Biostatistics (2013), **14**, 2, *pp.* 366–381
doi:10.1093/biostatistics/kxs037
Advance Access publication on October 19, 2012

Extensions of criteria for evaluating risk prediction models for public health applications

RUTH M. PFEIFFER

Biostatistics Branch, National Cancer Institute, Bethesda, MD 20892-7244, USA
pfeiffer@mail.nih.gov

Summary

We recently proposed two novel criteria to assess the usefulness of risk prediction models for public health applications. The proportion of cases followed, PCF(p), is the proportion of individuals who will develop disease who are included in the proportion p of individuals in the population at highest risk. The proportion needed to follow-up, PNF(q), is the proportion of the general population at highest risk that one needs to follow in order that a proportion q of those destined to become cases will be followed (Pfeiffer, R. M. and Gail, M. H., 2011. Two criteria for evaluating risk prediction models. *Biometrics* **67**, 1057–1065). Here, we extend these criteria in two ways. First, we introduce two new criteria by integrating PCF and PNF over a range of values of q or p to obtain iPCF, the integrated PCF, and iPNF, the integrated PNF. A key assumption in the previous work was that the risk model is well calibrated. This assumption also underlies novel estimates of iPCF and iPNF based on observed risks in a population alone. The second extension is to propose and study estimates of PCF, PNF, iPCF, and iPNF that are consistent even if the risk models are not well calibrated. These new estimates are obtained from case–control data when the outcome prevalence in the population is known, and from cohort data, with baseline covariates and observed health outcomes. We study the efficiency of the various estimates and propose and compare tests for comparing two risk models, both of which were evaluated in the same validation data.

Keywords: Area under the receiver operator characteristics curve (ROC); AUC; Discrimination; Discriminatory accuracy; Risk models; Study design.

1. Introduction

Statistical models that predict disease incidence (Freedman *and others*, 2009), disease recurrence (Stephenson *and others*, 2006), mortality following disease onset (Albertsen *and others*, 2005), or response to treatment (O'Brien *and others*, 2011) are used in clinical practice and decision making, for example, to inform choices for a prevention or treatment with serious side effects. These models also have public health applications. They can be used to target preventive interventions to those with high enough risks to justify an intervention that has adverse effects and to identify high-risk individuals for intensive screening for early detection of disease.

We recently proposed two measures of concentration of risk that are directly relevant to public health decisions. We defined the "proportion of cases followed", PCF(p), as the proportion of cases that would be followed in a program that followed the proportion p of the population at highest risk. We also proposed

Published by Oxford University Press 2012.

a complementary criterion, the "proportion needed to follow-up", PNF(q), namely the proportion of the general population at highest risk that one needs to follow in order that a proportion q of those destined to become cases will be followed. We also derived tests for comparing PCF and PNF for two risk models evaluated in the same validation data (Pfeiffer and Gail, 2011).

Here, we extend these criteria in two ways. In the previous work, p or q were prespecified and fixed numbers. First, we introduce two new criteria by integrating PCF and PNF over a range of values of p or q to obtain iPCF, the integrated PCF, and iPNF, the integrated PNF (Section 3). When integrating over the whole range of p and when the disease is rare, iPCF is similar to the area under the curve (AUC), the area under the receiver operating characteristic (ROC) curve (Pepe, 2003, p. 67). While the AUC is based on comparing ranks of the estimated risks in cases to those in non-cases, iPCF compares the risk in cases to risks in the whole population, which is a mixture of cases and non-cases. The AUC is ideal for measuring the discrimination accuracy in diagnostic applications, where one wants to distinguish cases from controls. For screening a general population, however, iPCF is more useful because it measures how different risks are in those destined to develop disease (or to have prevalent disease) from the population to be screened.

A key assumption in Pfeiffer and Gail (2011) was that the risk model is well calibrated. This assumption is also needed for novel estimates of iPCF and iPNF based on observed risks in a population alone (Section 4.1). The second extension is to propose and study estimates of PCF, PNF, iPCF, and iPNF that are consistent even if the risk model is not well calibrated. These new estimates are obtained from case–control data when the outcome prevalence in the population is known, and from cohort data, with baseline covariates and observed health outcomes (Sections 4.2 and 4.3). We propose testing differences between two risk models evaluated on the same dataset using iPCF and iPNF (Section 5). We then study the efficiency of the various estimates for these criteria and compare their performance for testing differences between two risk models evaluated on the same dataset in simulations (Section 6). A data example is presented in Section 7 before we close with a discussion (Section 8).

2. Notation and background

We are interested in predicting the probability of a binary event, $Y=1$ or $Y=0$. This event could denote the incidence of a particular disease over a given time period, for example, 5 years, or of dying before the end of a defined time interval after disease onset. The event could also refer to the response to a treatment in a population with a particular disease. Given a set of baseline predictors X, a risk prediction model $R(x)=P(Y=1\mid X=x)$ is a mapping from the set Ω of possible values of X to [0, 1]. In a specific population, the distribution of the covariates $F_X(x)$ induces the distribution F of risk R that has support on [0, 1] through

$$F(r)=P(R\leqslant r)=\int_{\{x:R(x)\leqslant r\}} \mathrm{d}F_X(x). \tag{2.1}$$

We let G be the distribution of risk in those who experience the event (cases, $Y=1$),

$$G(r)=P(R\leqslant r\mid Y=1),$$

and K be the distribution of risk in non-cases, or controls ($Y=0$),

$$K(r)=P(R\leqslant r\mid Y=0).$$

We denote risk realizations from F by r^F, and risk realizations from cases and non-cases by r^G and r^K, respectively.

3. Criteria to assess model performance and their estimation

3.1 *Review of the definition of PCF and PNF*

We recently proposed and studied two criteria to assess the usefulness of models that predict the risk of disease incidence for screening and prevention, or the usefulness of prognostic models for management following disease diagnosis (Pfeiffer and Gail, 2011). The first criterion, the proportion of cases followed, $\mathrm{PCF}(p)$ is the proportion of cases who are included in the proportion p of individuals in the population at highest risk, given by

$$\mathrm{PCF}(p) = 1 - G \circ F^{-1}(1-p) = 1 - G(\phi_{1-p}), \tag{3.1}$$

where $G \circ F(x) = G\{F(x)\}$ is the composition of G with F and $\phi_{1-p} = F^{-1}(1-p)$ denotes the $1-p$th quantile of F. The second criterion is the proportion needed to follow-up $\mathrm{PNF}(q)$ namely the proportion of the general population at highest risk that one needs to follow in order that a proportion q of cases will be followed, defined as

$$\mathrm{PNF}(q) = 1 - F \circ G^{-1}(1-q) = 1 - F(\gamma_{1-q}), \tag{3.2}$$

where $\gamma_{1-q} = G^{-1}(1-q)$ denotes the $1-q$th quantile of the distribution of risk in cases, G.

If risk is concentrated in a small proportion of the population at highest risk, then $\mathrm{PCF}(p)$ will be high, even for small p and $\mathrm{PNF}(q)$ will be small, even for large q.

3.2 *New criteria: iPCF and iPNF*

While $\mathrm{PCF}(p)$ and $\mathrm{PNF}(q)$ are useful criteria for model evaluation, they require the specification of thresholds p and q. To lessen the dependency of these criteria on the given thresholds, we define the iPCF as

$$\mathrm{iPCF}(p^*) = \int_{p^*}^{1} \mathrm{PCF}(p)\,\mathrm{d}W(p) = 1 - p^* - \int_0^{1-p^*} G(\phi_p)\,\mathrm{d}W(p), \tag{3.3}$$

where W is a probability measure on the unit interval. The iPNF is

$$\mathrm{iPNF}(q^*) = \int_{q^*}^{1} \mathrm{PNF}(q)\,\mathrm{d}W(q) = 1 - q^* - \int_0^{1-q^*} F(\gamma_q)\,\mathrm{d}W(q). \tag{3.4}$$

In what follows, we assume $\mathrm{d}W(p) = \mathrm{d}p$. In that case, using a change of variables, we obtain

$$\mathrm{iPCF}(p^*) = 1 - p^* - \int_0^{\phi_{1-p^*}} G(u)\,\mathrm{d}F(u) = 1 - p^* - \frac{1}{1-p^*} P\{R_G \leqslant R_F \mid R_F \in (0, \phi_{1-p^*})\} \tag{3.5}$$

and

$$\mathrm{iPNF}(q^*) = 1 - q^* - \int_0^{\gamma_{1-q^*}} F(u)\,\mathrm{d}G(u) = 1 - q^* - \frac{1}{1-q^*} P\{R_F \leqslant R_G \mid R_G \in (0, \gamma_{1-q^*})\}. \tag{3.6}$$

For the special case of $p^* = q^* = 0$, $\mathrm{iPCF}(0) = 1 - P(R_G \leqslant R_F) = P(R_F < R_G)$ and $\mathrm{iPNF}(0) = 1 - P(R_F \leqslant R_G) = P(R_F > R_G)$. We note that iPCF(0) is similar to the AUC, which can also be expressed as the probability that a randomly selected case has a higher projected risk than a randomly selected control, i.e. $\mathrm{AUC} = P(R_G > R_K)$. While the AUC is based on comparing ranks of the estimated risks in cases to those in non-cases, iPCF(0) compares risk in cases to risks in the whole population, which is a mixture of cases and non-cases. However, for a rare disease $K \approx F$, and the values of the AUC and iPCF(0) will be close. Figure 1 shows a PCF curve when the population distribution of risk F is a beta distribution with parameters $\alpha = 1.5$, $\beta = 28.5$, with $\mu \equiv P(Y = 1) = 0.05$. The area under this curve is $\mathrm{iPCF}(0) = 0.71$.

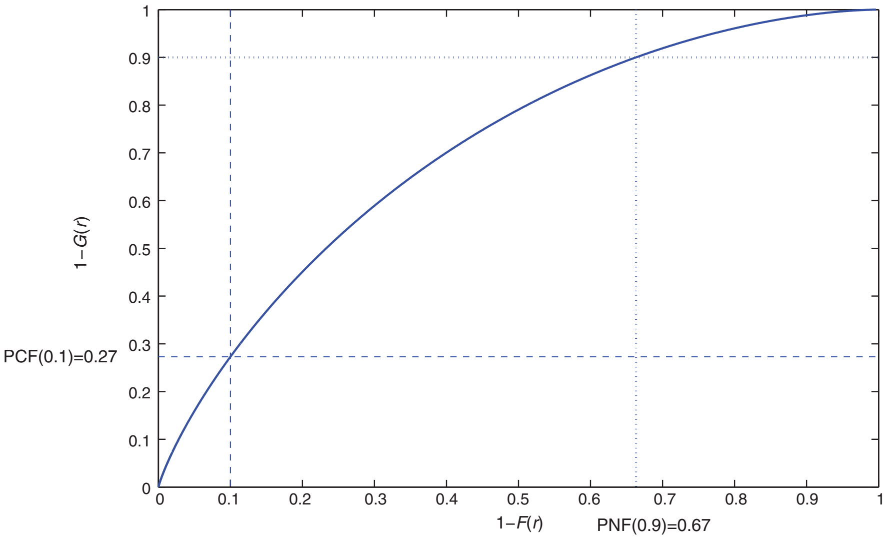

Fig. 1. PCF curve when the distribution F of risk is a beta distribution with parameters $\alpha = 1.5$, $\beta = 28.5$, with $\mu = 0.05$, corresponding to $\text{iPCF}(0) = 0.71$.

Equations (3.5) and (3.6) resemble expressions for the partial area under the ROC curve (pAUC; McGlish 1989) that focuses on the region of the ROC curve with a low false positive rate, which is often of prime interest for diagnostic tests. Likewise, $\text{iPCF}(p^*)$ and $\text{iPNF}(q*)$ can be used to focus on the high-risk portion of the population to be screened. However, (3.5) and (3.6) again use F as the reference population instead of K.

4. Estimation of $\text{PCF}(p)$, $\text{PNF}(q)$, iPCF and iPNF

We now study estimates of PCF, PNF, and their integrated versions, iPCF and iPNF, for three types of data and derive their asymptotic distributions. First, we assume that the risk model is well calibrated, and a random sample of risk estimates r_i^F, $i = 1, \ldots, N$, is observed. We first review the estimators of PCF and PNF used in Pfeiffer and Gail (2011) for this setting and propose novel estimators for iPCF and iPNF. We then estimate PCF, PNF, iPCF, and iPNF non-parameterically using random samples of risks in cases, $r_i^G \sim G$, $i = 1, \ldots, m$, and controls, $r_j^K \sim K$, $j = 1, \ldots, n$, assuming the event probability $\mu \equiv P(Y = 1)$ in the population is known from external sources. We also study the corresponding estimates when a random sample of risks in the population and the associated binary outcomes (r_i^F, Y_i), $i = 1, \ldots, N$, are available.

4.1 *Estimation using observed risks in a population*

If the risk model R is well calibrated, that is, $P(Y = 1 \mid r) = r$, i.e. among individuals with risk r the fraction of events is r, then $\mu \equiv P(Y = 1) = E(R) = \int_0^1 r \, \mathrm{d}F(r)$, and the distributions G and K of risk in cases and non-cases, respectively, can be derived from the population distribution F as

$$G(r) = P(R \leqslant r \mid Y = 1) = \frac{1}{\mu} \int_0^r t \, \mathrm{d}F(t) \tag{4.1}$$

and

$$K(r) = P(R \leqslant r \mid Y = 0) = \frac{1}{1-\mu} \int_0^r (1-t)\,\mathrm{d}F(t). \tag{4.2}$$

In this setting, using (4.1),

$$\mathrm{PCF}(p) = 1 - G(\phi_{1-p}) = 1 - \frac{1}{\mu} \int_0^{\phi_{1-p}} t\,\mathrm{d}F(t) = 1 - L(1-p), \tag{4.3}$$

where the L denotes the Lorenz curve of F (Goldie, 1977), and

$$\mathrm{PNF}(q) = 1 - F(\gamma_{1-q}) = 1 - L^{-1}(1-q), \tag{4.4}$$

where L^{-1} is the inverse of the Lorenz curve, also called the concentration curve (Goldie, 1977).

Thus, if the risk model is well calibrated, PCF and PNF can be estimated from a random sample $r_1^F, \ldots, r_N^F$ of risks from the continuous distribution F in a given population. To briefly summarize earlier work (Pfeiffer and Gail, 2011), let $r_{(1)}^F \leqslant \ldots \leqslant r_{(N)}^F$ denote the order statistics of the estimated risks, and $[x]$ be the largest integer less than or equal to x. Following Goldie (1977), and letting $S_i = \sum_{k=1}^i r_{(k)}$, an estimate of the Lorenz curve and thus PCF is

$$\widehat{\mathrm{PCF}}(p) = 1 - L_N(1-p) = 1 - S_{[N(1-p)]}/S_N. \tag{4.5}$$

Using the result of Goldie (1977) for the inverse function of the Lorenz curve, L_N^{-1}, for a fixed value of $1-q$, the PNF is estimated as

$$\widehat{\mathrm{PNF}}(q) = 1 - L_N^{-1}(1-q) = 1 - i/n, \quad S_i/S_N < 1-q \leqslant S_{i+1}/S_N, \;\; i = 0, \ldots, N. \tag{4.6}$$

By drawing on the distribution theory for the Lorenz cure and its inverse, we derived the asymptotic normality of the estimates in (4.5) and (4.6) and obtained their asymptotic variances using an influence function-based approach (Pfeiffer and Gail, 2011).

If the model is well calibrated, iPCF and iPNF also relate to the Lorenz curve and its inverse through

$$\mathrm{iPCF}(p^*) = 1 - p^* - \int_{p^*}^1 L(1-p)\,\mathrm{d}p$$

and

$$\mathrm{iPNF}(q^*) = 1 - q^* - \int_{q^*}^1 L^{-1}(1-q)\,\mathrm{d}q.$$

It is easy to see that a popular summary measure of the Lorenz curve, the Gini index (Gini, 1912), defined as $\mathrm{Gini} = 1 - 2\int_0^1 L(p)\,\mathrm{d}p$, which is commonly used to measure income inequality in economics, is related to iPCF(0) through $\mathrm{Gini} = 2\mathrm{iPCF}(0) - 1$.

Using ordered risk estimates $r_{(1)}^F \leqslant \cdots \leqslant r_{(N)}^F$ in the population, a non-parametric estimate of iPCF based on (4.5) is thus obtained by interpolation as

$$\widehat{\mathrm{iPCF}}(p^*) = 1 - p^* - \frac{1}{NS_N} \sum_{i=1}^{[(1-p^*)N]} S_i = 1 - p^* - \frac{1}{NS_N} \sum_{i=1}^{[(1-p^*)N]} ([(1-p^*)N] - i + 1) r_{(i)}. \tag{4.7}$$

For $p^* = 0$, this expression reduces to $\widehat{\mathrm{iPNF}}(0) = 1 - 1/(NS_N)\sum_1^N (N - i + 1) r_{(i)}$.

Similarly, we estimate iPNF using simple geometric arguments as

$$\widehat{\mathrm{iPNF}}(q^*) = 1 - q^* - \frac{1}{NS_N}\sum_{i=1}^{k^*} ir_{(i+1)}, \tag{4.8}$$

where k^* satisfies $S_{k^*}/S_N < q^* \leqslant S_{k^*+1}/S_N$.

In addition to being consistent, $\widehat{\mathrm{iPCF}}$ and $\widehat{\mathrm{iPNF}}$ also have asymptotically normal distributions. This follows directly from the fact that they are linear functionals of estimates of the Lorenz curve and its inverse, which are Gaussian stochastic processes (Goldie, 1977). The asymptotic variance estimates of $\widehat{\mathrm{iPCF}}$ and $\widehat{\mathrm{iPNF}}$ are given in Appendix A, see supplementary material available at *Biostatistics* online.

4.2 *Estimation using risks in a case–control sample when $\mu = P(Y = 1)$ is known*

We assume that risks $r_i^G \sim G$, $i = 1, \ldots, m$, from a random sample of cases and risks $r_j^K \sim K$, $j = 1, \ldots, n$, from a random sample of non-cases from a population are available, and that the event probability $\mu = P(Y = 1)$ in that population is known. We express the distribution of risk in the general population as $F = \mu G + (1 - \mu)K$, and estimate F using the empirical distribution functions

$$G_m(r^*) = \frac{1}{m}\sum_{i=1}^{m} I(r_i^G \leqslant r^*) \quad \text{and} \quad K_n(r^*) = \frac{1}{n}\sum_{i=1}^{n} I(r_i^K \leqslant r^*),$$

as

$$\hat{F}(r^*) = \mu G_m(r^*) + (1 - \mu)K_n(r^*).$$

Plugging G_m and $\hat{F}$ into (3.1) yields

$$\widehat{\mathrm{PCF}}(p) = 1 - G_m \circ \hat{F}^{-1}(1 - p). \tag{4.9}$$

The expression for PNF in (3.2) simplifies to $\mathrm{PNF}(q) = 1 - \mu(1 - q) - (1 - \mu)K \circ G^{-1}(1 - q)$ and thus

$$\widehat{\mathrm{PNF}}(q) = 1 - \mu(1 - q) - (1 - \mu)K_n \circ G_m^{-1}(1 - q). \tag{4.10}$$

Using $F = \mu G + (1 - \mu)K$ in (3.5), we express iPCF as

$$\mathrm{iPCF}(p^*) = 1 - p^* - \frac{\mu}{2}G^2(\phi_{1-p^*}) - (1 - \mu)P\{R_G \leqslant R_K;\, R_K \in (0, \phi_{1-p^*})\}. \tag{4.11}$$

We estimate iPCF using the empirical distribution functions G_m and K_n, and $\hat{F}$ and $\hat{\phi}_{1-p^*} = \hat{F}^{-1}$ $(1 - p^*)$ as

$$\widehat{\mathrm{iPCF}}(p^*) = 1 - p^* - \frac{\mu}{2}G_m^2(\hat{\phi}_{1-p^*}) - (1 - \mu)\frac{1}{mn}\sum_{i,j} I\{r_i^G \leqslant r_j^K;\, r_j^K \in (0, \hat{\phi}_{1-p^*})\}, \tag{4.12}$$

where $I(A)$ denotes the indicator function that is one if A is true and zero otherwise.

Similarly, iPNF given in (3.6) can be expressed as

$$\mathrm{iPNF}(q^*) = 1 - q^* - \frac{\mu}{2}(1-q^*)^2 - (1-\mu)P\{R_K \leqslant R_G,\, R_G \in (0, \gamma_{1-q^*})\}, \tag{4.13}$$

and thus an estimate is given by

$$\widehat{\mathrm{iPNF}}(1-q^*) = 1 - q^* - \frac{\mu}{2}(1-q^*)^2 - (1-\mu)\frac{1}{mn}\sum_{i,j} I\{r_i^K \leqslant r_j^G; r_j^G \in (0, \hat{\gamma}_{1-q^*})\}, \tag{4.14}$$

where $\hat{\gamma}_{1-q^*} = G_m^{-}1(1-q^*)$.

Consistency of $\widehat{\mathrm{iPCF}}(p^*)$ and $\widehat{\mathrm{iPNF}}$ follows immediately from the consistency of G_m and $\hat{\phi}_{1-p^*}$, $\hat{\gamma}_{1-q^*}$ and the fact that $EI\{r_i^G \leqslant r_j^K; r_j^K \in (0, \hat{\phi}_{1-p^*})\} = P\{R_G \leqslant R_K,\, R_K \in (0, \phi_{1-p^*})\}$ and $EI\{r_i^K \leqslant r_j^G; r_j^G \in (0, \hat{\gamma}_{1-q^*})\} = P\{R_K \leqslant R_G,\, R_G \in (0, \gamma_{1-q^*})\}$.

PCF, PNF, iPCF, and iPNF are functionals of the two distribution functions G and K that are estimated based on independent samples. We derive their asymptotic properties using a bivariate influence function approach (Pires and Branco, 2002) in Appendix B, see supplementary material available at *Biostatistics* online.

4.3 *Estimation using risks and outcomes in a population*

Here, a random sample of risks and the corresponding event outcomes in a population are available, that is, we observe the i.i.d. samples (r_i^F, Y_i), $i = 1, \ldots, N$. For a model that predicts disease incidence, these data would be comprised of risk estimates at baseline and observed outcomes at the end of the follow-up period, and for a model that predicts the prevalence of a disease, the risks and outcomes could be based on a cross-sectional sample.

We estimate PCF and PNF by plugging estimates of F, G and the corresponding quantiles ϕ and γ into the expressions (3.1) and (3.2), respectively. The distribution of risk in the general population, F, is estimated using the empirical distribution function in the whole population,

$$F_N(r^*) = \frac{1}{N}\sum_{i=1}^{N} I(r_i^F \leqslant r^*),$$

and G is estimated using the empirical distribution function among cases,

$$\hat{G}(r^*) = \frac{1}{N\bar{Y}}\sum_{i=1}^{N} I(r_i^F \leqslant r^*, Y_i = 1) = \frac{1}{\hat{n}_1}\sum_{i=1}^{\hat{n}_1} I(r_i^G \leqslant r^*),$$

where $\bar{Y} = \sum Y_i/N$ denotes the empirical mean of Y and $\hat{n}_1 = \sum Y_i$. Estimates of iPCF and iPNF are obtained in a similar way as

$$\widehat{\mathrm{iPCF}}(p^*) = 1 - p^* - \frac{1}{N\hat{n}_1}\sum_{i,j} I\{r_i^G \leqslant r_j^F; r_j^F \in (0, \hat{\phi}_{1-p^*})\} \tag{4.15}$$

and

$$\widehat{\mathrm{iPNF}}(q^*) = 1 - q^* - \frac{1}{N\hat{n}_1}\sum_{i,j} I\{r_i^F \leqslant r_j^G; r_j^G \in (0, \hat{\gamma}_{1-q^*})\}. \tag{4.16}$$

The asymptotic distributions of the estimators for PCF, PNF, iPCF, and iPNF based on risks and outcomes in a cohort differ from the asymptotic distributions of the estimators (4.9)–(4.11) and (4.14) based on the case–control data in the previous section, as they use different estimates of F, and also incorporate the variation arising from estimating the disease prevalence and the number of cases, $N\bar{Y}$, in the population. In Appendix C (see supplementary material available at *Biostatistics* online), we derive their asymptotic distributions by treating F and G as functions of the bivariate distribution function of (r^F, Y).

5. Comparing two risk models

We previously proposed test statistics based on PCF and PNF for two risk models, R^1 and R^2, both of which were applied to the same population. To test whether, for fixed p, $\mathrm{PCF}^1 = \mathrm{PCF}^2$, or for a fixed q, $\mathrm{PNF}^1 = \mathrm{PNF}^2$, we use the statistics

$$T_{\mathrm{PCF}}(p) = \frac{n\{\widehat{\mathrm{PCF}^1}(p) - \widehat{\mathrm{PCF}^2}(p)\}^2}{\hat{V}_{\mathrm{PCF}}} \quad \text{and} \quad T_{\mathrm{PNF}}(q) = \frac{n\{\widehat{\mathrm{PNF}^1}(q) - \widehat{\mathrm{PNF}^2}(q)\}^2}{\hat{V}_{\mathrm{PNF}}}, \tag{5.1}$$

where $\hat{V}$ are consistent estimates of the variance of the difference of the estimates.

Two new test statistics based on iPCF and iPNF to compare two models using correlated risk estimates (r^1, r^2) are

$$T_{\mathrm{iPCF}}(p^*) = \frac{\{\widehat{\mathrm{iPCF}^1}(p^*) - \widehat{\mathrm{iPCF}^2}(p^*)\}^2}{\hat{V}_{\mathrm{iPCF}}} \quad \text{and} \quad T_{\mathrm{iPNF}}(q^*) = \frac{\{\widehat{\mathrm{iPNF}^1}(q^*) - \widehat{\mathrm{iPNF}^2}(q^*)\}^2}{\hat{V}_{\mathrm{iPNF}}}, \tag{5.2}$$

where iPNF and iPCF for both models are evaluated at the same value p^* or q^*, respectively.

Asymptotically all test statistics, T_{PCF}, T_{PNF}, T_{iPCF}, and T_{iPNF} have a central χ^2_1 distribution under H_0. Under the alternative, the non-centrality parameters for the test statistics are $\delta_{\mathrm{PCF}} = n(\mathrm{PCF}^1 - \mathrm{PCF}^2)^2/V_{\mathrm{PCF}}$, $\delta_{\mathrm{PNF}} = n(\mathrm{PNF}^1 - \mathrm{PNF}^2)^2/V_{\mathrm{PNF}}$, $\delta_{\mathrm{iPCF}} = n(\mathrm{iPCF}^1 - \mathrm{iPCF}^2)^2/V_{\mathrm{iPCF}}$, $\delta_{\mathrm{iPNF}} = n(\mathrm{iPNF}^1 - \mathrm{iPNF}^2)^2/V_{\mathrm{iPNF}}$, respectively. The variances for all test statistics can be computed based on the respective influence functions ψ^{R_1} and ψ^{R_2} for models 1 and 2 as $V = \mathrm{Var}(\psi^{R_1} - \psi^{R_2})$, or alternatively, using a bootstrap variance estimate. In the simulations, we use the bootstrap variances in the formulas of the test statistics.

6. Simulations

6.1 *Efficiency of estimates of PCF, PNF, iPCF, and iPNF*

We use simulations to investigate the properties of the non-parametric estimates of PCF, PNF, iPCF, and iPNF defined in Sections 4.1–4.3 and to compare their efficiency. We assume that the population distribution of risk is a beta distribution with parameters α and β, $F(r) = B(r, \alpha, \beta)/B(\alpha, \beta)$, where $B(r, \alpha, \beta) = \int_0^r t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t$ and $B(\alpha, \beta) = B(1, \alpha, \beta)$. In this setting, the distributions of risk in cases and non-cases are also beta distributions, given by $G(r) = B(r, \alpha+1, \beta)/B(\alpha+1, \beta)$ and $K(r) = B(r, \alpha, \beta+1)/B(\alpha, \beta+1)$. The subscript R refers to estimates based on the population risks only, the subscript CC is used for estimates based on risks observed for a case–control sample with known disease prevalence μ, and the subscript (R, Y) refers to estimates based on risks and observed outcomes in a population. The efficiency of the various estimates is compared using their asymptotic relative efficiencies (AREs), computed as the ratios of the influence function-based variances given in the Appendices (see supplementary material available at *Biostatistics* online).

To create data for each of the study designs, we first simulated risk estimates r_i^F, $i = 1, \ldots, N$, and then generated the binary outcomes Y_i from a binomial distribution with probability r_i, $Y_i \sim \mathrm{binom}(1, r_i)$, $i =$

$1, \ldots, N$. For the estimates using the population-based risks and the risks and outcomes, we used all the observations of r_i^F or (r_i^F, Y_i), respectively. To create a case–control study, we split the population into cases and non-cases and used the risk estimates from all the cases and all the non-cases together, with the true value of the disease prevalence μ. Thus, all estimates in a given simulation are based on the same observations.

Table 1 gives results for 500 simulations each based on a random sample of size $N = 10\,000$ for a rare disease. The beta distribution parameters were $\alpha = 6.55$, 1, and 0.3 with corresponding $\beta = 124.45$, 19.0, and 5.7 and expected risk $\mu = E(R) = 0.05$ for each (α, β) pair. The AUC values for these parameter choices are 0.63, 0.79, and 0.98, respectively, corresponding to models with moderate to very high discriminatory ability. The mean estimates of PCF and PNF were very close to the theoretical values for all estimators. However, the estimates based on case–control data and cohort data with outcomes were much less precise than the corresponding estimates $\widehat{\text{PCF}}_R$ and $\widehat{\text{PNF}}_R$. The AREs of $\widehat{\text{PCF}}_{\text{CC}}$ compared with $\widehat{\text{PCF}}_R$ ranged from 22.86 to 387.2, with more loss of efficiency for parameter values corresponding to smaller AUCs. For example, for $p = 0.20$, ARE $= 367.45$ for $(\alpha, \beta) = (6.55, 124.45)$ and ARE $= 22.86$ for $(\alpha, \beta) = (0.3, 5.7)$. The AREs for the estimates based on risks and outcomes and based on case–control data were close to 1.00 in all cases. Estimates of PNF behaved very similarly to estimates of PCF in terms of efficiency (Table 1). Again, $\widehat{\text{PNF}}_{\text{CC}}$ and $\widehat{\text{PNF}}_{(R,Y)}$ were much less efficient than $\widehat{\text{PNF}}_R$ for all settings.

Supplementary material available at *Biostatistics* online, Table S1, gives results for a common disease, $\mu = 0.30$ with $\alpha = 6.55$, 1, and 0.3 and corresponding $\beta = 15.28$, 2.33, and 0.7, leading to the same AUC values as in Table 1. The patterns were very similar to those seen in Table 1, but the loss in efficiency for the case–control based estimates and for cohort data with outcome-based estimates was less pronounced than for a rare disease. $\widehat{\text{PCF}}_{\text{CC}}$ and $\widehat{\text{PNF}}_{\text{CC}}$ were somewhat more efficient than $\widehat{\text{PCF}}_{(R,Y)}$ and $\widehat{\text{PNF}}_{(R,Y)}$ for parameters resulting in larger AUC values. For example, for $p = 0.10$, the variance ratio of $\widehat{\text{PCF}}_{\text{CC}}$ compared with $\widehat{\text{PCF}}_{(R,Y)}$ was ARE $= 3.43$ for $(\alpha, \beta) = (0.3, 0.7)$.

Table 2 gives results for iPCF and iPNF for 500 simulations, each based on a random sample of size $N = 10\,000$ for the same simulation settings as in Table 1. Again, all methods had mean estimates that were very close to the theoretical values of iPCF and iPNF. Similar to PCF and PNF the estimates of iPCF and iPNF based on case–control data and (R, Y) were much less precise than corresponding estimates $\widehat{\text{iPCF}}_R$ and $\widehat{\text{iPNF}}_R$. For $\widehat{\text{iPCF}}_{\text{CC}}$ compared with $\widehat{\text{iPCF}}_R$, the AREs ranged from 316.73 to 46.58 and were lower for parameter values corresponding to larger AUCs. For example, for $p = 0.20$, ARE $= 311.36$ for $(\alpha, \beta) = (6.55, 124.45)$ and ARE $= 59.61$ for $(\alpha, \beta) = (1, 2.33)$. Estimates of iPNF_{CC} and $\text{iPNF}_{(R,Y)}$ were also less efficient than $\widehat{\text{iPNF}}_R$, but the loss of efficiency was less pronounced than for iPCF. $\widehat{\text{iPCF}}_{\text{CC}}$ and $\widehat{\text{iPNF}}_{\text{CC}}$ were slightly more efficient than estimated based on the cohort with outcomes. Results for iPCF and iPNF for a common disease are given in Table S2, see supplementary material available at *Biostatistics* online.

6.2 *Comparing two risk models using PCF, PNF, and iPCF and iPNF*

We examined the size and power of the tests (5.1) and (5.2) for comparing risk models 1 and 2 when PCF, PNF, iPCF, and iPNF are estimated from observed bivariate risks (r_i^1, r_i^2), $i = 1, \ldots, N$, and risks and outcomes in a population. To simulate bivariate risks with outcome data, we first drew a random number m of cases $(Y = 1)$ in a population of size N from a binomial distribution with parameter μ, and assigned the remaining $n = N - m$ individuals to be controls $(Y = 0)$. To obtain risk estimates from each model that have a marginal beta distribution and are correlated, we first generated bivariate normal random variables $(X_{i1}, X_{i2}) \sim MVN((0, 0), \Sigma)$ $i = 1, \ldots, N$, where $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = \rho$. We then separately computed risks for the n cases and m controls from $r_{i1} = F_1^{-1} \circ \Phi^{-1}(X_{i1})$ and $r_{i2} = F_2^{-1} \circ \Phi^{-1}(X_{i2})$ where F_k^{-1}, $k = 1, 2$, denotes the inverse of the beta distribution function with parameters

Table 1. *Mean values of PCF and PNF estimated using observed risks R in a population, assuming that the model is well calibrated; risk estimates in a case–control sampling when the disease prevalence μ is known, and based on observations of (R, Y) in the population*

		$\widehat{\text{PCF}}$			var($\widehat{\text{PCF}}$)			ARE		
p	PCF true	R	CC	(R, Y)	R	CC	(R, Y)	CC/R	$(R, Y)/R$	(R, Y)/CC
$\alpha = 6.55,\ \beta = 124.45^{\dagger}$										
0.10	0.18	0.18	0.18	0.18	0.01	2.67	2.67	386.71	387.23	1.00
0.20	0.32	0.32	0.31	0.31	0.01	4.03	4.03	367.45	368.10	1.00
0.30	0.44	0.44	0.44	0.44	0.01	4.63	4.65	362.90	363.84	1.00
0.40	0.55	0.55	0.55	0.55	0.01	4.70	4.68	369.90	368.69	1.00
$\alpha = 1,\ \beta = 19^{\ddagger}$										
0.10	0.32	0.32	0.32	0.32	0.06	3.80	3.81	59.41	59.64	1.00
0.20	0.51	0.51	0.51	0.51	0.08	4.60	4.61	59.27	59.39	1.00
0.30	0.65	0.65	0.65	0.65	0.07	4.30	4.31	62.52	62.70	1.00
0.40	0.76	0.76	0.76	0.76	0.05	3.53	3.53	68.18	68.19	1.00
$\alpha = 0.3,\ \beta = 5.7^{\S}$										
0.10	0.51	0.51	0.51	0.51	0.22	4.23	4.32	19.07	19.48	1.02
0.20	0.73	0.73	0.73	0.73	0.16	3.65	3.71	22.48	22.86	1.02
0.30	0.86	0.86	0.86	0.86	0.08	2.35	2.38	28.20	28.46	1.01
0.40	0.93	0.93	0.93	0.93	0.03	1.28	1.29	38.74	38.98	1.01
		$\widehat{\text{PNF}}$			var($\widehat{\text{PNF}}$)			ARE		
q	PNF true	R	CC	(R, Y)	R	CC	(R, Y)	CC/R	$(R, Y)/R$	(R, Y)/CC
$\alpha = 6.55,\ \beta = 124.45^{\dagger}$										
0.90	0.60	0.60	0.60	0.60	0.07	6.11	6.15	92.04	92.64	1.01
0.80	0.45	0.45	0.45	0.45	0.06	4.52	4.51	73.59	73.40	1.00
0.70	0.34	0.34	0.34	0.34	0.05	3.34	3.34	66.08	66.02	1.00
0.60	0.26	0.26	0.26	0.26	0.04	2.42	2.42	60.15	60.15	1.00
$\alpha = 1,\ \beta = 19.0^{\ddagger}$										
0.90	0.60	0.60	0.60	0.60	0.06	6.11	6.07	95.58	94.83	0.99
0.80	0.45	0.45	0.45	0.45	0.06	4.52	4.54	74.24	74.62	1.01
0.70	0.34	0.34	0.35	0.35	0.05	3.34	3.36	62.44	62.80	1.01
0.60	0.26	0.26	0.26	0.26	0.04	2.42	2.43	59.11	59.46	1.01
$\alpha = 0.3,\ \beta = 5.7^{\S}$										
0.90	0.35	0.35	0.38	0.38	0.11	3.58	3.65	32.19	32.81	1.02
0.80	0.25	0.25	0.26	0.26	0.08	1.89	1.93	24.23	24.75	1.02
0.70	0.18	0.18	0.19	0.19	0.05	1.14	1.14	21.79	21.79	1.00
0.60	0.13	0.13	0.14	0.14	0.04	0.71	0.72	20.15	20.33	1.01

Results are based on 500 simulations for each set of parameters (α, β) for the beta distribution and values of q and p. Each simulation has $N = 10\,000$ samples with $\mu = 0.05$. AREs are computed as the ratio of the influence function-based variances.

ARE = asymptotic relative efficiency, the ratio of the influence functions-based variances CC$/R = \text{var}(T_{\text{CC}})/\text{var}(T_R)$, $(R, Y)/R = \text{var}(T_{(R,Y)})/\text{var}(T_R)$, $(R, Y)/\text{CC} = \text{var}(T_{(R,Y)})/\text{var}(T_{\text{CC}})$, where $T =$ PCF or $T =$ PNF, respectively.

†AUC = 0.63.

‡AUC = 0.79.

§ AUC = 0.92.

Table 2. *Mean values of iPCF and iPNF estimated using: the population risk distribution $F(R)$ assuming that the model is well calibrated; risks in a case–control study and known disease prevalence μ; observations of (R, Y) in the population*

		$\widehat{\text{iPCF}}$			$\text{var}(\widehat{\text{iPCF}})$			ARE		
$p^\dagger$	iPCF true	R	CC	(R, Y)	R	CC	(R, Y)	CC/R	(R,Y)/R	(R, Y)/CC
$\alpha = 6.55,\ \beta = 124.45^\dagger$										
0.10	0.60	0.60	0.60	0.60	0	1.26	1.26	279.48	278.91	1.00
0.20	0.57	0.57	0.57	0.57	0	1.20	1.20	311.36	311.83	1.00
0.30	0.53	0.53	0.53	0.53	0	0.84	0.84	316.73	316.32	1.00
0.40	0.48	0.48	0.48	0.48	0	0.49	0.49	271.95	271.77	1.00
$\alpha = 1,\ \beta = 2.33^\ddagger$										
0.10	0.73	0.73	0.73	0.73	0.02	0.81	0.81	46.58	46.54	1.00
0.20	0.68	0.68	0.68	0.68	0.01	0.59	0.59	59.61	59.52	1.00
0.30	0.63	0.63	0.63	0.63	0.01	0.39	0.39	68.36	68.90	1.01
0.40	0.55	0.55	0.55	0.55	0	0.24	0.25	69.42	69.52	1.00
		$\widehat{\text{iPNF}}$			$\text{var}(\widehat{\text{iPNF}})$			ARE		
$q^\dagger$	iPNF true	R	CC	(R, Y)	R	CC	(R, Y)	CC/R	(R,Y)/R	(R, Y)/CC
$\alpha = 6.55,\ \beta = 124.45^\dagger$										
0.60	0.27	0.27	0.27	0.27	0.0014	0.44	0.44	311.04	310.85	1.00
0.70	0.22	0.22	0.22	0.22	0.0006	0.23	0.23	358.61	358.83	1.00
0.80	0.16	0.16	0.16	0.16	0.0003	0.10	0.10	353.88	353.58	1.00
0.90	0.09	0.09	0.09	0.09	0	0.02	0.02	506.42	506.83	1.00
$\alpha = 1,\ \beta = 19.0^\ddagger$										
0.60	0.19	0.19	0.19	0.19	0.01	0.49	0.49	58.48	58.39	1.00
0.70	0.16	0.16	0.16	0.16	0.01	0.27	0.27	53.43	53.46	1.00
0.80	0.12	0.12	0.12	0.12	0.0002	0.15	0.15	77.24	77.47	1.00
0.90	0.07	0.07	0.07	0.07	0	0.05	0.05	91.55	91.58	1.00
$\alpha = 0.3,\ \beta = 5.7^\ddagger$										
0.50	0.12	0.12	0.12	0.12	0.01	0.29	0.30	20.92	21.71	1.04
0.60	0.11	0.11	0.11	0.11	0.01	0.23	0.23	19.70	20.19	1.02
0.70	0.10	0.10	0.10	0.10	0.01	0.19	0.19	26.88	26.77	1.00
0.80	0.08	0.08	0.08	0.08	0	0.11	0.11	27.62	27.53	1.00
0.90	0.05	0.05	0.05	0.05	0	0.05	0.05	46.03	45.81	1.00

Results are based on 500 simulations for each set of parameters (α, β) for the beta distribution and values of q and p. Each simulation has $N = 10\,000$ samples with $\mu = 0.05$. AREs are computed as the ratio of the influence function-based variances.

ARE = asymptotic relative efficiency, the ratio of the influence functions-based variances $\text{CC}/R = \text{var}(T_{\text{CC}})/\text{var}(T_R)$, $(R, Y)/R = \text{var}(T_{(R,Y)})/\text{var}(T_R)$, $(R, Y)/\text{CC} = \text{var}(T_{(R,Y)})/\text{var}(T_{\text{CC}})$, where $T = \text{PCF}$ or $T = \text{PNF}$ respectively.

†AUC = 0.63.

‡AUC = 0.79.

§ AUC = 0.92.

$(\alpha_k + 1, \beta_k)$ for cases and parameters $(\alpha_k, \beta_k + 1)$ for controls, and Φ^{-1} is the inverse of the standard normal distribution.

Based on the random sample $(r_{i1}, r_{i2}, Y_i),\ i = 1, \ldots, N$, we computed the non-parametric estimates of T_{PCF}, T_{PNF}, T_{iPCF}, and T_{iPNF} using observed risks only as well as risks and outcome data in the population, with the bootstrap variance estimates based on $B = 500$ bootstrap samples. A standard way to assess the discriminatory ability of two models is to compare their AUC or partial area under the curve (pAUC) values. When the test statistics were estimated from risks and outcome data, we computed non-parametric estimates of pAUC,

$$\widehat{\mathrm{pAUC}}(p) = \frac{1}{mn}\sum_{i,j} I\{r_i^G > r_j^K, r_j^K \in (\hat{\phi}_{1-p}, 1)\} + 0.5I\{r_i^G = r_j^K, r_j^K \in (\hat{\phi}_{1-p}, 1)\},$$

and also used the test statistic

$$T_{\mathrm{pAUC}} = \{\widehat{\mathrm{pAUC}}_1(p) - \widehat{\mathrm{pAUC}}_2(p)\}^2/\widehat{\mathrm{var}}(\widehat{\mathrm{pAUC}}_1 - \widehat{\mathrm{pAUC}}_2).$$

Estimates of size and power were based on 100 simulations for each choice of parameter values. Each simulation is based on a random sample of $N = 1000$ bivariate risks. We show results for $\rho = 0.2$ and a common disease, $\mu = 0.3$, as other choices resulted in similar conclusions.

Tests based on PCF, iPCF, PNF, or iPNF had better power than T_{pAUC} for all settings in Table 3. Table 3 highlights again that estimates computed under the assumption of a well-calibrated model have better power than those relying on risks and outcome data. For example, for $(\alpha_1, \beta_1) = (1, 2.3)$ and $(\alpha_2, \beta_2) = (1.2, 2.8)$ the power ranged from 0.77 to 0.81 for iPCF_R, but was lower than 0.21 for $\mathrm{iPCF}_{(R,Y)}$ for all values of p, and from 0.79 to 0.82 for iPNF_R, and from 0.29 to 0.77 for $\mathrm{iPNF}_{(R,Y)}$; the power was lower for smaller values of q. Tests based on iPCF_R and iPNF_R had somewhat better power than tests based on PCF_R and PNF_R.

7. Data example

To illustrate the various estimates of PCF, PNF, iPCF, and iPNF, we used data from a validation study of a colorectal cancer (CRC) risk prediction model (Freedman *and others*, 2009) that was developed to aid decision making for colorectal cancer screening. The risk model R estimates the probability, or absolute risk (also called cumulative incidence), of the binary event defined as "developing CRC during the age interval $(a, b]$, given that one is alive and free of CRC at age a". Letting T denote the failure time, the absolute CRC risk R is

$$R(x, a, b) = P(T \in (a, b], \text{cause} = \text{CRC} \mid T > a) = \int_a^b h_{\mathrm{CRC}}(t, x)\mathrm{e}^{-\int_0^t h_{\mathrm{CRC}}(u,x)+h_{\mathrm{M}}(u,x)\,\mathrm{d}u}\,\mathrm{d}t,$$

where x denotes individual covariates, $h_{\mathrm{CRC}}(t, x)$ is the cause-specific hazard for CRC, and $h_{\mathrm{M}}(t, x)$ denotes the competing mortality hazard.

The validation data were from an independent sample of 108 057 women from a prospective cohort, the National Institutes of Health (NIH)-AARP Diet and Health Study (Park *and others*, 2009). For the ith woman in the validation cohort the starting age a_i of the prediction was her age at entry into the cohort, and the end of the prediction interval, b_i, the age the woman had at the sooner of the end of study or loss to follow-up. Note that b_i does not depend on whether the woman died or developed CRC. These events are accounted for in the calculation of absolute risk R. For the validation, we thus define the event as $Y_i = 1$ if woman i develops CRC in $(a_i, b_i]$ and $Y_i = 0$ otherwise. The mean follow-up time was 6.9 years, and 965 women were diagnosed with CRC during follow-up. For estimates of PCF, PNF, iPCF, and iPNF based on case–control data, the disease prevalence μ was the observed incidence of CRC in the validation cohort.

Based on risk predictions for all 108 057 women in the study, R had $\widehat{\mathrm{AUC}} = 0.605$. Estimates $\widehat{\mathrm{PCF}}_{\mathrm{CC}}$ and $\widehat{\mathrm{PCF}}_{(R,Y)}$ were basically identical, but the latter had slightly wider confidence intervals as the uncertainty from estimating μ is also accommodated (Table 4). For example, for $p = 0.10$, $\widehat{\mathrm{PCF}}_{\mathrm{CC}} = 0.178$ with

Table 3. *Power of the tests T_{pAUC}, T_{PCF}, T_{PNF}, T_{iPCF}, and T_{iPNF} for two beta-distributed risk models under the assumption of well-calibrated models, R_1 with parameters (α_1, β_1) and R_2 with parameters (α_2, β_2) for $q = 0.1$ for sample size $N = 1000$, with $\mu = 0.3$ and $c = \chi^2_{1,0.95} = 3.84$*

							Power for test based on					
								iPCF			iPNF	
	pAUC	PCF	PNF	iPCF	iPNF		PCF			PNF		
p, q	1, 2	1, 2	1, 2	1, 2	1, 2	pAUC	R	R	(R, Y)	R	R	(R, Y)
$(\alpha_1, \beta_1) = (0.3, 0.7)$, $(\alpha_2, \beta_2) = (0.5, 1.17)$												
0.1, 0.9	0.06, 0.04	0.31, 0.29	0.44, 0.53	0.88, 0.84	0.95, 0.96	0.86	0.89	1.00	0.94	1.00	1.00	1.00
0.2, 0.8	0.14, 0.11	0.56, 0.51	0.34, 0.40	0.93, 0.90	0.89, 0.91	0.92	0.99	1.00	0.97	1.00	1.00	0.99
0.3, 0.7	0.23, 0.19	0.75, 0.67	0.27, 0.32	0.97, 0.94	0.82, 0.85	0.90	1.00	1.00	0.91	1.00	1.00	0.96
0.4, 0.6	0.32, 0.28	0.87, 0.80	0.22, 0.25	0.98, 0.97	0.75, 0.77	0.90	1.00	1.00	0.80	1.00	1.00	0.92
$(\alpha_1, \beta_1) = (6.55, 15.3)$, $(\alpha_2, \beta_2) = (4.55, 10.62)$												
0.1, 0.9	0.01, 0.01	0.16, 0.17	0.82, 0.80	0.68, 0.70	0.99, 0.99	0.10	0.98	1.00	0.28	1.00	0.99	0.90
0.2, 0.8	0.04, 0.04	0.29, 0.31	0.69, 0.67	0.76, 0.77	0.97, 0.96	0.12	1.00	1.00	0.13	1.00	1.00	0.68
0.3, 0.7	0.08, 0.09	0.42, 0.44	0.58, 0.55	0.82, 0.84	0.93, 0.92	0.24	1.00	1.00	0.27	1.00	1.00	0.61
0.4, 0.6	0.13, 0.15	0.53, 0.55	0.47, 0.45	0.88, 0.89	0.88, 0.87	0.16	1.00	1.00	0.18	1.00	1.00	0.34
$(\alpha_1, \beta_1) = (6.55, 15.3)$, $(\alpha_2, \beta_2) = (8.55, 19.95)$												
0.1, 0.9	0.01, 0.01	0.16, 0.15	0.82, 0.83	0.68, 0.67	0.99, 0.99	0.03	0.86	0.98	0.08	0.94	0.91	0.88
0.2, 0.8	0.04, 0.04	0.29, 0.28	0.69, 0.71	0.76, 0.75	0.97, 0.97	0.10	0.96	0.99	0.10	0.97	0.93	0.69
0.3, 0.7	0.08, 0.08	0.42, 0.40	0.58, 0.59	0.82, 0.82	0.93, 0.93	0.11	0.94	0.97	0.14	0.96	0.95	0.46
0.4, 0.6	0.13, 0.13	0.53, 0.51	0.47, 0.49	0.88, 0.87	0.88, 0.89	0.10	0.98	0.99	0.09	0.99	0.99	0.21
$(\alpha_1, \beta_1) = (1, 2.3)$, $(\alpha_2, \beta_2) = (1.2, 2.8)$												
0.1, 0.9	0.03, 0.03	0.25, 0.24	0.64, 0.67	0.79, 0.78	0.98, 0.98	0.09	0.41	0.77	0.16	0.81	0.81	0.77
0.2, 0.8	0.08, 0.07	0.43, 0.41	0.50, 0.52	0.86, 0.85	0.93, 0.94	0.12	0.57	0.81	0.21	0.78	0.80	0.59
0.3, 0.7	0.15, 0.14	0.58, 0.56	0.40, 0.42	0.91, 0.90	0.88, 0.88	0.13	0.61	0.78	0.13	0.63	0.79	0.31
0.4, 0.6	0.23, 0.22	0.70, 0.68	0.31, 0.33	0.94, 0.94	0.81, 0.82	0.19	0.70	0.81	0.16	0.66	0.82	0.29
$(\alpha_1, \beta_1) = (1, 2.3)$, $(\alpha_2, \beta_2) = (1.3, 3.03)$												
0.1, 0.9	0.03, 0.02	0.25, 0.23	0.64, 0.68	0.79, 0.77	0.98, 0.98	0.24	0.81	0.99	0.31	0.99	0.98	0.89
0.2, 0.8	0.08, 0.07	0.43, 0.41	0.50, 0.54	0.86, 0.84	0.93, 0.94	0.25	0.84	1.00	0.25	1.00	1.00	0.71
0.3, 0.7	0.15, 0.14	0.58, 0.55	0.40, 0.43	0.91, 0.89	0.88, 0.89	0.24	0.89	0.96	0.33	0.93	0.98	0.54
0.4, 0.6	0.23, 0.21	0.70, 0.67	0.31, 0.34	0.94, 0.93	0.81, 0.83	0.30	0.96	0.97	0.25	0.93	0.98	0.42

Estimates are based on 100 simulations for each set of parameter values. Bootstrap variances were based on $B = 500$ replicates.

Table 4. *Estimates* $\widehat{\mathrm{PCF}}$, $\widehat{\mathrm{PNF}}$, $\widehat{\mathrm{iPCF}}$, *and* $\widehat{\mathrm{PNF}}$ (*with* 95% *bootstrap confidence intervals in parenthesis*) *based on the observed distribution of risks R and outcome data for CRC in AARP women*

p	PCF*	$\widehat{\mathrm{PCF}}_R$	$\widehat{\mathrm{PCF}}_{\mathrm{CC}}$	$\widehat{\mathrm{PCF}}_{(R,Y)}$	$\widehat{\mathrm{PCF}}_{R_c}$
0.10	0.178	0.245 (0.244, 0.246)	0.178 (0.155, 0.202)	0.178 (0.152, 0.205)	0.179
0.20	0.313	0.415 (0.413, 0.416)	0.313 (0.284, 0.342)	0.311 (0.277, 0.345)	0.307
0.30	0.423	0.551 (0.549, 0.552)	0.420 (0.387, 0.452)	0.420 (0.377, 0.462)	0.418
0.40	0.549	0.663 (0.662, 0.664)	0.547 (0.515, 0.579)	0.547 (0.500, 0.594)	0.518
q		$\widehat{\mathrm{PNF}}_R, \hat{q}_R$	$\widehat{\mathrm{PNF}}_{\mathrm{CC}}$	$\widehat{\mathrm{PNF}}_{(R,Y)}, \hat{q}_{(R,Y)}$	$\widehat{\mathrm{PNF}}^*_{R_c}, \hat{q}_{R_c}$
0.90		0.711 (0.710, 0.712), 0.832	0.797 (0.784, 0.810)	0.797 (0.737, 0.858), 0.898	0.857, 0.927
0.80		0.556 (0.555, 0.557), 0.704	0.660 (0.641, 0.679)	0.661 (0.600, 0.723), 0.800	0.725, 0.839
0.70		0.438 (0.437, 0.439), 0.585	0.543 (0.518, 0.569)	0.545 (0.490, 0.599), 0.691	0.602, 0.751
0.60		0.342 (0.341, 0.343), 0.473	0.446 (0.418, 0.473)	0.447 (0.393, 0.501), 0.597	0.488, 0.635
p		$\widehat{\mathrm{iPCF}}_R$	$\widehat{\mathrm{iPCF}}_{\mathrm{CC}}$	$\widehat{\mathrm{iPCF}}_{(R,Y)}$	$\widehat{\mathrm{iPCF}}^*_{R_c}$
0.10		0.667 (0.666, 0.668)	0.592 (0.589, 0.596)	0.592 (0.575, 0.609)	0.572
0.20		0.634 (0.633, 0.634)	0.568 (0.565, 0.572)	0.567 (0.552, 0.582)	0.547
0.30		0.585 (0.585, 0.586)	0.530 (0.527, 0.533)	0.530 (0.517, 0.543)	0.511
0.40		0.524 (0.524, 0.525)	0.482 (0.479, 0.486)	0.483 (0.472, 0.493)	0.464
q		$\widehat{\mathrm{iPNF}}_R$	$\widehat{\mathrm{iPNF}}_{\mathrm{CC}}$	$\widehat{\mathrm{iPNF}}_{(R,Y)}$	$\widehat{\mathrm{iPNF}}^*_{R_c}$
0.90		0.083 (0.082, 0.083)	0.086 (0.084, 0.089)	0.086 (0.084, 0.089)	0.093
0.80		0.145 (0.145, 0.146)	0.160 (0.154, 0.165)	0.160 (0.154, 0.165)	0.172
0.70		0.195 (0.195, 0.195)	0.216 (0.207, 0.224)	0.216 (0.208, 0.225)	0.238
0.60		0.234 (0.233, 0.234)	0.268 (0.258, 0.278)	0.269 (0.259, 0.279)	0.293

The corresponding observed proportion of cases PCF* found among the fraction q of the AARP population at highest risk and the proportions p_R and p_{CC} of cases found among the fractions $\widehat{\mathrm{PNF}}_R$ and $\widehat{\mathrm{PNF}}_{\mathrm{CC}}$ of the AARP population at highest risk are also shown. $\widehat{\mathrm{PCF}}_{R_c}$ and $\widehat{\mathrm{PNF}}_{R_c}$ are based means over five test sets of the model after recalibration using 5-fold cross-validation.

95% CI (0.155, 0.202) and $\widehat{\mathrm{PCF}}_{(R,Y)} = 0.178$ (0.152, 0.205); thus, 17.8% of the cases were in the 10% of women at highest risk. However, for $p = 0.10$, $\widehat{\mathrm{PCF}}_R = 0.245$ (0.244, 0.246), which noticeably overestimated the observed PCF $= 0.178$. For PNF, a fraction $\hat{q}_{(R,Y)} = 0.898$ of cases was found in the fraction $\widehat{\mathrm{PNF}}_{\mathrm{CC}} = \widehat{\mathrm{PNF}}_{(R,Y)} = 0.797$ of the population with the highest risk when $q = 0.90$, showing an unbiased estimation of PNF. However, only a fraction $\hat{q}_R = 0.832$ of cases was found in the fraction $\widehat{\mathrm{PNF}}_R = 0.711$ of the population with highest risk when $q = 0.90$. Thus, only 83.2% of cases instead of the desired 90% had risks in the highest 71.1% of the population, reflecting poor calibration. Similarly, estimates $\widehat{\mathrm{iPCF}}_R$ were higher than $\widehat{\mathrm{iPCF}}_{\mathrm{CC}}$ and $\widehat{\mathrm{iPCF}}_{(R,Y)}$, and $\widehat{\mathrm{iPNF}}_R$ were lower than estimates $\widehat{\mathrm{iPNF}}_{\mathrm{CC}}$ and $\widehat{\mathrm{iPNF}}_{(R,Y)}$ (Table 4).

To illustrate the importance of good calibration for estimates of PCF, PNF, and iPCF and iPNF based on observed risks alone, we recalibrated the model and recalculated these estimates using 5-fold cross-validation. That is, we split the AARP cohort randomly into five equal-sized datasets, and used four of them to recalibrate the model by fitting a logistic regression model to observed CRC outcomes with the risk estimate r as the independent variable (Cox, 1958). This simple recalibration requires estimating only two parameters, the logistic intercept β_0 and slope β_1. We then used $\mathrm{logit}(r_c) = \beta_0 + \beta_1 r$ to predict CRC outcomes for women in the remaining fifth of the data, the test set, to estimate the criteria with the recalibrated model. The last column of Table 4 shows averages of estimates of PCF, PNF, iPCF, and iPNF over the five test sets. After recalibration $\widehat{\mathrm{PCF}}_{R_c}$ was less biased, e.g. $\widehat{\mathrm{PCF}}_{R_c} = 0.179$ for $p = 0.10$. Similarly, a fraction $\hat{q}_{R_c} = 0.927$ of cases was found in the fraction $\widehat{\mathrm{PNF}}_{R_c} = 0.857$ of women at highest risk when $q = 0.90$, reflecting improved calibration of R_c. Estimates $\widehat{\text{iPCF}}_{R_c}$ and $\widehat{\text{iPNF}}_{R_c}$ were also noticeably closer to $\widehat{\text{iPCF}}_{\text{CC}}$ and $\widehat{\text{iPNF}}_{\text{CC}}$.

8. Discussion

We proposed two new criteria for model evaluation, iPCF and iPNF, respectively, which lessen the dependency of earlier criteria, PCF and PNF, on prespecified thresholds. $\text{PCF}(p) = 1 - G \circ F^{-1}(1 - p)$ resembles the ROC value, which can be expressed as $\text{ROC}(p) = 1 - G \circ K^{-1}(1 - p)$ (Huang and Pepe, 2009). Similarly, iPCF that also is based on comparing the distribution of risk in cases to the distribution of risk in the whole population instead of in non-cases, relates to the AUC, the partial area under the ROC curve, pAUC, or more generally, to the weighted area under the ROC curve (Li and Fine, 2010). For rare diseases, they tend to be very similar, but derivations of the asymptotic properties for criteria based on PCF and PNF are more involved, as unlike the estimates of the ROC curve and the (partial) AUC, the risks in the population and in cases are not independent. The comparison of risk in cases to non-cases is appropriate for diagnostic tests that are applied in a clinical setting. However, for risk models that may be used to identify high-risk individuals for screening or for assessing the impact of a screening program in a population, criteria based on comparing the risk of cases to the risk in the whole population are more relevant (Pharoah *and others*, 2002).

While decision making based on risk models for public health applications ultimately needs to incorporate cost considerations, the proposed criteria can aid in the initial assessment of the feasibility of a screening or intervention program. For example, assume that one can only afford to screen 10% of a population. If a particular model has a low PCF or high PNF, targeting those at highest risk based on that model will have limited preventive impact. In contrast, if a model has a high PCF or a low PNF, then a targeted screening program may identify a large proportion of the disease and reduce costs and the burden of screening. A more complete understanding is provided by iPCF and iPNF, which display the proportion of disease accounted for by cumulative proportions of individuals in the population ranked from the lowest to highest risk.

The new criteria are also useful for comparing two risk models evaluated on the same dataset. A test for comparing two models based on PCF had comparable power to a test based on pAUC. However, a test for model comparison based on iPCF had significantly better power than a test based on PCF, while the tests based on PNF and iPNF had similar power.

We also studied estimates of PCF, PNF, iPCF, and iPNF when either risk estimates alone, or risk estimates and outcomes in a case–control study with known prevalence or in a cohort study are available. Estimates that also used outcome data were less efficient than estimates that were based on only observed risks and the assumption that the model was well calibrated. The efficiency gain comes from the fact that, for a well-calibrated model, knowing F implies knowing G, the distribution of risk in the cases. All the observed risks in a population are thus used to estimate G. When one estimates G from the risks in observed cases in a cohort, i.e. based on (R, Y), the effective sample size is much reduced, leading to substantial losses in efficiency. However, as also highlighted by our real example, estimates of PCF, PNF, iPCF, and iPNF based on R alone are biased when the model is not well calibrated, and model comparisons can be misleading. In practice, if outcome data are not available, one is forced to use estimates such as $\widehat{\text{iPCF}}_R$. If outcome data are available, one can compute $\widehat{\text{iPCF}}_{(R,Y)}$. If there are large discrepancies, one must suspect miscalibration and rely on $\widehat{\text{iPCF}}_{(R,Y)}$. We thus recommend using estimates of the criteria based on R alone, but comparing them to estimates that also use Y to check unbiasedness.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The author thanks Mitchell Gail for useful discussions and the reviewers for helpful suggestions. *Conflict of Interest*: None declared.

Funding

Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute.

References

Albertsen, P. C., Hanley, J. A. and Fine, J. (2005). Twenty-year outcomes following conservative management of clinically localized prostate cancer. *Journal of the American Medical Association* **293**, 2095–2101.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika* **45**, 562–565.

Freedman, A. N., Slattery, M. L., Ballard-Barbash, R., Willis, G., Cann, B. J., Pee, D., Gail, M. H. and Pfeiffer, R. M. (2009). A colorectal cancer risk assessment tool. *Journal of Clinical Oncology* **27**, 686–693.

Gini, C. (1912). Variabilita e mutabilita. *Studi Economico-Giuridici dell'Universita di Cagliari*, Volume 3, pp. 1–158.

Goldie, C. M. (1977). Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Applied Probability* **9**, 765–791.

Huang, Y. and Pepe, M. S. (2009). A parametric ROC model-based approach for evaluating the predictiveness of continuous markers in case-control studies. *Biometrics* **65**, 1133–1144.

Li, J. and Fine, J. P. (2010). Weighted area under the receiver operating characteristic curve and its application to gene selection. *Journal of the Royal Statistical Society. Series C: Applied Statistics* **59**, 673–692.

Mcclish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* **9**, 190–195.

O'Brien, T. R., Everhart, J. E., Morgan, T. R., Lok, A. S., Chung, R. T., Shao, Y., Shiffman, M. L., Dotrang, M., Sninsky, J. J., Bonkovsky, H. L., Pfeiffer, R. M. and HALT-C Trial Group (2011). An IL28B genotype-based clinical prediction model for treatment of chronic hepatitis C. *PLoS One* **6**, e20904.

Park, Y., Freedman, A. N., Gail, M. H., Pee, D., Hollenbeck, A., Schatzkin, A. and Pfeiffer, R. M. (2009). Validation of a colorectal cancer risk prediction model among white patients age 50 years and older. *Journal of Clinical Oncology* **27**, 694–698.

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

Pfeiffer, R. M. and Gail, M. H. (2011). Two criteria for evaluating risk prediction models. *Biometrics* **67**, 1057–1065.

Pharoah, P. D. P., Antoniou, A., Bobrow, M., Zimmern, R. L., Easton, D. F. and Ponder, B. A. J. (2002). Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genetics* **31**, 33–36.

Pires, A. M. and Branco, J. A. (2002). Partial influence functions. *Journal of Multivariate Analysis* **83**, 451–468.

Stephenson, A. J., Scardino, P. T., Eastham, J. A., Bianco, F. J., Dotan, Z. A., Fearn, P. A. and Kattan, M. W. (2006). Preoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *Journal of the National Cancer Institute* **98**, 715–717.

[*Received May 17, 2012; revised August 28, 2012; accepted for publication August 28, 2012*]