Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–12

DOI: 10.1177/0962280217723945
journals.sagepub.com/home/smm
SAGE

Evaluating disease prediction models using a cohort whose covariate distribution differs from that of the target population

Scott Powers,[1] Valerie McGuire,[2] Leslie Bernstein,[3] Alison J Canchola[4] and Alice S Whittemore[2]

Abstract

Personal predictive models for disease development play important roles in chronic disease prevention. The performance of these models is evaluated by applying them to the baseline covariates of participants in external cohort studies, with model predictions compared to subjects' subsequent disease incidence. However, the covariate distribution among participants in a validation cohort may differ from that of the population for which the model will be used. Since estimates of predictive model performance depend on the distribution of covariates among the subjects to which it is applied, such differences can cause misleading estimates of model performance in the target population. We propose a method for addressing this problem by weighting the cohort subjects to make their covariate distribution better match that of the target population. Simulations show that the method provides accurate estimates of model performance in the target population, while un-weighted estimates may not. We illustrate the method by applying it to evaluate an ovarian cancer prediction model targeted to US women, using cohort data from participants in the California Teachers Study. The methods can be implemented using open-source code for public use as the R-package RMAP (Risk Model Assessment Package) available at http://stanford.edu/~ggong/rmap/.

Keywords

Cohort selection bias, calibration, concordance, personal predictive model, weighted-as-needed

1 Introduction

Personal predictive models for future adverse health outcomes provide important tools in the practice of preventive medicine. Such models use an individual's personal covariates to assign him/her a probability of developing an adverse outcome within a specified future time period. Accurate predictions are needed for rational decisions about preventive strategies, and for allocating preventive efforts to those who need them most. The models are often evaluated by comparing their assigned risks to outcome incidence among participants in longitudinal cohort studies. Specifically, the prediction model is used to assign, say, ten-year risks to the subjects based on the covariates they report at cohort entry, and these assigned risks are then compared to the subjects' outcome incidence during the following ten years.

However, the covariates of participants in large cohort studies may differ from those of the population for which the predictive model is targeted. Since estimates of predictive model performance can vary with the covariate distribution of the subjects,[1–4] these differences can cause misleading estimates of model performance in the target population.

[1]Department of Statistics, Stanford University, Stanford, CA, USA
[2]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA
[3]Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA, USA
[4]Cancer Prevention Institute of California, Fremont, CA, USA

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

There is a large literature on methods for checking the external validity of prediction models and on modifying them when they perform poorly in a population of interest.[3–6] These papers tend to focus on prediction models for outcomes occurring at or shortly after risk assignment. In contrast, here we are concerned with long-term outcomes such as breast cancer development within the next ten years, so that model validation requires longitudinal cohort data. This difference has three consequences: First, many of the sampled subjects decline to participate in such long, time-consuming cohort studies, so the covariates of those who do participate may be selected. Second, because cohort studies are costly they are sparse. For instance, the ovarian cancer prediction model we discuss is targeted to the entire adult female US population, but currently there are no long-term cohort data involving a population-based sample of this population. Third, the populations (e.g. female nurses, female teachers) from which the cohort subjects are sampled often differ from the population for whom the model is intended. Findings of disease association in these cohorts are nevertheless valuable because of the ''relative-risk robustness'' assumption (e.g. if smoking is found to double bladder cancer risk among nurses, it's likely to do so among all women). However, because prediction model performance depends on the covariate distribution of the population to which the model is applied,[4–6] this robustness cannot guarantee that a model's performance in an external validation cohort indicates how the model would perform in a target population with a different covariate distribution.

Here, we propose a weighting method for using cohort subjects to evaluate how a predictive model would perform if applied to a target population. The method involves the following steps. First, we supplement the cohort subjects' data with cross-sectional covariate data from an additional sample of subjects whose covariate distribution represents that of the target population. Second, we use the model to assign risks to the subjects in both cohort and population-based samples. Third, we classify both samples of subjects into joint covariate categories and then weight the cohort subjects to make their covariate distribution more closely resemble that of the target population. Finally, we evaluate model performance using the cohort's weighted distributions of assigned risks and outcomes.

The efficacy of the proposed weighting method requires that: (i) the relative risk robustness assumption holds for all associations relating outcome to model covariates; and (ii) all covariate combinations present in the target population also be represented in the cohort. If, for example, the target population contains older women with a history of oral contraceptive use but the cohort lacks such women, then model performance in this subgroup cannot be evaluated with the cohort data.

The weighted approach proposed here is analogous to other methods that use subject-specific weights to adjust for selection bias in observational studies. These include the inverse-probability-weighting method[7,8] and, for retrospective case-control studies, the bias-breaking-variable method.[9] However, the latter two methods use weights to reduce bias in estimates of regression coefficients relating risk factors to the outcome, while here the weighting is performed to reduce bias in estimates of predictive model performance. Thus while all three methods share the common objective of weighting study participants to make the distribution of their attributes better represent those of a particular target population, the specific goals and construction of weights differ.

We use simulations to show that weighting can change inferences about predictive model performance. In particular, it can increase power to detect poor model calibration to the target population, and it can provide more accurate estimates of model discrimination. We illustrate the method by applying it to evaluate an ovarian cancer prediction model targeted to the general US population, as applied to participants in the California Teachers Study (CTS).

2 Methods

We begin by reviewing two criteria for predictive model performance: model calibration and model discrimination. We then describe the weighted method for evaluating a model's calibration and discrimination as applied to the target population.

2.1 Performance measures

Predictive model performance is evaluated using two criteria: calibration and discrimination.[10,11] *Calibration* is the extent of agreement between model-assigned risks and observed outcome incidence. This can be done visually by partitioning the cohort into subgroups at different risk, and comparing observed to predicted outcome prevalence in the subgroups. If needed, this comparison can be formalized by testing the null hypothesis where a model is well-calibrated to the population underlying a given cohort sample. To implement the test we partition its subjects

into $L \geq 1$ subgroups $Q_1, \ldots, Q_L$ according to assigned risks or covariate values, and then assess the sum of squared and standardized differences between the weighted mean assigned risks $\bar{R}_\ell = \sum_{i \in Q_\ell} w_{Ci} r_i / \sum_{i \in Q_\ell} w_{Ci}$ and the weighted mean outcome incidence estimates $\hat{\pi}_\ell$

$$X_L^2 = \sum_{\ell=1}^{L} \hat{\gamma}_\ell \frac{\left(\hat{\pi}_\ell - \bar{R}_\ell\right)^2}{\widehat{\text{var}}(\hat{\pi}_\ell)} \tag{1}$$

Here, $\hat{\gamma}_\ell = N_C^{-1} \sum_{i \in Q_\ell} w_{Ci}$ is the weighted proportion cohort subjects in subgroup ℓ, and we obtain the cumulative incidence estimates $\hat{\pi}_\ell$ by applying standard survival data methods to the weighted subjects, as described in the Appendix. We used the bootstrap with 1000 bootstrap replications to estimate the subgroup-specific variances $\text{var}(\hat{\pi}_\ell)$, $\ell = 1, \ldots, L$. Given the proportions $\hat{\gamma}_1, \ldots, \hat{\gamma}_L$, the goodness-of-fit (GOF) statistic X_L^2 is asymptotically equivalent to a quadratic form in L Gaussian variables, whose null distribution is that of a mixture of central chi-squared distributions: $X_L^2 \sim \sum_{\ell=1}^{L} \hat{\gamma}_\ell \chi_{\ell 1}^2$. Several approximations to this distribution have been proposed.[12] Here, we use the exact method of Davies,[13] which has been found to perform well and can be implemented in the R-package CompQuadForm.

A predictive model's *discrimination* describes its ability to distinguish outcome-positive from outcome negative subjects (i.e. those who do and do not develop the outcome in the specified time period). The most commonly used discrimination measure for a predictive model is its *concordance*, defined as the probability that the risk assigned to an outcome-positive person exceeds that assigned to an outcome-negative one

$$AUC = \Pr\left(R_i > R_j | Y_i = 1 \ \& \ Y_j = 0\right) \tag{2}$$

Here $Y_i = 1$ if the ith subject is outcome-positive and $Y_i = 0$ if he/she is outcome negative. This measure is also called the area under the receiver operating characteristic (ROC) curve, or AUC.[10] To estimate AUC and its 95% confidence interval, we assume that subjects' censoring times are independent of both their assigned risks and their outcome times, and use the weighted version of the estimator proposed by Hung and Chiang[14] and Blanche et al.[15] as described in the Appendix.

2.2 The weighting method

To implement the method, we obtain an additional population-based sample of model covariates for N_P subjects from the target population, and use them in the model to assign risks to all subjects in this sample for comparison with the risks assigned to the cohort subjects. Then, we classify the two samples according to a common set of J joint covariate categories, and assign cohort subject i the weight

$$w_{Ci} = \sum_{j=1}^{J} \frac{\hat{\varphi}_{Pj}}{\hat{\varphi}_{Cj}} 1(i \in cat\ j), \quad i = 1, \ldots, N_C \tag{3}$$

Here, $\hat{\varphi}_{Cj}$ and $\hat{\varphi}_{Pj}$ denote the proportions of cohort and population subjects, respectively, with covariates in category j, $j = 1, \ldots, J$, and $1(E)$ is the indicator function assuming the value 1 if event E is true and zero otherwise. Note that: (a) the sum $\sum_{i=1}^{N_C} w_{Ci}$ of the weights (1) over all cohort subjects is their total count N_C; (b) the standard (unweighted) analyses correspond to weights $w_{Ci} = 1$ for all cohort subjects; and (c) a weighted analysis assigns weights $w_{Ci} < 1$ to subjects with overrepresented covariates, and weights $w_{Ci} > 1$ to those with underrepresented covariates, in the cohort compared to the population. Finally, we use the weights (3) to obtain weighted tests of calibration and weighted estimates of concordance. The weighted GOF tests X_L^2 involve subgroup-specific weighted mean assigned risk and weighted outcome incidence estimates described in the Appendix.

3 Simulation

We used simulated data to evaluate the proposed weighting strategy for evaluating predictive model performance using cohort subjects whose covariate distribution differs from that of the target population. Specifically, we generated and analyzed data in each of $E = 1200$ replications. In each replication, we generated covariate data for subjects in three cohorts and one population-based sample, and used the covariates of cohort subjects to generate censored outcome data. We then analyzed the data by using the subjects' covariates in hypothetical predictive models to assign them outcome risks, and if cohort and population risk distributions differed

significantly, we calculated the weighted GOF tests and weighted concordance estimates described in the Appendix. For comparison, we also evaluated these measures without using weighting. Finally, we used summary statistics averaged over the E replications to assess the impact of cohort selection bias on model performance with and without using weighting.

3.1 Data generation

In each replication, we generated cross-sectional covariate and corresponding censored outcome data for three cohort samples, labeled C_1–C_3, each containing $N_C = 10{,}000$ subjects, and cross-sectional covariate data for subjects in a population-based sample P containing $N_P = 5000$ subjects. We assumed that subjects' outcome probabilities depend on their values for a covariate vector $x = (x_1, x_2)$, whose population distribution is Gaussian with mean and variance

$$\mu = (-2.50, 0.50) \quad \text{and} \quad V = \text{diag}\left(\sigma_1^2, \sigma_2^2\right) = \text{diag}(0.640, 0.562) \tag{4}$$

For cohort C_1, we oversampled subjects with small values of x_1 (biased sampling). To do so, for each subject we: (a) randomly chose a vector $\mathrm{x} = (x_1, x_2)$ from the Gaussian distribution (4); (b) classified his/her standardized covariate vector $z = (z_1, z_2) = [(x_1 - \mu_1)/\sigma_1, (x_2 - \mu_2)/\sigma_2]$ into one of the four joint covariate categories shown in Table 1; and (c) sampled the subject with the category-specific selection probability shown in Table 1. We then repeated steps (a) to (c) until 10,000 subjects were selected. We generated covariates for subjects in cohort C_2 similarly, but now we oversampled subjects with small values of x_2. We used simple random sampling to generate covariates for subjects in cohort C_3 and the population-based sample P. (We included the unbiased cohort C_3 to assess the effects of weighting subjects when cohort and population risk distributions differed only by chance.)

We then used each cohort subject's covariate vector x to generate times t_0 and t_1 to censoring and outcome, respectively, according to independent exponential density functions of the form

$$f_\tau(t; x) = \lambda_\tau(x) e^{-\lambda_\tau(x)t}, \quad \tau = 0, 1, \quad \text{with} \quad \lambda_0(x) \equiv \lambda_0 = 0.056 \quad \text{and} \quad \lambda_1(x) = e^{x_1 + x_2} \tag{5}$$

In the absence of censoring, the density $f_1(t; x)$ gives the probability of outcome occurrence by time $t^* = 1$ for an individual with covariates x as $P(x) = 1 - \exp(-e^{x_1 + x_2})$. We recorded each subject's survival data as (t,y), where $t = \min(t_0, t_1, t^*)$, and the outcome status indicator $y = 1$ if $t = t_1$ (outcome-positive), $y = 0$ if $t = t^*$ (outcome-negative) with y unknown if $t = t_0$ (outcome-censored).

3.2 Data analysis

In each replication, we applied two hypothetical predictive models to the covariate data for cohort and population subjects. Model A assigns a subject with covariates x the actual outcome probability $R_A(x) = P(x) = 1 - \exp(-e^{x_1 + x_2})$ used to generate his/her uncensored survival data. In contrast, Model B assigns this subject an

Table 1. Distribution of covariates $x = (x_1, x_2)$ in $J = 4$ categories for subjects from three cohorts.

Covariate category		Covariate sampling scheme								
		Cohort 1. Oversampling small values of x_1			Cohort 2. Oversampling small values of x_2			Cohort 3. Unbiased sampling		
$\frac{x_1-\mu_1}{\sigma_1}$	$\frac{x_2-\mu_2}{\sigma_2}$	Sampling probability	Proportion of subjects	Weight[a]	Sampling probability	Proportion of subjects	Weight	Sampling probability	Proportion of subjects	Weight
<0	<0	1.00	0.4762	0.525	1.00	0.4762	0.525	1.00	0.25	1.00
<0	≥ 0	1.00	0.4762	0.525	0.05	0.0238	10.50	1.00	0.25	1.00
≥ 0	<0	0.05	0.0238	10.50	1.00	0.4762	0.525	1.00	0.25	1.00
≥ 0	≥ 0	0.05	0.0238	10.50	0.05	0.0238	10.50	1.00	0.25	1.00

[a]Weight for cohort subjects in category j is $\hat{\varphi}_{Pj}/\hat{\varphi}_{Cj}$, where $\hat{\varphi}_{Pj}\left(\text{or } \hat{\varphi}_{Cj}\right)$ denotes the proportion of subjects in the target population (cohort population) that belong to category j, $j = 1,\ldots,4$.

incorrect outcome probability $R_B(x) = 1 - \exp\left(-e^{x_1+0.5x_2}\right)$. We used each model to assign risks to subjects in cohorts C_1–C_3 and to the target population sample P. We then evaluated model calibration and discrimination with and without weighting the cohort subjects. To weight them, we classified each subject into one of the $J=4$ categories of Table 1 and obtained weights (3) using the empirical proportions of subjects in cohort and population-based samples.

3.3 Simulation results

To evaluate model calibration, we applied the GOF statistic X_L^2 of equation (1) to each of the 1200 data sets and assessed the proportion of data sets in which the hypothesis of good model calibration was rejected at the 5% significance level. For each data set, and for each of the two models, we evaluated X_L^2 for the entire cohort ($L=1$) and for $L=4$ subgroups determined by the subjects' assigned risks as: (1) [0.00,0.10), (2) [0.10,0.15), (3) [0.15,0.20), (4) 0.20,1.00]. The upper half of Table 2 shows that, for the well-calibrated Model A, the empirical test size of the overall test statistic X_1^2 was slightly lower than the nominal 5%, particularly when weighting was used with data from the two biased cohorts C_1 and C_2. However, agreement between actual and nominal size was better for the subgroup-specific test X_4^2. In contrast, the lower half of the table shows that weighting outperforms an unweighted analysis in detecting the poor calibration of Model B with data from a biased cohort. Specifically, when applied to cohort C_2 (obtained by oversampling subjects with low values of covariate x_2), the unweighted analyses had low power (39.3% for X_1^2 and 34.6% for X_4^2) to detect the poor model fit, while the weighted analyses had nearly 100% power.

We also evaluated model discrimination by estimating the concordance statistic (2) for each cohort and each of the two predictive models. We compared these estimates with the theoretical concordance of the two models in a hypothetically infinite uncensored population, given by $AUC_A = 0.790$ and $AUC_B = 0.776$ for Models A and B, respectively. Table 3 shows close agreement between empirical and theoretical concordance for both models when applied to the unbiased cohort C_3. However, the unweighted concordance estimates obtained from the two biased cohorts C_1 and C_2 were appreciably lower than their theoretical values, indicating substantial downward bias for both the well-calibrated Model A and the poorly calibrated Model B. In contrast, the weighted analysis showed close agreement with the theoretical values. As expected, the weighted estimates showed greater variability than did those of the unweighted analyses. This increased variability reflects the additional uncertainty involved in the weighting procedure.

4 Application to data

We illustrate the weighted method by using it to evaluate a predictive model for epithelial ovarian cancer occurrence within 10 years of risk assignment,[16] as applied to a target population of US white women aged

Table 2. Proportion of 1200 data sets rejecting (at 5% significant level) the hypothesis that Models A and B are well-calibrated.

	Overall calibration ($L=1$ group)		Subgroup-specific calibration ($L=4$ groups)[a]	
	Analysis type		Analysis type	
Cohort	Unweighted	Weighted	Unweighted	Weighted
Model A[b]				
1	0.035	0.016	0.047	0.061
2	0.033	0.015	0.048	0.046
3	0.028	0.020	0.055	0.052
Model B[c]				
1	1.000	1.000	1.000	0.991
2	0.393	1.000	0.346	0.980
3	1.000	1.000	1.000	1.000

[a]$L=4$ subgroups determined by PPM-assigned risks in intervals [0, 0.10], [0.10, 0.15], [0.15–0.20], [0.20, 1].
[b]Model A is well-calibrated to simulated population.
[c]Model B is poorly calibrated to the simulated population.

Table 3. Mean and standard deviation (SD) of 1200 estimates of PPM concordance[a] to population risks.

	Analysis type			
	Unweighted		Weighted	
Cohort	Mean	SD	Mean	SD
Model A				
1	0.750	0.008	0.788	0.014
2	0.756	0.008	0.789	0.014
3	0.791	0.006	0.791	0.006
Model B				
1	0.739	0.008	0.775	0.013
2	0.748	0.008	0.775	0.015
3	0.777	0.006	0.777	0.006

[a]True concordance is 0.790 for Model A and 0.776 for Model B.

50–79 years with at least one intact ovary and no prior ovarian cancer. We evaluated this model by application to participants in the California Teachers Study (CTS), a cohort whose distribution of ovarian cancer risk factors may differ from that of the target population. To estimate the population distribution, we also assembled covariate data from four cross-sectional National Health and Nutrition Surveys (NHANES) conducted by the National Center for Health Statistics (NCHS) in years 1999–2000, 2001–2002, 2003–2004, and 2005–2006. (https://wwwn.cdc.gov/nchs/nhanes/Search/Nhanes99_00.aspx). Our goals were to compare the assigned risk distributions of CTS and NHANES subjects, and: (a) evaluate model performance by weighting the CTS subjects to make their covariate distribution similar to that of the US population; and (b) determine whether the performance estimates differ from those obtained without weighting the CTS subjects.

4.1 Ovarian cancer predictive model

The predictive model of Pfeiffer et al.[16] assigns a woman a 10-year ovarian cancer probability based on her values $x = (x_1, \ldots, x_5)$ of five covariates reported at the time of risk assignment. Here $x_1, \ldots, x_5$ represent, respectively, her age (years), parity (0,1–2,3 + full-term pregnancies), years of menopausal hormone therapy (MHT) (0, < 10, 10+), history of oral contraceptive (OC) use for at least one year (yes,no), and first-degree family history (FH) of breast or ovarian cancer (yes, no) The values of these covariates determine her risk as

$$R(x) = 1 - exp\left[-e^{\beta_2 x_2 + \cdots + \beta_5 x_5} \int_{x_1}^{x_1+10} \lambda_0(u)\mathrm{d}u\right]$$

where $\lambda_0(u)$ is the baseline hazard rate at age u for a nulliparous woman with no MHT use, less than a year of OC use, and no FH of breast or ovarian cancer. (See Pfeiffer et al.[16] for a description of how the baseline rate and regression coefficients were obtained.) We imputed values for incomplete covariate data in both CTS and NHANES samples by using multiple imputation techniques[17] implemented in SAS 9.4 (SAS Institute, Cary NC). The FH covariate was missing for all NHANES subjects, so for each subject we imputed a probability of FH positivity using the complete data from an earlier NHANES sample of $N = 1986$ subjects[18] via logistic regression of FH positivity against age and the occurrence and timing of prior personal breast cancer. To evaluate the impact of uncertainty in the imputed FH covariate, we also imputed this covariate as the prediction of a random forest,[19] and then assigned each NHANES subject a corresponding Pfeiffer-model risk. We found high correlation ($r = 0.97$) between the two sets of Pfeiffer-model risks, suggesting that the distribution of the NHANES risks is robust to imputation uncertainty.

4.2 CTS cohort

In 1995–1996, the CTS enrolled 133,479 female California public school professionals (active or retired) ages 22 years or older at recruitment, and followed them for subsequent morbidity and mortality through

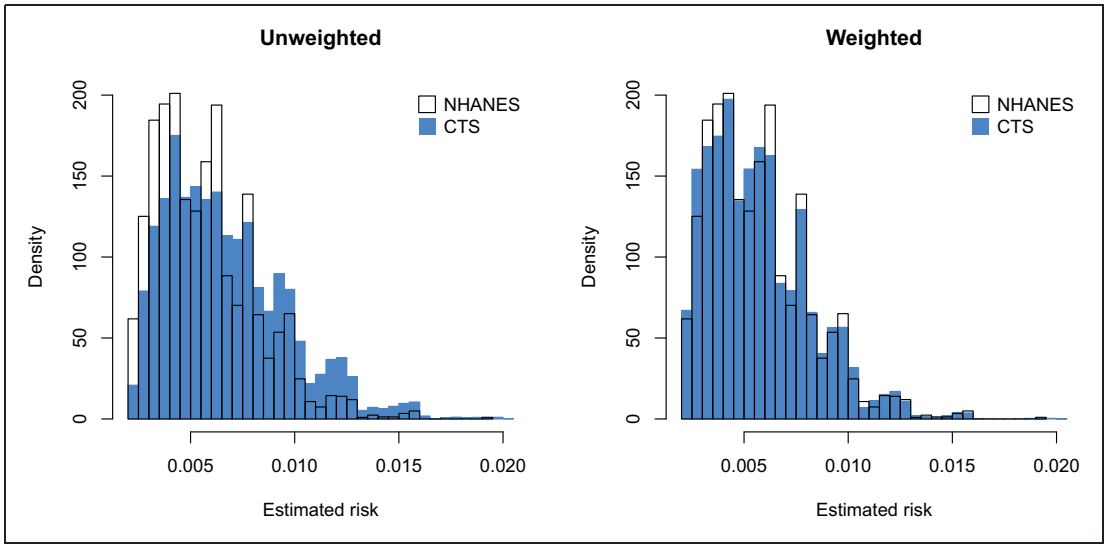

Figure 1. Unweighted (left panel) and weighted (right panel) histograms of assigned risks among California Teachers Study (CTS) subjects (shaded bars) compared to those among the NHANES subjects. The unweighted cohort risks were higher than those of the NHANES sample ($P < 10^{-4}$), while the weighted risk distributions were similar ($P = 0.81$).

31 December 2011. Further information about the cohort can be found in Bernstein et al.[20] The present analysis includes covariates and subsequent ovarian cancer incidence of $N_C = 46{,}743$ CTS subjects who at enrollment met the required eligibility criteria. We applied the Pfeiffer model to subjects' covariate data and thus obtained a distribution of assigned risks among these CTS subjects.

4.3 NHANES sample

Each of the four NHANES surveys is a stratified multistage sample of the non-institutionalized civilian US population.[21,22] Subjects were selected using a complex sampling design, to ensure unbiased and efficient representation of this population.[21] To accommodate the NHANES sampling design, we followed Johnson et al.[22] to extract from the NCHS website (http://wwwn.cdc.gov/nchs/nhanes/search/DataPage.aspx) a subject-specific sampling weight, s_i, for each of the $N_P = 2009$ NHANES subjects who met the age-, race- and ovarian eligibility criteria of the current target population. We then standardized each subject's sampling weight as $\varsigma_{Pi} = s_i/\bar{s}$, where $\bar{s} = N_P^{-1} \sum_{i'=1}^{N_P} s_i, i = 1, .., N_p$ is the mean of the sampling weights. In this way we obtained a joint covariate distribution and corresponding distribution of model-assigned risks among the NHANES subjects that provide unbiased estimates of these distributions among white females in the US target population.

4.4 Comparison of assigned risks in CTS and NHANES

The left panel of Figure 1 compares the unweighted distribution of model-assigned risks among the CTS subjects to that of the NHANES subjects. The figure shows higher risks for the CTS subjects than those for the NHANES sample; indeed a test of the null hypothesis of a common distribution for the two underlying populations was rejected ($P < 10^{-4}$). This difference is supported by comparison of the marginal distributions of the five Pfeiffer model covariates shown in Table 4: CTS subjects were more likely than NHANES subjects to be nulliparous, to have a family history of breast or ovarian cancer, and to have used MHT. Accordingly, we classified both samples into the $J = 3 \times 3 \times 2 \times 2 \times 3 = 108$ joint covariate categories corresponding to the marginal categories of Table 4, and used formula (3) to weight each of the CTS subjects. The right panel of Figure 1 shows the resulting weighted distributions, which do not differ significantly ($P = 0.807$).

4.5 Performance evaluation

We next evaluated the calibration and discrimination of the Pfeiffer model to the outcomes in the CTS cohort, with and without using the weighted method. Specifically, we evaluated the GOF statistic X_L^2 of equation (1) for the

Table 4. Distributions of Caucasian CTS and NHANES[a] subjects who at interview were aged 50–79 years and reported no prior history of ovarian cancer or bilateral oophorectomy.

Covariate	CTS	NHANES
No. of subjects	46,743	2009
Age (yrs)		
50–59	0.483	0.480
60–69	0.311	0.302
70–79	0.206	0.218
Parity		
0	0.195	0.115
1–2	0.461	0.405
3+	0.344	0.479
yrs OC use		
< 1	0.504	0.457
1+	0.495	0.543
Family hx		
yes	0.168	0.087
no	0.832	0.913
yrs MHT use		
0	0.336	0.616
<10	0.471	0.222
10+	0.193	0.162

[a]Cross-sectional NHANES data in years 1999–2000, 2001–2002, 2003–2004, and 2005–2006.

entire cohort ($L = 1$ subgroup) and for $L = 4$ subgroups determined by assigned risks in the intervals (0–0.004], (0.004–0.006], (0.006–0.008], and (0.008–1]. The overall GOF statistic was $X_1^2 = 3.84 (P = 0.053)$ for the unweighted analysis and $X_1^2 = 4.63$ ($P = 0.031$) for the weighted analysis, providing stronger evidence of poor fit. Figure 2 contrasts subgroup-specific mean assigned risks with outcome incidence estimates for each of the $L = 4$ subgroups, based on the unweighted (left panel) and weighted (right panel) analyses. The corresponding GOF statistics were $X_4^2 = 1.55$ ($P = 0.185$) for the unweighted analysis and $X_4^2 = 3.24$ ($P = 0.016$) for the weighted analysis. These results suggest that the weighted analysis detects poor fit to the population that is not evident in the unweighted analysis. Indeed, as shown in the right panel of Figure 2, the weighted mean assigned risk among subjects in the lowest assigned risk group substantially exceeded the weighted estimate of ovarian cancer incidence. While the unweighted proportion of subjects in this risk group was 25.0%, the weighted proportion was 36.7%, indicating a larger proportion of low risk subjects in the general population than in the CTS cohort. Thus for this example, the weighted and unweighted analyses yield different inferences about the calibration of the Pfeifer model to the US population. And their difference supports the simulation findings that an unweighted analysis of selected cohort subjects can lack power to detect problems in a model's calibration to the target population, while a weighted analysis can correctly detect it.

4.6 Discrimination

The weighted and unweighted analyses of the Pfeiffer model's discrimination based on the CTS data revealed nearly identical concordance estimates: $\widehat{AUC} = 0.621$ (95% confidence interval 0.588, 0.653) for the un-weighted analysis, and $\widehat{AUC} = 0.628$ (0.585, 0.670) for the weighted analysis. Thus for this example, adjusting for cohort selection bias has little effect on the model's discriminatory ability: both analyses indicate relatively poor model discrimination, which is characteristic of predictive models for ovarian cancer.[16]

In summary, the weighted analysis suggests that if the Pfeiffer predictive model were fit to both the covariates and the subsequent ovarian cancer incidence of a random sample of the general US target population, it would show poorer calibration, but similar discrimination to estimates based on women at higher risk of the disease.

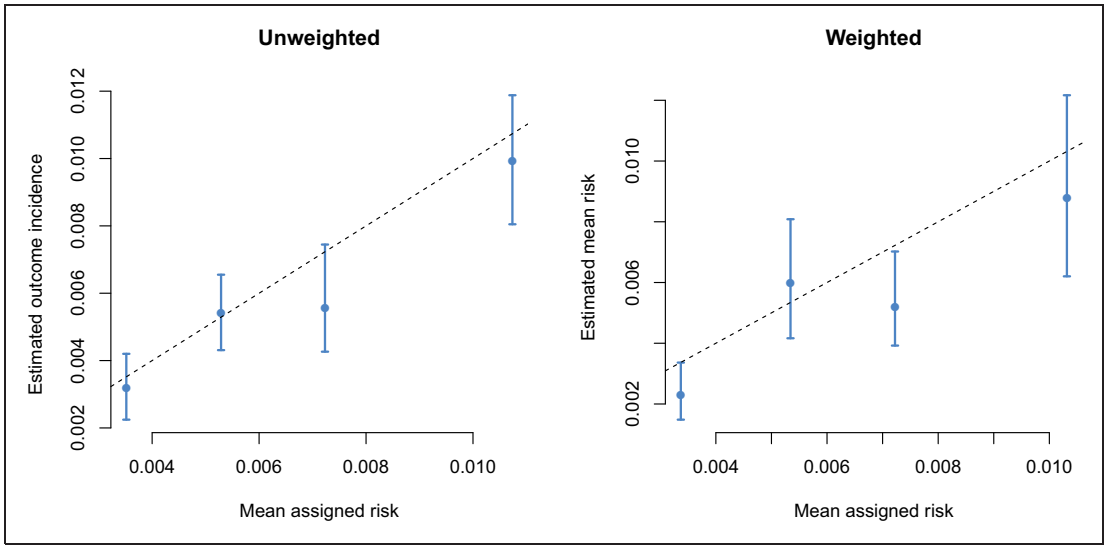

Figure 2. Plot of unweighted values of outcome incidence vs assigned risks (left panel) and corresponding weighted values (right panel).

5 Discussion

We have proposed a weighting strategy for using external cohort data to evaluate how a personal predictive model would perform if applied to a target population for which the model is needed, when the target and cohort samples may have different covariate distributions. (When the likelihood of different risk distributions in the two populations is uncertain, one could regress subjects' binary population membership indicators against their assigned risks, and check for significant association. However this strategy is limited by the sample-size dependence of detecting such an association.) The strategy involves partitioning both samples into a common set of joint covariate categories, and then weighting the cohort subjects in each category in proportion to the category-specific relative frequencies of subjects in target and cohort samples. The strategy's efficacy rests on two assumptions: (1) that the model covariates have the same outcome effect-sizes in each population; and (2) that each joint category represented in the target population also contains cohort subjects.

We used simulations to compare weighted and unweighted model performance assessments, when the distribution of covariates in cohort and target samples do and do not differ. We also illustrated the methods by application to a model for the probability of ovarian cancer diagnosis within ten years of risk assessment, as applied to the CTS cohort of white women. The distribution of ovarian cancer risks in this cohort differs from that of the US female white population represented by the NHANES samples: comparison of risks among women in the two samples shows that, in general, CTS participants have higher levels of ovarian cancer risk factors and higher ovarian cancer risks than do US white women, in agreement with the earlier observations.[20] Weighting the CTS cohort substantially reduced these risk differences.

For assessing model calibration, the simulations showed that an unweighted analysis of a biased cohort can miss poor model calibration to the target population when it exists, while a weighted analysis correctly detects the poor calibration. This phenomenon was also evident when using the CTS data to evaluate how well the Pfeiffer ovarian cancer model matches the risks of US women. Specifically, the unweighted calibration test produced little evidence for poor model fit, while the weighted test statistic showed statistically significant evidence for poor calibration. These findings suggest that in the presence of cohort/population covariate differences, a nonsignificant goodness-of-fit statistic obtained using an unweighted analysis cannot be interpreted as evidence that the model is well-calibrated to the target population. A more reliable assessment of such calibration would require a weighted analysis.

For assessing model discrimination, the simulations showed that using an unweighted analysis of a biased cohort can produce concordance estimates that are biased downward, while the weighted-based concordance estimates agreed well with the theoretical concordances. Yet for the ovarian cancer example, despite large differences in the risk distributions of the CTS participants and the US population (as estimated using NHANES data), the weighted and unweighted estimates for the Pfeiffer model's concordance were essentially the same. Since the concordance

measure C is invariant to rank-preserving risk transformations, this similarity suggests that the selection bias of CTS subjects causes a rank-preserving transformation of risks in the US population.

Acknowledgement

We express our appreciation to the California Teachers Study participants and to the California Teachers Study Steering Committee members responsible for the formation and maintenance of the cohort who did not directly contribute to the current paper: Hoda Anton-Culver, Jessica Clague deHart, Christina A Clarke, Dennis Deapen, James V Lacey Jr, Eunjung Lee, Huiyan Ma, David Nelson, Susan L Neuhausen, Hannah Park, Rich Pinder, Peggy Reynolds, Sophia S Wang, and Argyrios Ziogas.

Declaration of conflicting interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work of Scott Powers was supported by a Graduate Research Fellowship from the National Science Foundation. This work was supported by the US NIH grants R01CA094069 and R01CA 077398.

References

1. Austin PC, Pencinca MJ and Steyerberg EW. Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. *Stat Methods Med Res* 2017; **26**: 1053–1077.
2. Debray TP, Vergouwe Y, Koffijberg H, et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Elidemiol* 2015; **68**: 279–289.
3. Vergouwe Y, Moons KG and Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010; **172**: 971–980.
4. van Klaveren D, Gonen M, Steyerberg EW, et al. A new concordance measure for risk prediction models in external validation settings. *Stat Med* 2016; **35**: 4136–4152.
5. Vergouwe Y, Nieboer D, Oostenbrink R, et al. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med* 2016.
6. Riley RD, Ahmed I, Debray TP, et al. Summarising and validating test accuracy results across multiple studies for use in clinical practice. *Stat Med* 2015; **34**: 2081–2103.
7. Hernan MA, Hernandez-Diaz S and Robins JM. A structural approach to selection bias. *J Epidemiol* 2004; **15**: 615–625.
8. Haneuse S, Schildcrout J, Crane P, et al. Adjustment for selection bias in observational studies with application to the analysis of autopsy data. *Neuroepidemiology* 2009; **32**: 229–239.
9. Geneletti S, Richardson S and Best N. Adjusting for selection bias in retrospective, case-control studies. *Biostatistics* 2009; **10**: 17–31.
10. Pepe M. *The statistical evaluation of medical tests for calibration and prediction*. New York: Oxford University Press, 2003.
11. Gong G, Quante AS, Terry MB, et al. Assessing the goodness of fit of personal risk models. *Stat Med* 2014; **33**: 3179–3190.
12. Duchesne P and De Micheaux PL. Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Comput Stat Data Anal* 2010; **54**: 858–862.
13. Davies RB. Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *J Roy Stat Soc C* 1980; **29**: 323–333.
14. Hung H and Chiang CT. Estimation methods for time-dependent AUC models with survival data. *Can J Stat* 2010; **38**: 8–26.
15. Blanche P, Dartigues JF and Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med* 2013; **32**: 5381–8397.
16. Pfeiffer RM, Park Y, Kreimer AR, et al. Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: Derivation and validation from population-based cohort studies. *PLoS Med* 2013; **10**: e1001492.
17. Berglund P and Heeringa SG. *Multiple imputation of missing data using SAS*. Cary, NC: SAS Institute, 2014.
18. Ingram DD and Makuc DM. Statistical issues in analyzing the NHANES I Epidemiologic Followup Study. Series 2: Data evaluation and methods research. *Vital Health Stat 2* 1994; 1–30.

19. Friedman J, Hastie T and Tibshirani R. *The elements of statistical learning*. Vol 1, Springer, Berlin: Springer series in statistics, 2001.
20. 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.
21. Curtin LR, Mohadjer LK, Dohrmann SM, et al. National Health and Nutrition Examination Survey: Sample design, 2007-2010. *Vital Health Stat* 2013; **2**: 1–23.
22. Johnson CL, Paulose-Ram R, Ogden CL, et al. National health and nutrition examination survey: Analytic guidelines, 1999–2010. *Vital Health Stat* 2013; **2**: 1–24.
23. Kalbfleisch JD and Prentice RL. *The statistical analysis of failure time data*, 2nd ed. New York: John Wiley & Sons, 2002.

Appendix

Here we describe, for an arbitrary subset of the cohort subjects, weighted estimates of: (a) their cumulative incidence of outcome occurrence by time t^*; and (b) the concordance between a prediction model and outcome status (outcome occurrence by t^* vs. outcome-free at t^*). The weights are positive numbers satisfying $\sum_{i=1}^{N_C} w_{Ci} = N_C$, where N_C is the number of subjects in the entire cohort. The standard ''unweighted'' estimates are obtained by setting all cohort subjects' weights w_{Ci} equal to one, $i = 1, \ldots, N_C$.

Let T and Z denote a subject's times to outcome and censoring, respectively. We observe only $X = \min(T,Z)$ and the value of an indicator $\varepsilon = 1(X = T)$. We assume that a subject's time Z to censoring is independent of both his assigned risk and outcome time T. Let $S_T(t) = \Pr(T > t)$ denote the outcome survival function at time t after cohort entry, with similar notation $S_Z(t) = \Pr(Z > t)$ and $S_X(t) = \Pr(X > t)$ for the survival functions of censoring and the minimum occurrence time X of outcome and censoring. We used weighted versions of well-known nonparametric estimates of these survival functions.[23] To describe them, let $Y_i(t)$ denote the left-continuous random process taking value 1 if the ith subject is at risk of outcome or censoring at time t, and zero otherwise. Also let $N_i^{(1)}(t) = 1(X_i \le t, \varepsilon_i = 1)$ and $N_i^{(0)}(t) = 1(X_i \le t, \varepsilon_i = 0)$ denote the right-continuous counting processes taking value 1 if at time t the ith subject is outcome positive and censored, respectively, and zero otherwise. Finally, let $t_1 < \cdots < t_K$ and $z_1 < \cdots < z_J$ denote the distinct occurrence times of outcomes and censoring, respectively. Then for $t \le t*$ let

$$n(t) = \sum_{i=1}^{N_C} w_{Ci} Y_i(t) \tag{6}$$

denote the weighted count of subjects at risk just before time t, and let

$$\begin{aligned} d_k^{(1)} &= \sum_{i=1}^{N_C} w_{Ci} Y_i(t_k) N_i^{(1)}(t_k) \\ d_j^{(0)} &= \sum_{i=1}^{N_C} w_{Ci} Y_i(z_j) N_i^{(0)}(z_j) \end{aligned} \tag{7}$$

denote the weighted counts of subjects who develop the outcome at time t_k, $k = 1, \ldots, K$, or who are censored at time z_j, $j = 1, \ldots, J$. The weighted survival function estimates at time $t \le t*$ are

$$\begin{aligned} \hat{S}_T(t) &= \prod_{k:t_k \le t} \left(\frac{n(t_k) - d_k^{(1)}}{n(t_k)} \right) \\ \hat{S}_Z(t) &= \prod_{j:z_j \le t} \left(\frac{n(z_j) - d_j^{(0)}}{n(z_j)} \right) \\ \hat{S}_X(t) &= \frac{1}{N_C} \sum_{i=1}^{N_C} w_{Ci} 1(X_i > t) \end{aligned} \tag{8}$$

Note that $\hat{S}_T(t)$ and $\hat{S}_Z(t)$ are weighted Kaplan–Meier estimates for the survival functions for outcome and censoring, while $\hat{S}_X(t)$ is a weighted empirical survival function estimate.

Weighted estimates of outcome incidence. For any subgroup of cohort subjects, the weighted estimate of the cumulative outcome incidence π at time t* is $\hat{\pi} = 1 - \hat{S}_T(t^*)$, where $\hat{S}_T(t^*)$ is given by equation (8). In the absence

of censoring, the incidence estimates $\hat{\pi}$ reduce to weighted binomial proportions of subjects who develop the outcome during the risk period.

Weighted estimates of concordance. We rewrite a predictive model's concordance (2) as

$$\begin{aligned} \mathrm{AUC} &= \Pr\left(R_i > R_j | T_i \le t^*, T_j > t^*\right) = \frac{\Pr\left(T_i \le t^*, T_j > t^*, R_i > R_j\right)}{\Pr\left(T_i \le t^*, T_j > t^*\right)} \\ &= \frac{\Pr\left(Z_i > T_i, T_i \le t^*, X_j > t^*, R_i > R_j\right)}{\Pr(Z_i > T_i)[1 - S_T(t^*)]S_X(t^*)} \end{aligned} \tag{9}$$

where $S_T(\cdot), S_Z(\cdot)$ and $S_X(\cdot) = S_T(\cdot)S_Z(\cdot)$ are, respectively, the survival functions for T, Z, and X. Based on the right side of equation (9), Hung and Chiang[14] and Blanche et al.[15] proposed estimating AUC as

$$\widehat{\mathrm{AUC}} = \left\{ \frac{1}{N_C(N_C - 1)} \sum\nolimits_{i \neq j} \frac{w_{Ci} w_{Cj} 1\left(X_i = T_i \le t^*, X_j > t^*, R_i > R_j\right)}{\hat{S}_Z(X_i)} \right\} \Big/ \left\{ \left[1 - \hat{S}_T(t^*)\right] \hat{S}_X(t^*) \right\} \tag{10}$$

where $\hat{S}_T(\cdot), \hat{S}_Z(\cdot)$ and $\hat{S}_X(\cdot)$ are given by equation (8).

We used the bootstrap to estimate the variances of the estimates $\hat{\pi}$ and $\hat{C}$.