Published in final edited form as:

Biometrics. 2011 September ; 67(3): 1057–1065. doi:10.1111/j.1541-0420.2010.01523.x.

Two criteria for evaluating risk prediction models

R.M. Pfeiffer* and M.H. Gail

Biostatistics Branch, National Cancer Institute, 6120 Executive Blvd, Bethesda, MD 20892, USA

SUMMARY

We propose and study two criteria to assess the usefulness of models that predict risk of disease incidence for screening and prevention, or the usefulness of prognostic models for management following disease diagnosis. The first criterion, the proportion of cases followed $PCF(q)$, is the proportion of individuals who will develop disease who are included in the proportion q of individuals in the population at highest risk. The second criterion is the proportion needed to follow-up, $PNF(p)$, namely the proportion of the general population at highest risk that one needs to follow in order that a proportion p of those destined to become cases will be followed. $PCF(q)$ assesses the effectiveness of a program that follows 100 q % of the population at highest risk. $PNF(p)$ assess the feasibility of covering 100 p % of cases by indicating how much of the population at highest risk must be followed. We show the relationship of those two criteria to the Lorenz curve and its inverse, and present distribution theory for estimates of PCF and PNF . We develop new methods, based on influence functions, for inference for a single risk model, and also for comparing the $PCFs$ and $PNFs$ of two risk models, both of which were evaluated in the same validation data.

Keywords

Discrimination; Discriminatory accuracy; Lorenz curve; Risk models

1. Introduction

Statistical models that predict disease incidence, disease recurrence or mortality following disease onset have broad public health and clinical applications. Disease incidence models are available for many diseases, including coronary heart disease (Wilson et al., 1998), colorectal cancer (Freedman et al., 2009) and other cancers (http://riskfactor.cancer.gov/cancer_risk_prediction/). Disease incidence models can be used to target preventive interventions to those with high enough risks to justify an intervention that has adverse effects (Gail et al., 1999) and to identify high risk individuals for intensive screening to detect disease in its early stages. Prognostic models are useful for understanding the risk of cancer recurrence (e.g. Stephenson et al., 2006) or death following cancer diagnosis (e.g. Albertsen et al., 2005). Such information can inform choices for a treatment whose side effects may outweigh the risk of death from the initial disease.

Before a model can be recommended for practical use, its performance characteristics need to be understood. General criteria to evaluate prediction models for dichotomous outcomes include predictive accuracy, proportion of variation explained, calibration and

*pfeiffer@mail.nih.gov.

Supplementary Materials

The Web Appendix referenced in Sections 2 and 3 is available under the Paper Information link at the Biometrics website <http://www.tibs.org/biometrics>.

discrimination (Gail and Pfeiffer, 2005; Hand, 1981; Pepe, 2003). Most recent validation studies have emphasized calibration and discrimination. A model is called well calibrated (or unbiased) when the predicted probabilities agree with observed risk in subsets of the population and overall. Discriminatory accuracy, which assesses how much the distributions of risk predictions differ among those who subsequently did or did not develop disease, is often defined as the area under the receiver operating characteristic (ROC) curve (*AUC*) (see Pepe, 2003, page 67). Non-parametric inference on the *AUC* is based on the Mann-Whitney-Wilcoxon test.

We propose two measures of concentration of risk that are directly relevant to public health decisions. Only if most of the total population risk is concentrated in a small high risk subset will a prevention strategy that focuses on high risk groups be effective (Rose, 1992). If risk is concentrated, *AUC* will typically be large, but the reverse is not necessarily true (see Discussion). We define the “proportion of cases followed”, *PCF*(*q*), as the proportion of cases that would be followed in a program that followed the proportion *q* of the population at highest risk. If risk is concentrated in a small proportion of the population at highest risk, then *PCF*(*q*) will be high, even for small *q*. Pharoah et al. (2002) used this quantity to measure the usefulness of a risk model, but did not give methods of inference. We propose a new complementary criterion, the proportion needed to follow-up, *PNF*(*p*), namely the proportion of the general population at highest risk that one needs to follow in order that a proportion *p* of those destined to become cases will be followed. If risk is concentrated in a small proportion of the population at highest risk, *PNF*(*p*) will be small, even for large *p*.

By recognizing the relationship of *PCF* and *PNF* with the Lorenz curve of the distribution of risks in the population, we are able to draw on theory for the Lorenz curve and its inverse to derive their distribution theory. We develop new methods based on influence functions for inference for a single risk model, and also for comparing the correlated estimates of *PCF* and *PNF* for two risk models evaluated in the same validation data. In Section 2 we present the general framework, formally define *PCF* and *PNF*, and give nonparametric estimates for *PCF* and *PNF* with their asymptotic properties. We present results for simulated data (Section 3), a data example (Section 4) and discussion (Section 5).

2. Background, Notation, Criteria and Inference

2.1 Background and Notation

We present notation and methods for disease incidence models, but they also apply to models of disease prognosis. We want to predict whether an individual will be diagnosed with a particular disease over a given time period, for example five years ($Y_i = 1$) or not ($Y_i = 0$). Individual *i* has a true, unknown probability of the event, $\pi_i = P(Y_i = 1)$. Given a set of baseline predictors X_i for individual *i*, we approximate π_i by a risk prediction model $R(x_i) = P(Y_i = 1 | X_i = x_i)$, that is a mapping from the set Ω of possible values of *X* to $[0, 1]$. The distribution *F* of risk *R* in the population can be derived from the distribution of the covariates $F_X(x)$,

$$F(r) = \int_{\{x: R(x) \leq r\}} dF_X(x). \quad (1)$$

Because *X* includes continuous variables such as age, and length of the risk projection interval, we assume throughout that *F* is continuous, even though some results only require continuity at selected quantiles of *F*. From Gail and Pfeiffer (2005), we assume that the triplet (Y_i, π_i, X_i) arises from a superpopulation with unknown distribution that can be written as $\pi^Y (1 - \pi)^{1-Y} F(\pi|X) F_X(X)$, where *Y* is conditionally independent of *X* given π , and $F(\pi|X)$ is the unknown, and unobservable conditional distribution of the true, random probability of

disease, π , given the observed covariates X . We assume the model is perfectly calibrated, i.e. $R(x) = E(Y|x) = E(E(Y|x, \pi)|x) = E(\pi|x)$, as $E(Y|x, \pi) = E(Y|\pi) = \pi$. Thus, the overall disease prevalence satisfies $P(Y=1)=E(R)=\int_0^1 r dF(r) \equiv \mu$ and the distribution of risk in those who will and will not develop disease at the end of follow-up can be derived from the distribution F of risk in the population (Gail and Pfeiffer, 2005). Using that $P(R \leq r, Y=1)=\int_0^r t dF(t)$, the distribution of risk in diseased individuals ($Y = 1$) is given by

$$G(r)=P(R \leq r|Y=1)=\frac{1}{\mu}\int_0^r t dF(t). \quad (2)$$

Similarly, the distribution of risk in non-diseased persons ($Y = 0$) is given by

$$P(R \leq r|Y=0)=\frac{1}{1-\mu}\int_0^r (1-t) dF(t). \quad (3)$$

2.2 Criteria

Suppose the risk model R can be applied to an entire population to assign a disease risk r to each person. The model-based risks are then ranked, and a proportion of those individuals with highest risks of developing disease is followed up. Depending on the specific application, follow-up could consist of a diagnostic test, screening, or a preventive intervention. For example, the risk of colorectal cancer could be computed based on a risk model, and then a proportion of those with highest risk would receive more intensive screening with colonoscopy. Another example of “follow-up” arises in high risk preventive strategies (Rose, 1992) in which a proportion of the population at highest risk is given a preventive intervention, such as statins to prevent heart disease. In the context of prognosis after disease diagnosis, those at highest risk of dying might be followed-up by giving a special treatment.

Proportion of cases followed (PCF)—The proportion of cases followed, $PCF(q)$, is the proportion of individuals who will develop disease who are included in the proportion q of individuals in the population at highest risk. Let ξ_{1-q} denote the $(1-q)$ th quantile of the population distribution F , that is $P(r > \xi_{1-q}) = 1 - F(\xi_{1-q}) = q$. To obtain $PCF(q)$ we apply (2) to the corresponding population quantile, $\xi_{1-q} = F^{-1}(1 - q)$, and get

$$PCF(q)=1 - G(\xi_{1-q})=1 - G \circ F^{-1}(1 - q)=1 - \frac{1}{\mu}\int_0^{\xi_{1-q}} t dF(t). \quad (4)$$

Note that after a change of variable,

$$G(\xi_{1-q})=\frac{1}{\mu}\int_0^{\xi_{1-q}} t dF(t)=\frac{1}{\mu}\int_0^{1-q} F^{-1}(x) dx=L(1 - q), \quad (5)$$

which is also the Lorenz curve L of F evaluated at $1 - q$. Thus $PCF(q) = 1 - L(1 - q)$, where $L(1 - q)$ is the ratio of the total population risk below the $(1 - q)$ th quantile of F to the total risk in the population (Gastwirth, 1971; Goldie, 1977). Figure 1 (a) plots the Lorenz curve when F is a Beta(1,49) distribution for which $PCF(0.1) = 0.325$.

Proportion needed to follow-up (PNF)—The proportion needed to follow-up, $PNF(p)$, is the fraction of the general population at highest risks one needs to follow to assure that a

fraction p of those destined to become cases will be followed. Hence the quantile $\xi_{1-PNF} = F^{-1}(1 - PNF)$ satisfies

$$1 - G(\xi_{1-PNF(p)}) = 1 - G \circ F^{-1}(1 - PNF(p)) = p.$$

Thus, from (5),

$$PNF(p) = 1 - L^{-1}(1 - p). \quad (6)$$

Using a result by Goldie (1977), who called expression (2), $G(t) = \mu^{-1} \int_0^t x dF(x)$, the “first-order moment distribution”, we can also calculate $PNF(p)$ as

$$PNF(p) = 1 - \mu \int_0^{1-p} \{1/G^{-1}(x)\} dx, \quad 0 \leq p \leq 1. \quad (7)$$

Figure 1 (b) plots the Lorenz curve and shows the corresponding value of $PNF(0.9) = 0.591$ when F is Beta(1,49).

2.3 Estimation and Inference

Assume that a validation cohort contains a random sample of individuals with measured covariates. Corresponding estimates $r_1, \dots, r_n$ are a random sample from the continuous risk distribution F . We then can find nonparametric estimates of PCF and PNF as follows. Let $r_{(1)} \leq \dots \leq r_{(n)}$ denote the order statistics of the estimated risks, and $[x]$ be the largest integer

less or equal to x . Let $S_i = \sum_{k=1}^i r_{(k)}$. An estimate of the Lorenz curve and thus $PCF(q)$ is

$$\widehat{PCF}(q) = 1 - L_n(1 - q) = 1 - S_{[n(1-q)]}/S_n.$$

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

$$\widehat{PNF}(p) = 1 - L_n^{-1}(1 - p) = i/n, \quad S_i/S_n < 1 - p \leq S_{i+1}/S_n, i = 0, \dots, n. \quad (8)$$

In what follows, we often suppress the arguments q and p for ease of exposition.

For continuous F with finite mean, $L_n(1 - q)$ converges almost surely to $L(1 - q)$ and hence $\widehat{PCF}$ converges almost surely to PCF (Goldie, 1977). As $0 \leq r_i \leq 1$, F has finite second moments, and for a fixed q ,

$$\sqrt{n}\{\widehat{PCF}(q) - PCF(q)\} = \sqrt{n}\{L(1 - q) - L_n(1 - q)\} \rightarrow N(0, \sigma_{PCF}^2). \quad (9)$$

In the Web Appendix we derive the influence function $\phi(r)$ for L_n ,

$$\phi(r) = -\frac{1}{\mu}(r - \mu)L(q) - \frac{1}{\mu}[F^{-1}(q) - r]I\{r \leq F^{-1}(q)\} + \frac{1}{\mu}F^{-1}(q)q - L(q). \quad (10)$$

and use it to compute $\sigma_{PCF}^2 = E\phi^2$, that is given in the Web Appendix. By substituting empirical estimates L_n , F_n and $\hat{\mu} = (1/n) \sum r_i$, we can estimate the influence of risk estimate r_i as $\hat{\phi}(r_i)$. Then σ_{PCF}^2 can be estimated as $\hat{\sigma}_{PCF}^2 = (1/n) \sum \hat{\phi}^2(r_i)$. In Section 3, however, we estimate σ_{PCF}^2 using a bootstrap procedure.

Almost sure convergence of $\widehat{PNF}$ to PNF for a given p holds for continuous F with finite mean (Goldie, 1977). For a fixed p , and assuming that a condition on the variation of the quantile process F^{-1} near 0 (equation 2.23 in Csörgö and Yu, 1999) holds,

$$\sqrt{n}\{\widehat{PNF}(p) - PNF(p)\} = \sqrt{n}\{L^{-1}(1-p) - L_n^{-1}(1-p)\} \rightarrow N(0, \sigma_{PNF}^2). \quad (11)$$

We evaluate $\sigma_{PNF}^2 = E\psi^2$ by deriving the influence function

$$\psi(r) = -L^{-1}(1-p) + \frac{r}{G^{-1}(1-p)} [(1-p) - I\{r \leq G^{-1}(1-p)\}] + I\{r \leq G^{-1}(1-p)\}, \quad (12)$$

and $E\psi^2$ (Web Appendix). Substitution of empirical estimates for μ , L^{-1} and G^{-1} in equation (12) leads to $\hat{\psi}(r_i)$ and $\hat{\sigma}_{PNF}^2 = (1/n) \sum \hat{\psi}_i^2$. As before, we use a bootstrap estimate of the variance in numerical studies.

Motivated by the asymptotic theory, we study 95% confidence intervals with the bootstrap variance estimates, $\widehat{PCF} \pm 1.96\hat{\sigma}_{PCF}$ and $\widehat{PNF} \pm 1.96\hat{\sigma}_{PNF}$. The influence function representation is also valuable for computing the efficiency of the non-parametric estimates, compared to parametric estimates, and for comparing two models with respect to PCF and PNF , as shown next.

2.4 Comparing two risk models using PCF and PNF

AUC values are often used to compare the discriminatory ability of two models. The model with larger AUC value better separates the distributions of risk in cases and non-cases. For small values of q and high values of p that are practically relevant, risk is more concentrated at high levels for a model that has larger $PCF(q)$ or a smaller $PNF(p)$.

We assume that risk models 1 and 2 are evaluated in the same population, and we observe paired risk projections (r_i^1, r_i^2) for individuals $i = 1, \dots, n$ in the population, from the joint distribution function of risks (R^1, R^2) . We estimate $\widehat{PCF}^1$ and $\widehat{PNF}^1$, and $\widehat{PCF}^2$ and $\widehat{PNF}^2$ from r_i^1 and r_i^2 , $i = 1, \dots, n$, respectively.

To test whether, for fixed q , $PCF^1 = PCF^2$, we use the statistic

$$T_{PCF} = \frac{n(\widehat{PCF}^1 - \widehat{PCF}^2)^2}{\widehat{V}_{PCF}}. \quad (13)$$

We denote the influence for subject i for risk model 1 by ϕ_{i1} , and the corresponding influence for risk model 2 by ϕ_{i2} . The influence function representation

$\sqrt{n}(\widehat{PCF}^1 - \widehat{PCF}^2) = \sqrt{n} \sum_i (\phi_{1,i} - \phi_{2,i}) + o_p$, and application of the central limit theorem prove

asymptotic normality of $\sqrt{n}(\widehat{PCF}^1 - \widehat{PCF}^2)$, as $var(\phi_{1,i} - \phi_{2,i})$ exists because ϕ is bounded. We estimate $\widehat{V}_{PCF}$ using a bootstrap in simulations and the data example. For each bootstrap sample we draw n bivariate observations with replacement from (r_i^1, r_i^2) , $i = 1, \dots, n$, and then compute the empirical variance estimate based on the B bootstrap samples of the differences $\sqrt{n}(\widehat{PCF}_b^1 - \widehat{PCF}_b^2)$, $b = 1, \dots, B$. $\widehat{V}_{PCF}$ can also be estimated as

$$\widehat{var}(\widehat{\phi}_1 - \widehat{\phi}_2) = n \sum_{i=1}^n (\widehat{\phi}_{i1} - \widehat{\phi}_{i2})^2, \quad (14)$$

where $\widehat{\phi}_1$ and $\widehat{\phi}_2$ are obtained as in Section 2.3. Using asymptotic normality, 95% confidence intervals for the difference can be computed as $(\widehat{PCF}^1 - \widehat{PCF}^2) \pm 1.96(\widehat{V}_{PCF})^{1/2}$. Alternatively, confidence intervals for the differences could be computed based on the 0.025th and 0.0975th quantiles from the bootstrap distribution. Similar remarks apply to

$$T_{PNF} = \frac{n(\widehat{PNF}^1 - \widehat{PNF}^2)^2}{\widehat{V}_{PNF}}, \quad (15)$$

where $\widehat{V}_{PNF}$ estimates $var(\psi_{1,i} - \psi_{2,i})$.

Asymptotically both T_{PCF} and T_{PNF} have a central χ_1^2 distribution under H_0 . Under the alternative, the non-centrality parameters are $\delta_{PCF} = n(PCF^1 - PCF^2)^2 / var(\phi_{i1} - \phi_{i2})$ and $\delta_{PNF} = n(PNF^1 - PNF^2)^2 / var(\psi_{i1} - \psi_{i2})$, respectively.

3. Simulations

3.1 Operating characteristics and efficiency of PCF and PNF

We used simulations to investigate the properties of the nonparametric estimates $\widehat{PCF}$ and $\widehat{PNF}$ defined in Section 2.3. We also used the influence functions in Section 2.2 to compute the variances of $\widehat{PCF}$ and $\widehat{PNF}$ and to compare their efficiencies to maximum likelihood estimates (MLEs) under a fully parametric model. We assumed that observed risks r_i , $i = 1, \dots, n$ are a random sample from F , a beta distribution with parameters α and β . The cumulative distribution function can be written as $F(x) = B(x, \alpha, \beta) / B(\alpha, \beta)$, where $B(\alpha, \beta)$ denotes the beta function and $B(x, \alpha, \beta)$ stands for the incomplete beta-function with parameters α and β . In this setting the distribution of risk in the cases, $G(x) = B(x, \alpha + 1, \beta) / B(\alpha + 1, \beta)$, and thus

$$PCF(q) = 1 - B(F^{-1}(1 - q, \alpha, \beta), \alpha + 1, \beta) / B(\alpha + 1, \beta),$$

and

$$PNF(p) = 1 - B(G^{-1}(p, \alpha + 1, \beta), \alpha, \beta) / B(\alpha, \beta).$$

To investigate the inefficiency of non-parametric estimates, we also estimated PCF and PNF by finding the MLEs $\hat{\alpha}$ and $\hat{\beta}$ based on r_i , $i = 1, \dots, n$ and then computing $PCF(\hat{\alpha}, \hat{\beta})$ and $PNF(\hat{\alpha}, \hat{\beta})$. Using the corresponding Fisher information matrix $I(\alpha, \beta)$ and applying the delta method yields the asymptotic variances of the MLEs $PCF(\hat{\alpha}, \hat{\beta})$ and $PNF(\hat{\alpha}, \hat{\beta})$. To provide

context for our parameter choices, we also compute the *AUC*. We study parameters $\alpha = 6.55$, 4.23 and 1 with corresponding $\beta = 320.95$, 207.27, and 49; the expected risk $E(R) = 0.02$ for each (α, β) pair. A 2% risk might be found, for example, for colorectal cancer over a ten year period. The *AUC* values for these choices of parameters are 0.609, 0.635 and 0.753, respectively.

Table 1 gives results for 1000 independent simulations each based on a random sample of size $n = 500$. Bootstrap variance estimates were based on $B = 1000$ replicates. The mean estimates of $\widehat{PCF}$ and $\widehat{PNF}$ were very close to the theoretical values, and the coverage of the confidence intervals was close to the nominal 0.95 value. The bootstrap estimates of the standard deviations of $n^{1/2}\widehat{PCF}$ and $n^{1/2}\widehat{PNF}$ agreed closely with the theoretical estimates based on the influence functions. For example, for $\widehat{PCF}$ with $(\alpha, \beta) = (4.23, 207.27)$ and $q = 0.10$, the theoretical calculation and mean bootstrap estimates were 0.114 and 0.111 respectively. For $q = 0.3$ the theoretical and mean bootstrap estimates of standard deviation were both 0.144. Similar good agreement was found for $\widehat{PNF}$. Very similar results were found for parameters that yielded the same *AUC* values as Table 1, but with $E(R) = 0.2$ (Web Table 1).

The model with the least discriminatory accuracy, $(\alpha, \beta) = (6.55, 320.95)$, detects only a fraction $PCF = 0.178$ of cases when the fraction $q = 0.10$ of those at highest risk is studied, whereas the model with the most discriminatory accuracy $(\alpha, \beta) = (1, 49)$, detects $PCF = 0.325$. Likewise, to follow a fraction $p = 0.90$ of cases requires that the fraction $PNF = 0.808$ of the general population at highest risk be followed for the least discriminating model and $PNF = 0.591$ for the most discriminating model studied.

The non-parametric estimates $\widehat{PCF}$ and $\widehat{PNF}$ were less precise than corresponding parametric estimates, as assessed by the asymptotic relative efficiency (ARE), or variance ratio. For small q , $\widehat{PCF}$ has an ARE less than 0.5, or a variance more than twice as large as the parametric estimates. For example, $ARE = 0.473$ for $(\alpha, \beta) = (4.23, 207.27)$ for $q = 0.10$. The AREs of $\widehat{PNF}$ compared to the MLE ranged from 0.682 to 0.895 over the parameters and choices of q studied. Even though the non-parametric procedures are less efficient, they yield consistent estimates, whereas incorrectly specified parametric methods need not.

3.2 Size and power of tests for comparing two risk models

We examined the size and power of tests T_{PCF} in equation (13) and T_{PNF} in equation (15) for comparing risk models 1 and 2 with data from the same validation sample. To obtain risk estimates from each model that have a marginal beta distribution and are correlated, we first generated n bivariate normal random variables $(X_{i1}, X_{i2}) \sim MVN((0, 0), \Sigma)$, $i = 1, \dots, n$, where $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = \rho$. We then computed $r_{i1} = F_1^{-1} \circ \phi^{-1}(X_{i1})$ 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 (α_k, β_k) , and ϕ^{-1} is the inverse of the standard normal distribution. Based on the random sample (r_{i1}, r_{i2}) , $i = 1, \dots, n$, we computed the non-parametric estimates of T_{PCF} and T_{PNF} with the bootstrap variance estimates based on $B = 1000$ bootstrap samples. Estimates of size and power were based on 1000 independent simulations for each choice of parameter values and use of T_{PCF} and T_{PNF} . Each simulation used a random sample of $n = 500$ bivariate risks.

The estimated correlations for the risk estimates (r_{i1}, r_{i2}) generated under this mechanism are close to the correlation used in the bivariate normal model (Table 2). For $\rho = 0.25$ the estimated correlations ranged from 0.19 to 0.29, and for $\rho = 0.50$ the estimated correlations ranged from 0.41 to 0.54 (Table 2).

Table 2 shows the size for T_{PCF} and T_{PNF} with nominal 5% significance level, for various choices of correlations, ρ , and common parameters for the two beta distributions. Theoretical size, $P_t(T_{PCF} > c)$ and $P_t(T_{PNF} > c)$, and empirical estimates, $P_e(T_{PCF} > c)$ and $P_e(T_{PNF} > c)$, are presented. Again, parameters were chosen so that $ER_1 = ER_2 = 0.02$. The tests had near nominal (0.05) size regardless of ρ or choice of beta distribution parameters (Table 2).

We compared the theoretical power $P_t(T_{PCF} > c)$ or $P_t(T_{PNF} > c)$ based on the influence-function calculation of non-centrality to power estimated from simulations for non-parametric estimates of T_{PCF} and T_{PNF} when risk estimates were generated from the normal-beta model. The estimated power agreed well with theoretical power estimates for T_{PCF} and T_{PNF} over a range of correlations ρ and marginal beta distributions of risk. Note that with a sample of $n = 500$ bivariate observations there is good power to detect subtle differences in PCF and PNF . For example, for $\rho = 0.10$, $(\alpha_1 = 6.55, \beta_1 = 320.95)$, $(\alpha_2 = 9.50, \beta_2 = 465.5)$, the estimated power was 0.834 to detect a difference in PCF of only $PCF^1 - PCF^2 = 0.178 - 0.163 = 0.015$. Likewise, for $\rho = 0.25$, the estimated power was 0.862 to detect a difference of $PCF^1 - PCF^2 = 0.806 - 0.826 = -0.02$.

For a few sets of parameters we compared the power of T_{PCF} and T_{PNF} with influence function based variance estimates to their power with the bootstrap variance estimates and found close agreement (unreported data).

4. Data Example

Freedman et al. (2009) developed a model to predict colorectal cancer (CRC) incidence. We now compute PCF and PNF for that model in independent data on 108,016 women from the National Institutes of Health (NIH)-AARP Diet and Health Study, a prospective cohort study (Park et al., 2009).

Let T denote the time to CRC. An individual's CRC risk over the age interval (a, b) is $r(x) = P(T \in (a, b) | T > a) = \int_a^b \lambda_1(t, X) S^*(t-) dt$, where $S^*(t) = \exp[-\int_0^t \{\lambda_1(u, X) + \lambda_2(u)\} du]$. Here λ_1 denotes the CRC hazard and λ_2 the hazard for competing causes of death. As CRC incidence differed among three different sites in the colon, Freedman et al. modeled $\lambda_1 = \lambda_{11} + \lambda_{12} + \lambda_{13}$, as the sum of proximal (λ_{11}), distal (λ_{12}) and rectal (λ_{13}) colon cancer hazard rates, with $\lambda_{1i} = RR_i(t, X) \lambda_{1i}^*(t)$, $i = 1, 2, 3$, where λ_{1i}^* denotes the age specific baseline hazard rate and RR_i the relative risk, which was estimated from case-control data. The $\lambda_{1i}^*(t)$ s were obtained by multiplying the composite hazards from SEER cancer registries by one minus the attributable risk, which was estimated from the relative risks and the risk factor distribution in cases in the case-control study. Risk factors included an individual's age, sex, sigmoidoscopy/colonoscopy history, polyps and other factors.

While the model was well calibrated overall among women in the AARP Diet and Health Study it showed lack of fit in some subgroups, possibly because the AARP questionnaire assessed sigmoidoscopy/colonoscopy history differently (Park et al., 2009). Because good estimates of PCF and PNF require good model calibration, we recalibrated the model. We first split the women in the AARP cohort randomly into a training and a test set, each comprised of 54,008 women, with 458 CRC cases in the training set and 467 in the test set. Following Cox (1958), we fit a logistic regression model with the log-transformed absolute risk estimate r as the independent variable to the training set. We then used $\text{logit}(r_c) = -4.7019 + 0.5335 \log(r)$ to predict disease outcomes for women in the test set. Estimates r_c were well calibrated overall and in groups defined by deciles of risk in the test set. The recalibration does not affect AUC , as r_c is a monotone transformation of r .

In the test data, the model r_c had $\widehat{AUC}=0.624$. The modest discriminatory ability was also reflected in $\widehat{PCF}$ and $\widehat{PNF}$. For example, for $q = 0.10$, $\widehat{PCF}=0.170$; thus 17% of the cases were predicted to be in the 10% of the population at highest risk. From observed disease outcome in the test data, we computed that the proportion of cases with risks above the 90th percentile of F was 0.180 (95% $CI : 0.145-0.215$). Similarly, for PNF , for $p = 0.90$, $\widehat{PNF}=0.808$; and in fact a fraction 0.914 with 95% $CI(0.884, 0.940)$ of cases were found in the fraction $\widehat{PNF}=0.808$ of the population with highest risk. Thus $\widehat{PCF}$ and $\widehat{PNF}$ were consistent with the observed outcome data in the test sample (Table 3).

We then compared the previous model (model 1) to model 2 that used a single hazard, $\lambda_1(t, X)=RR(X, t)\lambda_1^*(t)$. Model 2 was fit to the original case control data in Freedman et al. We applied the same recalibration to model 2, resulting in a logistic model, $\text{logit}(r_c^{(2)}) = -4.6959 + 0.5206 \log(r_2)$. Estimates $r_c^{(2)}$ were well calibrated in the test set.

Model 2 had $\widehat{AUC}_2=0.622$ in the test data, which was not significantly different from $\widehat{AUC}_1=0.624$ ($p = 0.23$). For $q = 0.1$, model 2 yielded a significantly lower $\widehat{PCF}_2=0.168$, 95% $CI(0.167, 0.168)$ than model 1 ($\widehat{PCF}_1=0.170$, $p < 0.0001$) (Table 3). Likewise, for $p = 0.9$, model 2 resulted in a significantly larger $\widehat{PNF}_2=0.811$, 95% $CI(0.810, 0.811)$ than model 1 ($\widehat{PNF}_1=0.808$, $p < 0.0001$). These small differences are statistically significant because the estimates are based on 54, 008 women in the test data, resulting in high precision.

Even though model 1 is statistically significantly superior to model 2, the practical improvement is small. For example, to assure that 90% of future cases are screened in the test cohort, model 1 would require screening of $0.808 \times 54, 008 = 43, 638$ women, compared to $0.811 \times 54, 008 = 43, 800$ for model 2.

5. Discussion

We considered two criteria for evaluating risk models that measure the concentration of risk in a population and that have potential application in public health and clinical epidemiology. $PCF(q)$, the proportion of cases who will be followed if the proportion q of the general population at highest risk is followed has been recommended by Pharoah et al. (2002). Hand and Henley (1997) used the “bad risk rate amongst accepts” that is analogous to $1 - PCF$, in credit risk models, but neither Pharoah et al. nor Hand and Henley described methods of inference. For models of disease incidence, PCF can determine the proportion of those destined to develop disease who will receive screening or preventive interventions. Following disease diagnosis, it can predict the proportion of all patients destined to have a bad outcome among patients selected for a treatment based on high risk. We introduced the criterion $PNF(p)$, namely the proportion of the population at highest risk that needs to be followed in order that a proportion p of cases be followed as a complimentary guide to screening or intervention applications. PNF could be adapted to high risk preventive interventions (Rose, 1992), such as deciding what proportion of a population should be given statins to assure that at least 80% of all persons destined to have a myocardial infarction shall have received statins beforehand. The quantity $1 - PCF(q)$ assesses what proportion of cases will fail to be followed if a proportion $1 - q$ at lowest risk is not followed. Thus, PCF is useful for evaluating the effectiveness of an ongoing or proposed screening or prevention program by determining what proportion of future cases shall have participated. A “high risk strategy” requires that q be small, say 20% or less. Whether that strategy will be effective in covering cases depends on whether $PCF(q)$ is large enough. PNF is useful in assessing the feasibility of the program required to cover a proportion p of future cases.

Typically, one will want p to be 80% or more. Whether such a p is feasible depends on whether $PNF(p)$ is small enough.

AUC , unlike PCF and PNF , does not measure risk concentration. For example, suppose that risk is zero in non-cases and uniformly distributed in $[0, 1]$ in cases, and that half of the population develops disease. Then $AUC = 1$, but $PCF(0.1) = 0.2$ and $PNF(0.9) = 0.45$.

By relating PCF and PNF to the Lorenz curve of the risk distribution and its inverse, we adapted and developed the theory needed for inference on these quantities for risks from a single model applied to the distribution of covariates in an independent validation sample. We developed methods for testing whether $PCF^1 = PCF^2$ or $PNF^1 = PNF^2$ for two models evaluated on the same independent sample with bivariate risk estimates (r_{1i}, r_{2i}) . Our methods allow for correlations between r_{1i} and r_{2i} . The work of Zheng and Cushing (2001) can be used to compare two $PCFs$ with dependent data, but we are unaware of similar results for PNF . Methods are available for comparing two Lorenz curves from independent samples (e.g. Dardanoni and Forcina, 1999). When using such tests, care should be taken to assure that each model is well calibrated. Otherwise, $PCF^1 > PCF^2$ or $PNF^1 < PNF^2$ may simply reflect miscalibration of one or both models. We plan to explore the sensitivity of our methods to miscalibration in further simulations in future research.

The ideal data for estimating PCF and PNF are from a random sample from the distribution of covariates in an independent population of interest. This sample can be used to derive the distribution of risks for one or more models. If random samples of cases ($Y = 1$) and controls ($Y = 0$) are available from a population, with corresponding distributions of risk G and K , and if disease is rare so that K nearly represents F , then $\widehat{PCF} = 1 - \widehat{G} \circ \widehat{K}^{-1}(1 - q)$ has a simple distribution theory because $\widehat{K}$ and $\widehat{G}$ are independent (Greenhouse and Mantel, 1950). The same is true for $\widehat{PNF} = 1 - \widehat{K} \circ \widehat{G}^{-1}(1 - p)$. If the disease is not rare but the disease risk μ is known, one can use $F = \mu G + (1 - \mu)K$ in expressions for PCF and PNF . However, the corresponding distribution theory is not developed.

Other criteria have been proposed to evaluate risk models, apart from calibration and the AUC. If loss functions can be specified, they can determine the optimal risk threshold for intervention, t^* , and indicate which risk model has smaller expected losses (Vickers and Elkin, 2006; Gail and Pfeiffer, 2005) or larger $PCF(q)$ for $q = 1 - F(t^*)$. Cook (2007) proposed reclassification criteria based on risk thresholds, and Pencina et al. (2008) proposed a related measure, the net reclassification index. Another criterion they recommended, the integrated discrimination improvement, is not tied to a particular risk threshold, and can be regarded as a global criterion, like the AUC. Pepe et al. (2008) point out that the ROC curve, which is a plot of $1 - G(r)$ against $1 - K(r)$ as r varies, suppresses information on r . Instead, they recommend use of the predictiveness curve, a plot r versus $F(r)$, together with another plot with two curves: $1 - G(r)$ versus $F(r)$ and $1 - K(r)$ versus $F(r)$. Together, these three curves summarize the information on the risk distribution in the population, $F(r)$, and the effects of various risk thresholds r on $P(R > r | Y = 1)$ and $P(R > r | Y = 0)$.

Huang and Pepe (2009) showed that $ROC(q) = 1 - G \circ K^{-1}(1 - q)$ and that $dROC(q)/dq$ can be used to estimate the predictiveness curve and the "total gain" statistic, introduced by Bura and Gastwirth (2001), if the disease prevalence μ is known. For rare diseases, $F \approx K$, and $PCF(q)$ approximates $ROC(q)$, but more work is needed to prove the conjecture that $PCF(q)$ and its derivative can be used to approximate the predictiveness curve and total gain.

The risk distribution F is central to the evaluation of risk models, because criteria such as expected loss, measures of relative dispersion of risk such as the Gini index, risk

distributions G and K , and the AUC are functionals of F (Gail and Pfeiffer, 2005), and because of its value in displaying information in the "integrated predictiveness and classification" plots of Pepe et al. (2008). We believe that two functionals of particular public health interest are $PCF(q)$ and $PNF(p)$.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the reviewers for many helpful suggestions.

References

- Albertsen PC, Hanley JA, Fine J. 20-year outcomes following conservative management of clinically localized prostate cancer. *Journal of the American Medical Association*. 2005; 293:2095–2101. [PubMed: 15870412]
- Bura E, Gastwirth JL. The binary regression quantile plot: Assessing the importance of predictors in binary regression visually. *Biometrical Journal*. 2001; 43:5–21.
- Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007; 115:928–935. [PubMed: 17309939]
- Cox DR. Two Further Applications of a Model for Binary Regression Two Further Applications of a Model for Binary Regression. *Biometrika*. 1958; 45:562–565.
- Csörgő M, Yu H. Weak approximations for empirical Lorenz curves and their Goldie inverses of stationary observations. *Advances in Applied Probability*. 1999; 31:698–719.
- Dardanoni V, Forcina A. Inference For The Lorenz Curve Ordering. *Econometrics Journal*. 1999; 2:48–74.
- Freedman AN, Slattery ML, Ballard-Barbash R, Willis G, Cann BJ, Pee D, Gail MH, Pfeiffer RM. Colorectal cancer risk prediction tool for white men and women without known susceptibility. *Journal of Clinical Oncology*. 2009; 27:686–693. [PubMed: 19114701]
- Gail MH, Gastwirth JL. Scale-free goodness-of-fit test for exponential distribution based on Lorenz curve. *Journal of the American Statistical Association*. 1978; 73:787–793.
- Gail MH, Costantino JP, Bryant J, Croyle R, Freedman L, Helzlsouer K, Vogel V. Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. *Journal of the National Cancer Institute*. 1999; 91:1829–1846. [PubMed: 10547390]
- Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risks. *Biostatistics*. 2005; 6:227–239. [PubMed: 15772102]
- Gastwirth JL. A General Definition of the Lorenz Curve. *Econometrica*. 1971; 39:1037–1039.
- Goldie CM. Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Applied Probability*. 1977; 9:765–791.
- Greenhouse SW, Mantel N. The evaluation of diagnostic tests. *Biometrics*. 1950; 6:399–412. [PubMed: 14791576]
- Hand, DJ. *Discrimination and classification*. Chichester: Wiley; 1981.
- Hand DJ, Henley WE. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society Series A*. 1997; 160:523–541.
- Huang Y, Pepe MS. A parametric ROC model-based approach for evaluating the predictiveness of continuous markers in case-control studies. *Biometrics*. 2009; 65:1133–1144. [PubMed: 19459841]
- Huber PJ. *Robust Statistical Procedures*. Society for Industrial and Applied Mathematics. 1996
- Park Y, Freedman AN, Gail MH, Pee D, Hollenbeck A, Schatzkin A, Pfeiffer RM. Validation of a Colorectal Cancer Risk Prediction Model Among White Patients Age 50 Years and Older. *Journal of Clinical Oncology*. 2009; 27:694–698. [PubMed: 19114700]

- Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*. 2008; 27(2):157–172. [PubMed: 17569110]
- Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series: Oxford University Press; 2003.
- Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*. 2008; 167:362–368. [PubMed: 17982157]
- Pharoah PDP, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BAJ. Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genetics*. 2002; 31:33–36. [PubMed: 11984562]
- Rose, G. *The Strategy of Preventive Medicine*. Oxford Medical Publications, Oxford University Press; 1992.
- Stephenson AJ, Scardino PT, Eastham JA, Bianco FJ, Dotan ZA, Fearn PA, Kattan MW. Preoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *Journal of the National Cancer Institute*. 2006; 98:715–717. [PubMed: 16705126]
- Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*. 2006; 26:565–574. [PubMed: 17099194]
- Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998; 97:1837–1847. [PubMed: 9603539]
- Zheng BH, Cushing BJ. Statistical inference for testing inequality indices with dependent samples. *Journal of Econometrics*. 2001; 101:315–335.

Figure 1.

Lorenz curve when the distribution F of risk is a Beta distribution with parameters $\alpha = 1$, $\beta = 49$ for $PCF(0.1)$ (Panel a) and $PNF(0.9)$ (Panel b). The vertical broken line in Panel a and the horizontal broken line in Panel b correspond to $q = 0.1$ and $L(1 - p) = 0.1$ respectively. $PCF(0.1)$ is the segment of the line $q = 0.1$ above the Lorenz curve in Panel a. $PNF(q)$ is the segment of the line $L(1 - 0.9) = L(0.1)$ to the right of its intersection with the Lorenz curve in Panel b.

Table 1

Mean values of $\widehat{PCF}$ and $\widehat{PNF}$ and of theoretical and bootstrap estimates of $\{nV \ar(\widehat{PCF})\}^{1/2}$ and $\{nV \ar(\widehat{PNF})\}^{1/2}$ with estimated coverage of 95% confidence intervals and asymptotic relative efficiency (ARE) compared to parametric analysis under the beta distribution. Results are based on 1000 simulations for each set of parameters (α, β) for the beta distribution and values of q and p . Each simulation has size $n = 500$ samples, and $B = 1000$ bootstrap replicates were used. All beta distributions have mean risk 0.02

α, β	q	PCF	$\widehat{PCF}$	standard error	coverage	ARE
$\sqrt{E\phi^2}$ bootstrap						
6.55, 320.95*	0.10	0.178	0.178	0.086	0.085	0.502
6.55, 320.95	0.20	0.317	0.317	0.106	0.107	0.694
6.55, 320.95	0.30	0.441	0.440	0.115	0.115	0.777
4.23, 207.27**	0.10	0.199	0.199	0.114	0.111	0.473
4.23, 207.27	0.20	0.349	0.348	0.139	0.136	0.634
4.23, 207.27	0.30	0.476	0.476	0.144	0.144	0.751
1, 49***	0.10	0.325	0.324	0.271	0.271	0.369
1, 49	0.20	0.517	0.516	0.290	0.288	0.505
1, 49	0.30	0.657	0.656	0.267	0.267	0.609
α, β	p	PNF	$\widehat{PNF}$	standard error	coverage	ARE
$\sqrt{E\psi^2}$ bootstrap						
6.55, 320.95	0.90	0.808	0.809	0.093	0.092	0.760
6.55, 320.95	0.80	0.671	0.672	0.108	0.110	0.873
6.55, 320.95	0.70	0.553	0.554	0.115	0.115	0.845
4.23, 207.27	0.90	0.778	0.779	0.118	0.121	0.846
4.23, 207.27	0.80	0.634	0.636	0.135	0.138	0.895
4.23, 207.27	0.70	0.515	0.517	0.139	0.141	0.862
1, 49	0.90	0.591	0.592	0.261	0.263	0.783
1, 49	0.80	0.442	0.444	0.255	0.254	0.682

α, β	q	PCF	$\widehat{PCF}$	standard error	coverage	ARE
				$\sqrt{E\phi^2}$	bootstrap	
1, 49	0.70	0.337	0.339	0.234	0.231	0.598

* AUC=0.609,
 ** AUC=0.635,
 *** AUC=0.753

Table 2

Power of the tests T_{PCF} and T_{PNF} to compare $PCF(q)$ and $PNF(p)$ for two beta-distributed risk models, R_1 with parameters (α_1, β_1) and R_2 with parameters (α_2, β_2) for $q = 0.1$ and $p = 0.9$ for sample size $n = 500$ and $c = \chi^2_{1,0.95} = 3.84$. Estimates are based on 1000 simulations for each set of parameter values. Bootstrap variances were based on $B = 1000$ replicates. All numbers are theoretical values, except for empirical estimates of risk correlations (column 2), and empirical estimates $P_e(T > c)$ of power (column 10). $P_l(T > c)$ stands for theoretical values based on influence function calculations of variance.

ρ	$\widehat{cor}(r_1, r_2)$	α_1, β_1	α_2, β_2	AUC_1	AUC_2	PCF_1	PCF_2	$P_l(T_{PCF} > c)$	$P_e(T_{PCF} > c)$
0.10	0.085	6.55, 320.95	6.55, 320.95	0.609	0.609	0.178	0.178	0.050	0.066
0.10	0.153	1.00, 49.00	1.00, 49.00	0.753	0.753	0.325	0.325	0.050	0.065
0.10	0.105	4.23, 207.27	4.23, 207.27	0.635	0.635	0.199	0.199	0.050	0.064
0.50	0.452	1.00, 49.00	1.00, 49.00	0.753	0.753	0.325	0.325	0.050	0.052
0.50	0.455	6.55, 320.95	6.55, 320.95	0.609	0.609	0.178	0.178	0.050	0.065
0.50	0.531	4.23, 207.27	4.23, 207.27	0.635	0.635	0.199	0.199	0.050	0.061
0.10	0.085	6.55, 320.95	8.50, 416.50	0.609	0.596	0.178	0.167	0.543	0.550
0.25	0.290							0.556	0.554
0.50	0.517							0.620	0.619
0.10	0.053	6.55, 320.95	9.50, 465.50	0.609	0.591	0.178	0.163	0.834	0.854
0.25	0.219							0.847	0.862
0.50	0.466							0.895	0.883
0.10	0.017	1.00, 49.00	1.30, 63.70	0.753	0.727	0.325	0.294	0.492	0.472
0.25	0.234							0.508	0.489
0.50	0.476							0.565	0.557
0.10	0.101	1.00, 49.00	1.40, 68.60	0.753	0.721	0.325	0.286	0.700	0.704
0.25	0.193							0.716	0.724
0.50	0.487							0.776	0.786
ρ	$\widehat{cor}(r_1, r_2)$	α_1, β_1	α_2, β_2	AUC_1	AUC_2	PNF_1	PNF_2	$P_l(T_{PNF} > c)$	$P_e(T_{PNF} > c)$
0.10	0.074	6.55, 320.95	6.55, 320.95	0.609	0.609	0.808	0.808	0.050	0.047
0.10	0.072	1.00, 49.00	1.00, 49.00	0.753	0.753	0.591	0.591	0.050	0.074
0.10	0.096	4.23, 207.27	4.23, 207.27	0.635	0.635	0.778	0.778	0.050	0.051

ρ	$\widehat{cor}(r_1, r_2)$	a_1, β_1	a_2, β_2	AUC_1	AUC_2	PCF_1	PCF_2	$P_c(T_{PCF} > c)$	$P_c(T_{PCF} > c)$
0.50	0.452	1.00, 49.00	1.00, 49.00	0.753	0.753	0.591	0.591	0.050	0.042
0.50	0.438	4.23, 207.27	4.23, 207.27	0.635	0.635	0.778	0.778	0.050	0.062
0.50	0.542	6.55, 320.95	6.55, 320.95	0.609	0.609	0.808	0.808	0.050	0.046
0.10	0.111	6.55, 320.95	8.50, 416.50	0.609	0.596	0.808	0.822	0.729	0.721
0.25	0.219							0.759	0.740
0.50	0.509							0.820	0.831
0.10	0.110	6.55, 320.95	9.50, 465.50	0.609	0.591	0.808	0.827	0.954	0.962
0.25	0.294							0.966	0.957
0.50	0.531							0.988	0.977
0.10	0.090	1.00, 49.00	1.30, 63.70	0.753	0.727	0.591	0.637	0.843	0.834
0.25	0.192							0.853	0.876
0.50	0.408							0.919	0.915

Table 3

Estimates $\widehat{PCF}$ and $\widehat{PNF}$ (with 95% confidence intervals in parenthesis) based on the observed distribution F of risks for colorectal cancer in AARP women. The corresponding observed proportion of cases PCF^* (with 95% confidence interval) found among the fraction q of the AARP population at highest risk and the proportion p^* of cases (with 95% confidence interval) found among the fraction $\widehat{PNF}$ of the AARP population at highest risk are also shown.

q	$\widehat{PCF}$	PCF^* in AARP test data
Model 1		
0.10	0.170 (0.170, 0.171)	0.180 (0.145, 0.215)
0.20	0.310 (0.310, 0.311)	0.323 (0.281, 0.366)
0.30	0.435 (0.434, 0.436)	0.465 (0.419, 0.510)
0.40	0.548 (0.547, 0.549)	0.572 (0.527, 0.617)
0.50	0.649 (0.648, 0.650)	0.685 (0.643, 0.727)
Model 2		
0.10	0.168 (0.167, 0.168)	0.197 (0.161, 0.233)
0.20	0.307 (0.306, 0.308)	0.311 (0.269, 0.353)
0.30	0.432 (0.431, 0.433)	0.460 (0.415, 0.506)
0.40	0.544 (0.543, 0.545)	0.5706 (0.525, 0.615)
0.50	0.646 (0.645, 0.646)	0.681 (0.639, 0.723)
$\widehat{PNF}$	p	p^* in AARP test data
Model 1		
0.808 (0.807, 0.809)	0.90	0.914 (0.889, 0.940)
0.671 (0.670, 0.672)	0.80	0.820 (0.785, 0.855)
0.554 (0.554, 0.555)	0.70	0.726 (0.686, 0.766)
0.450 (0.450, 0.451)	0.60	0.632 (0.588, 0.675)
0.357 (0.356, 0.357)	0.50	0.514 (0.469, 0.559)
Model 2		
0.811 (0.810, 0.811)	0.90	0.914 (0.889, 0.940)
0.675 (0.674, 0.676)	0.80	0.816 (0.781, 0.851)
0.558 (0.557, 0.559)	0.70	0.720 (0.679, 0.760)
0.454 (0.453, 0.455)	0.60	0.634 (0.590, 0.678)
0.361 (0.359, 0.361)	0.50	0.527 (0.482, 0.572)