\boldsymbol{\pi}}$ and $\mathbf{r} = (r_1, \ldots, r_L)$ with r_ℓ denoting the mean assigned risk in group $\ell, \ell = 1, \ldots, L$. Model discrimination can be assessed via the concordance statistic[11]

$$\widehat{AUC} = \widehat{AUC}\left(\hat{\theta}\right) = \sum_{\ell=1}^{L} \sum_{\ell' > \ell} \hat{\gamma}_\ell \hat{\gamma}_{\ell'} \frac{\hat{\pi}_{\ell'}(1 - \hat{\pi}_\ell)}{\hat{\pi}(1 - \hat{\pi})} + \frac{1}{2} \sum_{\ell=1}^{L} \hat{\gamma}_\ell^2 \frac{\hat{\pi}_\ell (1 - \hat{\pi}_\ell)}{\hat{\pi}(1 - \hat{\pi})} \tag{2}$$

where $\hat{\pi} = \sum_{\ell=1}^{L} \hat{\gamma}_\ell \hat{\pi}_\ell$ is an estimate of the outcome prevalence in the population. Estimates for the variances of these and other measures of model performance can be obtained from the estimated covariance matrix of $\hat{\theta}$ using a standard Taylor series expansion (the 'delta method').

2.1 Complete cohort sampling

Suppose we have used a risk model to assign to each of n subjects a probability of outcome development during a future period $[0, t_*)$, and that we use these assigned risks to classify the subjects into L risk groups. We observe the risk group and survival data (T, ε) of each subject, where T is the time from risk assignment to the first of outcome occurrence, death, t_* or last observation, and $\varepsilon = (\varepsilon_1, \varepsilon_2)$, where ε_1 and ε_2 are indicators for outcome occurrence and death, respectively. Times to outcome or death are unobserved for censored subjects, who at time $T < t_*$ were last observed alive and outcome-free ($\varepsilon = (0, 0)$). Our goal is to use these data to estimate the

risk-group-specific outcome probabilities $\pi_1, \ldots, \pi_L$, where $\pi_\ell = \Pr[T < t_*, \varepsilon_1 = 1]$ among subjects in risk group $\ell, \ell = 1, \ldots, L$. In terms of competing risk theory, $\pi_\ell = F_{\ell 1}(t_*)$, where

$$F_{\ell\tau}(t) = \int_0^t \lambda_{\ell\tau}(x) \exp\left\{-\int_0^x [\lambda_{\ell\tau}(u) + \lambda_{\ell\tau}(u)]\mathrm{d}u\right\}\mathrm{d}x \tag{3}$$

is the cumulative incidence function for an event of type $\tau, \tau = 1, 2$, with $\lambda_{\ell\tau}(\cdot)$ denoting the event τ hazard in risk group ℓ. It is straightforward to show that, conditional on the counts n_ℓ of subjects in the L risk groups, the survival data within different risk groups are independent, and their distribution does not depend on the vector γ of multinomial group membership probabilities. Therefore, the asymptotic covariance matrix Ω of the maximum-likelihood estimate $\hat{\theta}$ is a block diagonal, with blocks Γ and Σ, where $\Gamma = diag(\gamma) - \gamma\gamma^T$ is the covariance matrix of the multinomial estimate $\hat{\gamma} = n^{-1}(n_1, \ldots, n_{L-1})$ and Σ is the diagonal matrix whose diagonal entries are the asymptotic variances of the $\hat{\pi}_\ell, \ell = 1, \ldots, L$.

As the cumulative incidence function $F_{\ell\tau}(t)$ of equation (3) is completely specified by the event-specific hazard functions $\lambda_{\ell\tau}, \tau = 1, 2$, we can estimate the outcome probabilities π_l by obtaining NPMLEs of these hazards and using them in equation (3)[18,19] (see also Kalbfleisch and Prentice,[14] Section 8.23, p. 254). Specifically, let M_l denote the number of distinct event times in subgroup l, with $M = \sum_{\ell=1}^{L} M_\ell$. The risk-group- and event-specific hazard functions are replaced by the $2M$-dimensional vector $\lambda = (\lambda_1, \ldots, \lambda_L)$, where

$$\lambda_\ell = \left(\lambda_{\ell 11}, \lambda_{\ell 21}, \ldots, \lambda_{\ell 1 M_l}, \lambda_{\ell 2 M_\ell}\right) \tag{4}$$

is the $2M_\ell$-dimensional vector of discrete hazards taking values at the M_ℓ distinct event times $0 < t_{\ell 1} < \cdots < t_{\ell M_\ell} < t_*$ among subjects in group $\ell, \ell = 1, \ldots, L$. We show in section 1 of the Supplement that the NPMLE is $(\hat{\gamma}, \hat{\lambda})$ with $\hat{\gamma} = n^{-1}(n_1, \ldots, n_{L-1})$ and $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_L)$, where $\hat{\lambda}_\ell$ is given by equation (4) with $\lambda_{\ell\tau m}$ replaced by $\hat{\lambda}_{\ell\tau m} = d_{\ell\tau m}/n_{\ell m}, m = 1, \ldots, M_\ell, \ell = 1, \ldots, L, \tau = 1, 2$. Here $n_{\ell m}$ and $d_{\ell\tau m}$ are the counts of subjects in group ℓ who at time $t_{\ell m}$ are at risk and fail of event τ, respectively. We now estimate θ as $\hat{\theta} = (\hat{\gamma}, \hat{\pi})$, where $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_L)$ with

$$\hat{\pi}_\ell = \varphi_\ell\left(\hat{\lambda}_{\ell l}\right) = \sum\nolimits_{m=1}^{M_\ell} \hat{\lambda}_{\ell 1 m} \prod\nolimits_{m'=1}^{m-1} \left(1 - \hat{\lambda}_{\ell 1 m} - \hat{\lambda}_{\ell 2 m}\right), \ell = 1, \ldots, L$$

and empty products equal one. In the absence of censoring, $\hat{\pi}_l$ reduces to the simple binomial proportion of subjects in group ℓ who develop the outcome during the risk period.

The large sample properties of the $\hat{\pi}_\ell$ were derived by Gill[20] using the theory of counting processes and stochastic integrals (see the summaries in Chapter IV of Andersen et al.[18] and Chapter 5 of Kalbfleisch and Prentice[14]). Specifically, $\sqrt{n}(\hat{\pi}_\ell - \pi_\ell)$ is asymptotically normally distributed with mean 0 and variance Σ_l given by a Taylor series expansion of $\pi_\ell = \varphi_\ell(\lambda_\ell)$ about λ_ℓ of equation (4), as described in section 1 of the Supplement. The resulting variance estimate, $\hat{\Sigma}_\ell$ agrees with that of Dinse and Larson[19] and Gaynor et al.[21]

2.2 Extension to two-stage sampling

The two-stage sampling design as applied to epidemiological data has been described elsewhere[22,23] and we review it briefly in the context of risk model validation. In stage 1, we obtain enough easily available covariate and outcome data to classify subjects into C sampling categories. In stage 2 we independently sample subjects in category c with prespecified Bernoulli probability $p_c > 0$,

$c = 1, \ldots, C$. We then assemble all covariates (both inexpensive and costly), assign model risks, and classify into risk groups only the subjects sampled at stage 2.

We use this design to obtain weighted estimates of the group membership probabilities γ_ℓ and the event-specific hazards λ_ℓ of (4). As described for complete data, we then estimate $\theta = (\gamma, \pi)$ by transforming these estimates to cumulative incidence estimates. Specifically, let Q_c denote the set of subjects in category c, and let $\bar{Q}_c$ denote the subset of subjects sampled in stage 2, with $\hat{p}_c = |\bar{Q}_c|/|Q_c|, c = 1, \ldots, C$. The weighted estimate for (γ, λ) satisfies the Horvitz–Thompson estimating equation $\sum_{i=1}^{n} a_i u_i = 0$, where u_i is the efficient score for the nonparametric likelihood equation described in equation (S.3) of the Supplement, and the weight

$$a_i = \sum_{c=1}^{C} \hat{p}_c^{-1} I\big(i \in \bar{Q}_c\big) \tag{5}$$

is the inverse of the observed sampling probability for those sampled at stage 2 and 0 for other subjects.[14] (In this equation, $I(E)$ is the indicator function, taking value 1 if event E occurs and zero otherwise. It has been shown[24,25] that it is asymptotically more efficient to use the observed sampling probability $\hat{p}_c$ rather than the chosen value p_c.) In clear analogy with the complete sampling case, the solution $(\hat{\gamma}, \hat{\lambda})$ to this estimating equation has components $\hat{\gamma}_\ell = \tilde{n}_\ell / n$ and $\hat{\lambda}_{\ell\tau m} = \tilde{d}_{\ell\tau m}/\tilde{n}_{\ell m}$, where $\tilde{n}_\ell$ is the up-weighted count of subjects in group ℓ, and $\tilde{n}_{\ell m}$ and $\tilde{d}_{\ell\tau m}$ are up-weighted counts of subjects who at time t_m are at risk and fail of type τ, respectively. It is well known that the estimating equation $\sum_{i=1}^{n} a_i u_i = 0$ is unbiased; thus the estimate $\hat{\theta} = (\hat{\gamma}, \hat{\pi})$ obtained by transforming $(\hat{\gamma}, \hat{\lambda})$ is consistent for θ; moreover, the asymptotic covariance matrix $\hat{\theta}$ can be estimated consistently as described in section 1 of the Supplement. While general theory is lacking, empirical evidence and results in special cases[14,18,20,23] suggest that the usual asymptotic normality holds.

The asymptotic covariance matrix of the two-stage estimate $\hat{\theta}$ may be unreliable when the number L of risk groups is large and some of them are sparsely populated. This may be a particular problem with two-stage sampling using small (<20%) second-stage sampling probabilities. In these circumstances, bootstrap covariance estimates provide a practical alternative. Section 1.3 of the Supplement contains specifications for a bootstrap estimate of the covariance matrix of the weighted estimate $\hat{\theta}$.

2.3 Choosing a two-stage design

An important question is how to choose a partition C of C sampling categories and corresponding sampling probabilities $p_1, \ldots, p_C$ to minimize the variance of an estimate $f(\hat{\theta})$ for some key parameter. For example, $f(\hat{\theta})$ might be the model's bias statistic $B(\hat{\theta})$ of equation (1) or its concordance statistic $AUC(\hat{\theta})$ of equation (2). We show in the Supplement (section S.1.2) that the variance of a statistic $f(\hat{\theta})$ obtained from two-stage sampling is the sum of its variance under complete sampling plus a penalty term $P = P(C, p)$ that depends on the parameter (γ, λ), the partition C, and the vector p of sampling probabilities $p_1, \ldots, p_C$. Therefore, minimizing the variance of $f(\hat{\theta})$ is equivalent to minimizing this penalty. To do so, we must:

(a) specify an approximate value for (γ, λ); (b) specify a set of one or more possible partitions C and for each calculate the penalty $P(C, p)$ as a function of the sampling probabilities $p = p_1, \ldots, p_C$; and (c) for each partition C, search in the unit cube of dimension $C-1$ for the sampling probabilities $p_1, \ldots, p_C$ that minimize P, subject to a constraint on the total size of the stage 2 sample.

As noted by Neyman,[17] specifying the parameter (γ, λ) in step (a) presents a dilemma: if we knew this parameter we would not need the study! This catch-22 is similar to the problem of designing a study to maximize the probability of rejecting a null hypothesis about the mean of a distribution,

when the optimal design depends on the value of the unknown mean. Such problems are typically addressed by using external information and available data to approximate the parameters, which are then used to select an approximately optimal design. For the current problem, the investigator has access to the inexpensive covariates and survival data for all subjects sampled in stage 1; only their costly covariates are unknown. Thus, all available relevant information can be used to impute the subjects' costly covariates. This imputation allows assignment of model risks, classification of subjects into the L risk groups, and use of subjects' survival data to specify an approximate value for (γ, λ) as needed for step (a). We illustrate this approach in the application to ovarian cancer in section 4.

For step (b), we know of no formal scheme for choosing partitions likely to yield precise estimates for any given target parameter $f(\theta)$. As illustrated in sections 3 and 4, different target parameters can require different types of partition. For rare outcomes, an intuitively good design is one that samples all the outcome-positive subjects, but only a fraction of the outcome-negative ones, and an even smaller fraction of the less informative outcome-unknown subjects. In other circumstances it may be advantageous to oversample subjects whose available covariates suggest high outcome risk. After determining the best sampling probabilities for each of several target parameters and each of several partitions C, one can then select a partition with acceptably small variances for the parameters of greatest interest.

Step (c) involves choosing sampling probabilities $p_1, \ldots, p_C$ to allow in stage 2 only a fraction k/n of the n stage 1 subjects. Thus, we require $\sum_{c=1}^{C} \hat{w}_c p_c = k/n$, where $\hat{w}_c$ is the proportion of stage 1 subjects in category c. For given values $\hat{w}_c$, this constraint implies that any $C-1$ sampling probabilities determine the remaining one. So, finding the optimal probabilities for a partition C involves searching within the $(C-1)$-dimensional unit cube using constrained minimization software. In the following sections, we used Brent's one-dimensional search[26] for $C=2$ and the simplex method of Nelder and Meade[27] for $C=3$, both implemented with the *R*-routine *constrOPTIM*.

3 Numerical studies

We used simulations and constrained optimization to examine the performance of complete and two-stage sampling for evaluating risk models. In the simulations we generated data for cohorts sampled from two hypothetical populations. For Population 1 the outcome is rare (overall prevalence of 1%), and for Population 2 it is more common (prevalence of 10%). We assumed that potential times to outcome occurrence, death and censoring were independently and exponentially distributed. The risk groups were determined by two covariates x_1 and x_2 taking five discrete values: $(x_1, x_2) = (1,1), (1,2), (1,3), (2,1), (2,2)$. Table 1 (panel A) gives the population distribution across the five risk groups, and the corresponding outcome hazards. We assumed a single competing mortality hazard $\lambda_2 = 0.01$ for all individuals in both populations, and we took the risk period to be the time interval from $t = 0$ to $t = t_* = 1$. The risk-group-specific outcome probabilities shown in Table 1 (panel A) were determined from equation (3) as $\pi_\ell = F_{\ell 1}(1) = [\lambda_{\ell 1}/(\lambda_{\ell 1} + 0.01)]\left[1 - e^{-(\lambda_{\ell 1}+0.01)}\right], \ell = 1, \ldots, 5$.

3.1 Simulations

In each of the 1000 replications, we generated censored times to outcome development or death for a cohort of size $n = 30{,}000$ (Population 1) or $n = 3000$ (Population 2). Times to censoring were exponentially distributed with hazard $\lambda_3 = 0.30$, independent of times to outcome and death. Outcome-positive subjects were those who developed the outcome before the minimum of death,

Table 1. Performance of outcome probability estimates $\hat{\pi}$ for simulated cohorts classified into $L = 5$ risk groups by two covariates x_1, x_2.

Panel A: Population covariate distribution, outcome probabilities, and model-assigned risks, multiplied by 100					
Risk group	1	2	3	4	5
Covariates x_1, x_2	1,1	1,2	1,3	2,1	2,2
Individuals (%)	64	16	10	2	8
		Population 1, outcome prevalence $\pi = 1\%$			
Outcome hazards	0.0035	0.0068	0.0101	0.0368	0.0642
Outcome probabilities π_ℓ (%)	0.35	0.68	1.00	3.59	6.18
Assigned risks (%)					
Well-calibrated model	0.35	0.68	1.00	3.59	6.18
Biased model	0.48	0.74	1.00	3.05	5.10
		Population 2, outcome prevalence $\pi = 10\%$			
Outcome hazards	0.0360	0.0704	0.1059	0.4477	0.9206
Outcome probabilities π_ℓ (%)	3.52	6.75	10.00	35.92	61.84
Assigned risks (%)					
Well-calibrated model	3.52	6.75	10.00	35.92	61.84
Biased model	4.88	7.44	10.00	30.48	50.96
Panel B: Outcome probability estimates $\hat{\pi}$[a] (TSD, ESD)[b], multiplied by 100					
		Population 1, outcome prevalence $\pi = 1\%$, $n = 30{,}000$			
Complete sampling	0.35 (0.046, 0.046)	0.68 (0.128, 0.128)	1.00 (0.196, 0.190)	3.62 (0.823, 0.831)	6.19 (0.531, 0.522)
Two-stage sampling					
Outcome-based[b]	0.35 (0.046, 0.047)	0.67 (0.128, 0.124)	1.00 (0.200, 0.199)	3.59 (0.884, 0.887)	6.16 (0.586, 0.570)
Covariate-based[b]	0.34 (0.108, 0.106)	0.68 (0.301, 0.282)	1.02 (0.464, 0.442)	3.58 (1.340, 1.370)	6.17 (0.868, 0.885)
		Population 2, outcome prevalence $\pi = 10\%$, $n = 3000$			
Complete sampling	3.52 (0.45, 0.46)	6.75 (1.23, 1.23)	10.02 (1.88, 1.82)	36.02 (6.73, 6.78)	61.98 (3.42, 3.47)
Two-stage sampling					
Outcome-based[b]	3.52 (0.47, 0.47)	6.97 (1.52, 1.49)	10.21 (2.49, 2.55)	40.10 (13.64, 15.13)	62.49 (7.61, 7.76)
Covariate-based[b]	3.56 (1.07, 1.02)	6.85 (2.91, 2.89)	10.11 (4.43, 4.24)	36.12 (11.07, 11.37)	62.00 (5.61, 5.49)

[a]Mean of 1000 replicated estimates.

[b]TSD, mean of 1000 theoretical standard deviation estimates, obtained from estimated covariate matrix of $\hat{\theta}$; ESD, square root of empirical variance, obtained as the variance of 1000 replicated estimates.

[c]Using sampling probabilities that minimize $\text{var}\left(\widehat{AUC}\right)$.

censoring, and risk period termination at time $t_* = 1$. The expected proportions of subjects positive, negative, and unknown for outcome development during the period were, respectively, 0.9%, 73.5%, and 25.6% for Population 1; 8.8%, 66.8%, and 24.4% for Population 2.

Table 1 (panel B) shows the means across replications of the estimated outcome probabilities and their theoretical standard deviations (TSDs) for complete and two-stage sampling. The TSDs were obtained from the estimated covariance matrix of $\hat{\theta}$ described in the Supplement. The two-stage sampling included $k/n = 1/5$ of the stage 1 subjects sampled in stage 2. We chose an outcome-based partition with $C = 3$ categories containing subjects whose outcomes were positive, negative, and unknown, and a covariate-based partition with $C = 2$ categories containing subjects whose values for the covariate x_1 were $x_1 = 1$ and $x_1 = 2$ (the rationale for these partitions is described in section 3.2). For each partition, the sampling probabilities were chosen to minimize the variance of the concordance statistic, as described in section 2.3.

Comparison of the risk estimates in Table 1 (panel B) with their true values in Table 1 (panel A) suggests that they are unbiased; moreover the TSDs agree well with the empirical standard deviations (ESDs), i.e. the SDs of the $\hat{\pi}_\ell$ across the 1000 replications. We also computed bootstrap-based SD estimates for both complete and two-stage sampling and found that they agreed well with both the TSDs and the ESDs (data not shown). As expected, the two-stage estimates are less precise than those obtained from complete sampling. The precision loss is relatively mild for the outcome-based partition, but more serious for the covariate-based partition. These results do not differ strongly between the two populations, with one exception. In Population 1 (rare outcome), the risk estimates from outcome-based sampling are more precise in all risk groups, while in Population 2 (more common outcome), they are less precise than those of covariate-based sampling in the highest two risk groups. This difference suggests that although outcome-based sampling will generally perform well for rare outcomes (prevalence $<10\%$), its performance for more common outcomes is less predictable.

3.2 Constrained optimization

The simulations address how well the risk-group-specific outcome probabilities are estimated with complete and two-stage sampling. However, summary performance measures such as model bias and concordance involve not only the risk-group-specific outcome probabilities π_ℓ, but also the risk-group membership probabilities γ_ℓ, and there is need to evaluate the performance of two-stage designs in estimating these measures. Our objective now is to examine and compare the optimal sampling probabilities and corresponding true SDs of estimated performance measures across different design choices and outcome prevalences when the model generating the data is known. Accordingly, we used the populations and generating models of Table 1 (panel A) to obtain the optimal sampling probabilities and calculate the corresponding true SDs for the bias and concordance statistics, using the methods described in section 1.4 of the Supplement. In practice, the true model is unknown and must be approximated using the covariate and survival data available for stage 1 subjects, as we illustrate with CTS ovarian cancer data in section 4.

We consider the SD of bias and concordance statistics obtained from a cohort of size 30,000 sampled from Population 1 (outcome prevalence of 1%) and a cohort of size 3000 sampled from Population 2 (outcome prevalence of 10%). We again consider the outcome- and covariate-based partitions shown in Table 1 (panel B). We chose the outcome-based partition to investigate the relative contribution to performance estimates of subjects with the three types of outcome. We chose the covariate-based partition to investigate the circumstance when one covariate (x_1) is available for all cohort subjects but another (x_2) must be ascertained. Table 1 (panel A) shows that subjects with

Table 2. Two-stage designs, optimal sampling probabilities and corresponding SDs of bias and concordance statistics.

Panel A: Two-stage sampling probabilities

	$C=3$ Outcome-based categories			$C=2$ Covariate-based categories	
	Outcome positive	Outcome negative	Outcome unknown	$x_1=1$	$x_1=2$
	Population 1, $\pi=1\%$, $n=30{,}000$				
Variance minimized [a]					
var($\hat{B}$)	1.00	0.22	0.12	0.11	1.00
var($\widehat{AUC}$)	1.00	0.21	0.15	0.19	0.31
	Population 2, $\pi=10\%$, $n=3000$				
Variance minimized					
var($\hat{B}$)	0.52	0.18	0.13	0.11	1.00
var($\widehat{AUC}$)	0.82	0.15	0.12	0.19	0.30

Panel B: SD of bias ($\hat{B}$) and concordance ($\widehat{AUC}$) statistics

		SD($\hat{B}$)	SD($\widehat{AUC}$)
Population 1, $\pi=1\%$, $B=0.0034$, AUC $=0.778$			
Complete sampling, $n=30{,}000$		0.0014	0.016
Two-stage sampling[a]	Variance minimized		
Outcome-based categories	var($\hat{B}$)	0.0015	0.021
	var($\widehat{AUC}$)	0.0015	0.017
Covariate-based categories	var($\hat{B}$)	0.0014	0.044
	var($\widehat{AUC}$)	0.0025	0.036
Population 2, $\pi=10\%$, $B=0.0336$, AUC $=0.805$			
Complete sampling, $n=3000$		0.0087	0.016
Two-stage sampling[a]	Variance minimized		
Outcome-based categories	var($\hat{B}$)	0.0160	0.035
	var($\widehat{AUC}$)	0.0166	0.026
Covariate-based categories	var($\hat{B}$)	0.0094	0.057
	var($\widehat{AUC}$)	0.0159	0.035

Note: [a]With 20% of stage 1 subjects sampled at stage 2.

$x_1=2$ comprise the highest two risk groups and only 10% of the population; therefore, we wanted to determine the relative contribution of these subjects to precision measures.

Table 2 (panel A) shows the optimal two-stage sampling probabilities for these two partitions and Table 2(panel B) shows the corresponding SDs of the bias statistic $\hat{B}$ and concordance statistic $\widehat{AUC}$. For comparison, Table 2 (panel B) also includes the SDs corresponding to complete sampling. The results in Table 2 (panel B) for the optimal outcome-based designs indicate that it is best to sample all outcome-positive subjects for rare outcomes but not for more common ones. In the latter case, however, the precision loss from sampling all outcome-positive subjects is not great – for example, SD($\hat{B}$) increased from 0.016 to 0.017, an increase of 6%, and SD($\widehat{AUC}$) increased from 0.0256 to 0.0260, an increase of 1% (data not shown). As expected, outcome-negative subjects are sampled

more heavily than censored ones, but the optimal sampling ratio depends on outcome prevalence. In contrast, results for the two covariate-based designs do not vary strongly with outcome prevalence. Comparison across the outcome-based and covariate-based partitions shows that the former gives more precise estimates of both performance measures for rare outcomes, but not for more common ones. Specifically, when outcome prevalence is 10%, model bias is estimated more precisely by the best covariate-based partition, which oversamples subjects with $x_1 = 2$, that is, subjects in the sparsely populated highest two risk groups.

We conclude this section with two observations about complete sampling. First, as shown in Table 2 (panel B), the SD of the concordance statistic is the same for a cohort from Population 1 of size 30,000 and a cohort from Population 2 of 3000. This suggests that the precision of this statistic depends more on the number of outcome-positive subjects (which is roughly 270 for both types of cohort) than on the total number of subjects. Second, as expected, other simulations (data not shown) revealed severe upward bias for outcome probability estimates obtained by excluding censored subjects and performing a standard binomial-based analysis only on the remaining outcome-positive and outcome-negative subjects.

4 Example

We illustrate the complete and two-stage sampling methods by using them to assess the performance of ovarian cancer risk models applied to subjects in the CTS described in section 1. We first assess the accuracy and discrimination of a nongenetic model (Model 1) in the entire cohort, and we use this complete data assessment as the gold standard for results using the two-stage design. We then show how available covariates, outcome data, and Model 1 risks can be used to find two-stage designs that minimize the variance of estimated gains in sensitivity and specificity associated with an expanded model (Model 2) involving additional costly genetic covariates.

4.1 CTS cohort subjects

Among all 133,429 subjects, $n = 40{,}139$ met the eligibility criteria described in section 3 of the Supplement and comprise the cohort. We used the nongenetic risk model described in the Supplement (hereafter called Model 1) to assign each subject a 12-year ovarian cancer risk. We took the follow-up time for each subject to be the number of days between cohort entry and the first occurrence of ovarian cancer diagnosis, death, last observation, or 12 years of follow-up. Among eligible subjects, 227 (0.6%) developed invasive epithelial ovarian cancer within 12 years of cohort entry (outcome-positive subjects) and 26,887 (67%) died from other causes or survived the period without ovarian cancer (outcome-negative subjects). An additional 13,025 subjects (32%) were alive and free of ovarian cancer at last observation, but had not accrued the full 12 years of follow-up. These subjects were classified as outcome unknown. Since most of them were observed for at least 10 years of the 12-year follow-up period, their inclusion in the analysis substantially reduces the estimated outcome probabilities.

4.2 Assessment of Model 1

Figure 1 shows two plots of points $(\bar{r}_\ell, \hat{\pi}_\ell)$ corresponding to $L = 5$ estimated quintiles of Model 1 risk. Here $\bar{r}_\ell$ is the mean assigned risk and $\hat{\pi}_\ell$ is the estimated outcome probability among subjects in quintile $\ell, \ell = 1, \ldots, 5$. The points in the upper panel (Figure 1A) were estimated from the complete sample of all $n = 40{,}139$ subjects, and those in the lower panel (Figure 1B) from two-

stage sampling with $C = 2$ sampling categories: (1) all 227 ovarian cancer cases, sampled with probability $p_1 = 1$, and (2) all other subjects, sampled with probability $p_2 = 0.1$. As evident in the figure, the two designs yield nearly identical estimates for the outcome probabilities and their SDs. We also found good agreement between the theoretical and bootstrap SD estimates (data not shown). There was evidence of model bias (Hosmer–Lemeshow statistic $\chi^2_5 = 11.7$ ($P = 0.04$) for complete sampling and $\chi^2_5 = 14.8$ ($P = 0.01$) for the two-stage sampling). Model discrimination was poor: the estimated AUCs were 0.59 (0.55–0.62) and 0.58 (0.54–0.62) for complete and two-stage sampling, respectively. The poor discrimination reinforces other observations[28] that existing markers for increased ovarian cancer risk cannot adequately discriminate those who will develop the disease from those who will not.

4.3 Choice of two-stage design for assessing an expanded model

Model 1 involves only existing CTS covariates and not costly biospecimen analysis. Of clinical interest is the discrimination gained from an expanded model that also includes genotypes for ovarian cancer susceptibility alleles. However, cost issues prevent genotyping the entire cohort and a two-stage design is needed. Here, we illustrate how the available covariates and survival data of the 40,139 stage 1 subjects, together with external information, can be used to partition these subjects into sampling categories and select sampling probabilities that minimize the variance of a parameter estimate of interest. For example, we have seen that the discrimination of Model 1 is poor, and we may want to estimate the improvement gained by augmenting Model 1 with subjects' carrier statuses for rare pathogenic mutations of the breast/ovarian cancer susceptibility genes BRCA1 and BRCA2 in an expanded Model 2, such as the one described in the Supplement. Moreover, suppose budgetary constraints limit the genetic testing to 7000 of the 40,139 subjects. The question is: how to choose sampling categories and stage 2 sampling probabilities to estimate the discriminatory gain for Model 2, using the 7000 stage 2 subjects who are genotyped and assigned risks according to both models?

We address this question by determining two-stage designs that minimize the variances of estimates for the gains in sensitivity and specificity associated with Model 2. We took a model's sensitivity to be the probability that it designates an outcome-positive subject as high risk (12-year risk $\geq 3\%$), and its specificity to be the probability that it designates an outcome-negative subject as low risk (12-year risk $<3\%$). We then classified the women in $L = 4$ risk groups indexed as j,k, $j = 1,2$, $k = 1,2$, as determined by their risk statuses (1 = low risk, 2 = high risk) according to Models 1 and 2, respectively. The Supplement contains expressions for the sensitivity and specificity gains for Model 2, as functions of $\theta = (\gamma, \pi)$, with $\gamma = (\gamma_{11}, \gamma_{12}, \gamma_{21})$ and $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$.

As noted in section 2.3, finding an optimal design involves three steps: (a) choosing an approximate value for θ; (b) choosing partition(s) C and calculating for each the penalty term $P(C, p)$ of equation (S.11) as a function of p; and (c) searching for the sampling probabilities p that minimize P. To accomplish step (a), we used each stage 1 subject's available covariates to impute her BRCA1 and BRCA2 carrier statuses. Specifically, we used her personal and family cancer history in the software BRCAPRO[29] to assign her probabilities of carrying a mutation of BRCA1 and BRCA2, and then used these probabilities to randomly assign her a carrier status for each gene. We then used her BRCA1 and BRCA2 carrier statuses and available covariates to assign her a Model 2 risk and, coupled with her Model 1 risk, classified her in one of the four risk groups. Table 3 shows the ovarian cancer risk distribution resulting from this classification. The table also shows the approximate outcome probabilities π_{jk} obtained from the event-specific survival data of the stage 1 subjects in each risk group. These results provided approximate parameter values for

Table 3. Distribution of 40,139 CTS subjects according to ovarian cancer risks assigned by Models 1 and 2.

	Model 2 risk [a]								
	Low (<3%)			High (≥3%)			Total		
Model 1 risk[a]	$n\hat{\gamma}$	$\hat{\pi}^a$	SD	$n\hat{\gamma}$	$\hat{\pi}$	SD	$n\hat{\gamma}$	$\hat{\pi}$	SD
Low (<3%)	39,475	0.6	0.04	74	22.6	5.15	39,549	0.6	0.04
High (≥3%)	94	1.2	1.18	496	0.8	0.42	590	0.9	0.40
Total	39,569	0.6	0.04	570	3.6	0.81	40,139	0.6	0.04

CTS: California Teachers Study.
Note: [a]Probability of developing ovarian cancer within 12 years of baseline, multiplied by 100.

step (a). For these parameters, Model 1 sensitivity and specificity were, respectively, 2.2% and 98.5%, with corresponding values for Model 2 of 8.4% and 98.6%. These values give a sensitivity gain of 6.2% and a small specificity gain of 0.1% for Model 2 compared with Model 1.

In step (b) we selected two partitions. The first is outcome-based with $C = 3$ sampling categories (positive, negative, and unknown ovarian cancer status). The second is based on both outcome and nongenetic covariates, with $C = 3$ categories: outcome-positive, other subjects in the highest Model 1 risk quintile, and all other subjects. For each of these two partitions and each of the two target measures (sensitivity gain and specificity gain), we determined the penalty P as a function of the sampling probabilities p, as described in section 1.3 of the Supplement. In step (c) we determined the sampling probabilities that minimize P.

Table 4 shows the optimal sampling probabilities and corresponding SDs for each target measure and each partition. These results illustrate several points. First, the two partitions differ little in their optimal performances, with a slight edge for the second partition compared with the first. Second, regardless of the partition chosen, the designs optimal for sensitivity gain include all outcome-positive subjects, while the ones optimal for specificity gain include only 6–7% of these subjects. Thus, even for outcomes of low prevalence, some parameters may be estimated more efficiently by sampling only a fraction of the outcome-positive subjects. However, because we typically want precise estimates for gains in both sensitivity and specificity, and for rare outcomes the SDs of sensitivity gain estimates substantially exceed the SDs of specificity gain estimates, the designs that optimize sensitivity at the expense of slight suboptimality for specificity will tend to dominate design decisions.

5 Discussion

We have used cohort data to evaluate the accuracy and discrimination of personal risk models in the presence of censoring due to incomplete follow-up. In so doing, we have adopted the common practice of classifying subjects into discrete groups determined by assigned risk, and estimating the actual outcome probabilities within these groups. Such discretization facilitates assessment of model accuracy using the bias statistic and analogous Hosmer–Lemeshow chi-squared test statistic. Moreover, using quintiles of assigned risk as the risk groups avoids groups with too few subjects for adequate assessment. However, these advantages are offset by the costs of clumping individual risks into broad groups, with consequent loss of sharpness in model assessment. Alternatives to the quintile approach based on nearest-neighbor methods have been proposed,[30,31] but their properties need further evaluation in simulations and data. In the absence of censoring, the

Table 4. Optimal sampling probabilities and standard deviations of estimated gains in sensitivity and specificity from expanding an ovarian cancer model with genetic covariates.

Panel A: Optimal sampling probabilities

	Outcome-based categories			Outcome/risk-based categories		
Variance minimized	Positive	Negative	Unknown	Outcome positive	Others, Model 1 quintile 5	Others, Model 1 quintiles 1–4
var($\hat{\Delta}_{SN}$)	1.00	0.22	0.14	1.00	0.24	0.15
var($\hat{\Delta}_{SP}$)	0.06	0.19	0.15	0.06	0.41	0.12

Panel B: Standard deviations of estimated performance gains for Model 2

		SD($\hat{\Delta}_{SN}$)	SD($\hat{\Delta}_{SP}$)
Complete sampling		0.0170	0.0034
Two-stage sampling	Variance minimized		
Outcome-based categories	var($\hat{\Delta}_{SN}$)	0.0177	0.0077
	var($\hat{\Delta}_{SP}$)	0.0700	0.0072
Outcome/risk-based categories	var($\hat{\Delta}_{SN}$)	0.0177	0.0066
	var($\hat{\Delta}_{SP}$)	0.0638	0.0060

SD: standard deviation; $\hat{\Delta}_{SN}$: estimated gains in sensitivity; $\hat{\Delta}_{SP}$: estimated gains in specificity.

maximum-likelihood estimates of the discrete outcome probabilities are simple binomial proportions of subjects who develop the outcome during the risk period. With censoring, however, this approach leads to biased risk estimates, because it fails to account for all subjects' times at risk, and competing risk analysis is needed. For this, Dinse and Larson[19] argue against direct estimation of the desired outcome probabilities, but instead recommend estimating the event-specific hazards and converting these hazard estimates to outcome probability estimates.

We have proposed a two-stage sampling design for using cohort data to assess the performances of externally derived risk models in circumstances when assigning risks to all subjects is infeasible. We also describe methods for using the resulting data to efficiently estimate relevant performance parameters. When the sampling categories are based on subjects' outcomes, the two-stage design is similar in spirit to the case–cohort design of Prentice[32] for regression analysis of survival data under a Cox proportional hazards model. If used to evaluate risk models, the case–cohort design would assign risks to all outcome-positive subjects and to a randomly selected subset of the entire cohort. This design yields outcome probability estimates identical to those of the two-stage design with sampling categories consisting of: (a) outcome-positive subjects; and (b) all other subjects. However, the variances of the two estimates differ, because those from the case–cohort design must accommodate any overlap between the sampled subcohort and the set of outcome-positive subjects. Moreover, the two-stage design is more flexible for evaluating risk models as it allows stage 2 sampling from an arbitrary partition of subjects into sampling categories.

We have provided consistent estimates for the theoretical covariance matrices of the weighted estimates, and we have proposed corresponding bootstrap covariance estimates. General asymptotic theory is lacking but the empirical evidence suggests that the usual normality results hold.

The proposed theoretical and bootstrap dispersion estimates both agreed well with the empirical ones used as the gold standard. The simulations, numerical calculations, and application to data from the CTS illustrate several points. First, determining optimal design choices requires estimating performance gains associated with the costly covariates, which are unobserved at completion of the first sampling stage. Thus, one needs reasonable estimates of the subjects' values for these covariates, and the relative optimality of different design choices depends on the accuracy of these estimates. Second, the design with optimal efficiency for one parameter need not be so for other parameters of interest, and trade-offs will be required. For example, optimal precision for some parameter estimates may require sampling only a fraction of outcome-positive subjects, even with uncommon outcomes. As a general rule, however, sampling all outcome-positive subjects and more outcome-negative than censored subjects can be expected to perform reasonably well in most circumstances.

An R program entitled 'Risk Model Assessment Program (RMAP)', which provides estimates for outcome probabilities, the associated covariance matrices and graphical examination of model performance, is freely available at <http://www.stanford.edu/ggong/rmap/index.html>.

Acknowledgements

The authors thank Gail Gong and David Johnston for developing a user-friendly R program that implements the methods, Joseph Keller and Nicole Ng for help with computations, Pamela Horn-Ross, Valerie McGuire, Anna Felberg, Alison Canchola, Leslie Bernstein and James V. Lacey Jr for help with the CTS data, and the CTS Steering Committee responsible for establishing and maintaining the CTS cohort.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the National Institutes of Health (grants R01 CA09045, R03 CA150136, R01 CA77398) and the California Breast Cancer Research Fund (contract 97-10500).

References

1. Costantino JP, Gail MH, Pee D, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst* 1999; **91**: 1541–1548.
2. Shariat SF, Karakiewicz PI, Roehrborn CG, et al. An updated catalog of prostate cancer predictive tools. *Cancer* 2008; **113**: 3075–3099.
3. Benichou J and Gail MH. Estimates of absolute cause-specific risk in cohort studies. *Biometrics* 1990; **46**: 813–826.
4. Gail MH and Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics* 2005; **6**: 227–239.
5. Janes H, Pepe MS and Gu W. Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med* 2008; **149**: 751–760.
6. Whittemore AS. Evaluating health risk models. *Stat Med* 2010; **29**: 2438–2452.
7. Bernstein L, Allen M, Anton-Culver H, et al. High breast cancer incidence rates among California teachers: results from the California Teachers Study (United States). *Cancer Causes Control* 2002; **13**: 625–635.
8. Chang ET, Lee VS, Canchola AJ, et al. Dietary patterns and risk of ovarian cancer in the California Teachers Study cohort. *Nutr Cancer* 2008; **60**: 285–291.
9. Hsu W and Murphy AH. The attributes diagram. A geometrical framework for assessing the quality of probability forecasts. *Int J Forecast* 1986; **2**: 285–293.
10. Hosmer DW and Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Commun Stat Theory Methods* 1980; **9**: 1043–1069.

11. Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**: 29–36.
12. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press, 2003.
13. Rockhill B, Speigelman D, Byrne C, et al Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. *J Natl Cancer Inst* 2001; **93**: 358–366.
14. Kalbfleisch JD and Prentice RL. *The statistical analysis of failure time data*, 2nd ed. New York, NY: Wiley, 2002.
15. Choudhury JB. Nonparametric confidence interval estimation for competing risks analysis: application to contraceptive data. *Stat Med* 2002; **21**: 1129–1144.
16. Putter H, Fiocco M and Geskus RB. Tutorial in biostatistics: competing risks and multistate models. *Stat Med* 2007; **26**: 2389–2430.
17. Neyman J. Contribution to the theory of sampling human populations. *J Am Stat Assoc* 1938; **33**: 101–116.
18. Andersen PK, Borgan O, Gill RD, et al. *Statistical models based on counting processes*. New York, NY: Springer Verlag, 1992.
19. Dinse GE and Larson MG. A note on semi-Markov models for partially censored data. *Biometrika* 1986; **73**: 379–386.
20. Gill RD. Nonparametric estimation based on censored observations of a Markov renewal process. *Zeitschrift ftir Wahrscheinlichkeitstheorie und verwandte Gebiete* 1980; **53**: 97–116.
21. Gaynor JJ, Feuer EJ, Tan CC, et al. On the use of event-specific event and conditional event probabilities: examples from clinical oncology data. *J Am Stat Assoc* 1993; **88**: 400–409.
22. Kalbfleisch JD and Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. *Stat Med* 1988; **7**: 140–160.
23. Whittemore AS. Multistage sampling designs and estimating equations. J R Stat Soc *Ser B, Stat Methodol* 1997; **59**: 589–602.
24. Robins JM, Rotnitzky A and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc* 1994; **89**: 846–866.
25. Robins JM, Rotnitzky A and Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995; **90**: 106–121.
26. Brent R. *Algorithms for minimization without derivatives*. New Jersey: Prentice-Hall, 1973. Reprinted by New York, NY: Dover, 2002.
27. Nelder JA and Meade R. A simplex method for function minimization. *Comput J* 1965; **7**: 308–313.
28. Palmer C, Duan X, Hawley S, et al. Systematic evaluation of candidate blood markers for detecting ovarian cancer. *PLoS One* 2008; **3**: e2633.
29. Parmigiani G, Berry D and Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *Am J Hum Genet* 1998; **62**: 145–158.
30. Akritas MG. Nearest neighbor estimation of a bivariate distribution under random censoring. *Ann Stat* 1994; **22**: 1299–1327.
31. Saha P and Heagerty PJ. Time-dependent predictive accuracy in the presence of competing risks. *Biometrics* 2010; **66**: 999–1011.
32. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73**: 1–11.