STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:3474–3486
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2299

Estimation of time-dependent area under the ROC curve for long-term risk prediction

Lloyd E. Chambless*,† and Guoqing Diao

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27514, U.S.A.

SUMMARY

Sensitivity, specificity, and area under the ROC curve (AUC) are often used to measure the ability of survival models to predict future risk. Estimation of these parameters is complicated by the fact that these parameters are time-dependent and by the fact that censoring affects their estimation just as it affects estimation of survival curves or coefficients of survival regression models. The authors present several estimators that overcome these complications. One approach is a recursive calculation over the ordered times of events, analogous to the Kaplan–Meier approach to survival function estimation. Another is to first apply Bayes' theorem to write the parameters of interest in terms of conditional survival functions that are then estimated by survival analysis methods. Simulation studies demonstrate that the proposed estimators perform well in practical situations, when compared with an estimator (*c*-statistic, from logistic regression) that ignores time. An illustration with data from a cardiovascular follow-up study is provided. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: ROC curves; area under the curve; sensitivity; specificity; risk prediction; Kaplan–Meier estimator

1. INTRODUCTION

The use of receiver operating characteristic (ROC) curves to evaluate the performance of a continuous score to predict a medical outcome is well established [1], and has been extended to the case when that score is a linear combination of several factors, using coefficients from a logistic regression model [2, 3]. This use of a logistic regression model is not well suited to analysis of probability of disease onset when disease onset is observed over follow-up periods that vary in length by person, since probability of onset usually varies by length of observation period. Sensitivity and specificity and area under the ROC curve (AUC) are all defined in terms of probability of disease onset, so they are also time-dependent when follow-up period

*Correspondence to: Lloyd E. Chambless, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27514, U.S.A.
†E-mail: wchambless@unc.edu

Contract/grant sponsor: National Heart, Lung, and Blood Institute; contract/grant numbers: N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, N01-HC-55022

Received April 2004
Accepted April 2005

is not fixed. AUC has been used [3–12] to assess goodness of prediction from multiple factors of the risk of mortality or disease onset in settings where survival analysis was used to assess association, and the connection between the two methodologies was not always clear. In some studies it seems likely that the AUC analysis was totally independent of the survival analysis. In some others when it does appear that the coefficients from the survival model were used in risk prediction it is not clear how the analyses accounted for censoring in the computation of AUC or for its variation over time. The purpose of this paper is to provide algorithms for time-dependent computations related to assessment of performance of a risk score estimated from survival data in the presence of censoring. We compare estimates of AUC from this survival analysis approach with an approach that simply ignores time, say by the use of logistic regression, or by simple calculation of proportions. Examples are presented from the Atherosclerosis Risk in Communities (ARIC) study [13, 14] and from simulations.

2. BACKGROUND

A risk score for a person may be calculated from a regression model of the association between risk factors and disease incidence as the sum of products of the individual's level for each risk factor (X) times the model's beta coefficient (BETA) associated with that factor. We call this sum XBETA. Alternatively, some monotonic transform of XBETA might be used, such as the probability of disease onset within a specified time period. To assess the usefulness of such a risk score, sensitivity and specificity of risk prediction are computed for predicting disease as a function of XBETA being above a specified cutpoint. The plot of sensitivity *versus* 1-specificity over all possible cutpoints is called the receiver operating characteristic (ROC) curve [1]. Sensitivity is the probability of risk score being above the specified cutpoint for those with onset of disease in the period of observation, and specificity is the probability of score being below the cutpoint among those with no disease onset within the period. The area under the ROC curve and above the horizontal axis is a standard summary measure of the predictability of the risk score, and can be interpreted as the probability that a person with disease onset has higher score than a person without such onset [1]. In logistic regression this AUC is called the c-statistic [15]. $\mathrm{AUC} = 1$ indicates perfect prediction, while $\mathrm{AUC} = 0.5$ indicates prediction no better than chance.

The procedure described above is straightforward and appropriate when all persons are observed for the same length of time. When persons in the sample used to fit the model have variable follow-up time, i.e. when there is censoring, probability of event by a given time, or alternatively, one minus this probability, the survival probability, must be estimated by survival analysis techniques, and varies by time. This applies also to sensitivity, specificity, the ROC curve, and AUC. Thus, for example, $\mathrm{AUC}(t)$ is the probability that the probability that a person with disease onset by time t has a higher score than a person with no event by time t. This dependence of AUC on t is implicit in the work by Harrell *et al.* [15].

3. METHODS

3.1. Recursive formulae for AUC, sensitivity, and specificity

Suppose we have a continuous score Z which can be used to predict whether a person had disease onset by time t, where we predict YES if Z is above a cutpoint and NO if it is not.

In most of our applications Z will be XBETA from a survival model, but it could also be a published score in common use, such as a Framingham risk score for incident coronary heart disease (CHD) [4]. The AUC at time t for this score function can be shown to be the probability that those persons with an event by time t have a greater score than do those without an event by time t. We write $D_i(t)$ for the indicator of whether person i had an event by time t, 1 if YES, and 0 if NO. Thus,

$$\begin{aligned}\mathrm{AUC}(t) &= P(Z_i > Z_j | D_i(t) = 1,\ D_j(t) = 0) \\ &= P(Z_i > Z_j,\ D_i(t) = 1,\ D_j(t) = 0)/(P(D_i(t) = 1)P(D_j(t) = 0))\end{aligned}$$

where i, j represent independent observations. The problem with direct estimation of AUC(t), and similarly for time-dependent sensitivity and specificity, is that we do not know the event status at time t for persons censored before t. One method around this problem is similar to Kaplan–Meier estimation of the survival probability [16], namely to calculate AUC(t) recursively using the risk sets at each time of event. If we condition on the observed event times $t_1 < t_2 < \cdots < t_n$, we can use Kaplan–Meier-like arguments to derive, for $1 \leqslant m \leqslant n$, the value of AUC at time t_m is

$$\begin{aligned}\mathrm{AUC}(t_m) = &\left(\sum \gamma_k \lambda(t_k)(1 - \lambda(t_k))S(t_{k-1})^2 - \sum \tau_k \lambda(t_k)\right. \\ &\left. \times\ (1 - S(t_{k-1}))S(t_{k-1})\right) \big/ \left(S(t_m)(1 - S(t_m))\right) \qquad (1)\end{aligned}$$

both sums being for $k \leqslant m$, where S and λ are survival and hazard functions (later to be estimated by Kaplan–Meier methods conditional only on the times of events), $t_0 = 0$, $\tau_0 = 0$, and

$$\begin{aligned}\gamma_k &= P(Z_i > Z_j | D_i(t_k) = 1,\ D_i(t_{k-1}) = 0,\ D_j(t_k) = 0) \\ \tau_k &= P(Z_i > Z_j | D_i(t_{k-1}) = 1,\ D_j(t_{k-1}) = 0,\ D_j(t_k) = 1)\end{aligned}$$

The sensitivity and specificity of the prediction rule at time t for some cutpoint K can be shown to be

$$\mathrm{sens}(t_m, K) = P(Z_i > K | D_i(t_m) = 1) = \sum \rho_k(K)\lambda(t_k)S(t_{k-1})/(1 - S(t_m)) \qquad (2)$$

$$\begin{aligned}\mathrm{spec}(t_m, K) &= P(Z_i \leqslant K | D_i(t_m) = 0) \\ &= \left(P(Z_i \leqslant K) - \sum (1 - \rho_k(K))\lambda(t_k)S(t_{k-1})\right) \big/ S(t_m) \qquad (3)\end{aligned}$$

where $\rho_k(K) = P(Z_i > K | D_i(t_k) = 1,\ D_i(t_{k-1}) = 0)$ and the sums are for $k \leqslant m$. Note that when there is only one event at each time of event, $\rho_k(K)$ is either 0 or 1. See the appendix for the proof of (1). Proofs of (2) and (3) are similar.

We could also calculate the positive predictive value (PPV) and the negative predictive value (NPV) in the above setting by

$$\begin{aligned}\mathrm{PPV}(t_m,K) &= P(D_i(t_m)=1|Z_i>K)\\ &= P(Z_i>K|D_i(t_m)=1)P(D_i(t_m)=1)/(P(Z_i>K)\\ &= \mathrm{sens}(t_m,K)(1-S(t_m))/(P(Z_i>K) \qquad (4)\\ \mathrm{NPV}(t_m,K) &= P(D_i(t_m)=0|Z_i\leqslant K)\\ &= P(Z_i\leqslant K|D_i(t_m)=0)P(D_i(t_m)=0)/(P(Z_i\leqslant K)\\ &= \mathrm{spec}(t_m,K)S(t_m)/(P(Z_i\leqslant K) \qquad (5)\end{aligned}$$

Note that $\mathrm{PPV}(t,K)$ and $\mathrm{NPV}(t,K)$ could also be estimated directly from Kaplan–Meier estimates on the subsets $\{Z_i>K\}$ and $\{Z_i\leqslant K\}$. We will consider these parameters no further.

The ROC curve for time t is the plot of $\mathrm{sens}(t,K)$ *versus* 1-$\mathrm{spec}(t,K)$ over all possible values of K, and it can be shown that the area under this curve and above the horizontal axis, estimated by the trapezoidal rule, is $\mathrm{AUC}(t)$ [1].

3.2. Estimates of τ_k, γ_k, and $\rho_k(K)$

Let $\mathbb{R}_k$ be the risk set at time t_k, the persons not censored and without disease onset by t_k, and R_k the number of persons in that set. We will assume that there is only one event at each time, and operationally when this is not the case the times for tied events can be separated randomly by a small time. Write $Z_{d(k)}$ for the score for the event at time t_k. Then our estimators are, using # for the size of the set, are

$$\hat{\tau}_k = \#\{i: 1\leqslant i\leqslant k-1, Z_{d(i)}>Z_{d(k)}\}/(k-1) \qquad (6)$$

$$\hat{\gamma}_k = \#\{j\in\mathbb{R}_k: Z_{d(k)}>Z_j\}/(R_k-1) \qquad (7)$$

$$\hat{\rho}_k(K) = (Z_{d(k)}>K)(\text{i.e. 1 if } Z_{d(k)}>K, 0 \text{ if } Z_{d(k)}\leqslant K) \qquad (8)$$

Estimates for $\mathrm{AUC}(t)$, $\mathrm{sens}(t,K)$, and $\mathrm{spec}(t,K)$ can then be obtained by using (1)–(3), these estimates for $\hat{\tau}_k$, $\hat{\gamma}_k$, and $\hat{\rho}_k(K)$, and the Kaplan–Meier estimates for S and λ, $\hat{\lambda}(t_k)=1/R_k$ (under the assumption of only one event at each time of event, which could be generalized, replacing the 1 by the number of events at time t_k), and $\hat{S}(t_{k+1})=\hat{S}(t_k)^*(1-\hat{\lambda}(t_{k+1}))$. Note that the estimation of $\mathrm{AUC}(t)$ did not involve fitting any other survival model to the data of interest, beyond these Kaplan–Meier estimates unrelated to any exposure variables in the sample to be evaluated for AUC. The risk score Z being evaluated for AUC may or may not have been derived from models fit from the data at hand. Variances and covariances of $\hat{\tau}_k$ and $\hat{\gamma}_k$ could be calculated from standard binomial methods, leading to a variance estimator for the estimator of $\mathrm{AUC}(t)$, though it is quite complex, even more so if the score Z is

considered as estimated from observed data and we wanted to account also for this additional variance, as in our main application when Z is estimated from survival models. Instead we use bootstrapping [17] for variances and tests. Also, for any t, consistency of $\hat{\mathrm{AUC}}(t)$, conditional on $Z, t_1, \ldots, t_n$, follows from $\hat{\mathrm{AUC}}(t)$ being sums of products and quotients of consistent estimators, namely the Kaplan–Meier estimators of hazard rates and survival functions and the sample proportions $\hat{\tau}_k$ and $\hat{\gamma}_k$. We will also show in simulations that the estimators have very little bias in practical situations. Asymptotic normality of $\hat{\mathrm{AUC}}(t)$ could be approached in a similar manner, through the delta method, though we will make no use of this approach, instead using bootstrapping for statistical tests [17].

When Z is not continuous and ties in Z can be expected to occur in the data, an additional term could be added to each of (6)–(8): $\#\{i : 1 \leqslant i \leqslant k-1, Z_{d(i)} = Z_{d(k)}\}/(2(k-1))$, $\#\{j \in \mathbb{R}_k : Z_{d(k)} = Z_j\}/(2(R_k - 1))$, and $(Z_{d(k)} = K)/2$ (i.e $\frac{1}{2}$ if $Z_{d(k)} = K$) [15, 18]. To keep the presentation simpler we will omit this modification, in effect assuming a continuous Z.

Note that estimated $\hat{\mathrm{sen}}(t, K)$ is monotonic in terms of K (since the $\hat{\rho}_k(K)$ are non-decreasing with increasing K), but the same is not necessarily true for $\hat{\mathrm{spe}}(t, K)$. (It is a difference between monotonic estimators, but this does not make it monotonic.) For plotting purposes this may not be a problem if the K values are not chosen too densely. Alternatively, after sorting by increasing K moving medians smoothing can be applied to $\hat{\mathrm{spe}}(t, K)$. We show in our simulations that estimation of $\mathrm{AUC}(t)$ by integrating the actual estimated ROC curve by the trapezoidal rule with the smoothed $\hat{\mathrm{spe}}(t, K)$ gives virtually the same results as applying a trapezoidal estimation formula simply ignoring the non-monotonicity of $\hat{\mathrm{spe}}(t, K)$, and that both these values of AUC are quite close to those from (1) or (9).

Estimators of AUC, sensitivity, and specificity ignoring time and censoring would be as follows:

$$\begin{aligned}
\hat{\mathrm{AUC}} &= \#\{(i,j)|Z_i > Z_j, D_i = 1,\ D_j = 0\}/\#\{(i,j)|D_i = 1,\ D_j = 0\} \\
\hat{\mathrm{sen}}(K) &= \#\{i|Z_i > K,\ D_i = 1\}/\#\{i|D_i = 1\} \\
\hat{\mathrm{spe}}(K) &= \#\{i|Z_i \leqslant K,\ D_i = 0\}/\#\{i|D_i = 0\}
\end{aligned}$$

where $D_i = 1$, for example, means that the ith person had a disease onset at some time in his observation period.

3.3. Alternative estimators

An alternative estimator for $\mathrm{AUC}(t)$, which is derived directly from a fitted survival function $S(t|Z)$, can be derived as follows. Writing $g(z)$ for the density of Z, the conditional density for Z is

$$\begin{aligned}
g(z|D(t) = 1) &= P(D(t) = 1)|Z = z)g(z)/P(D(t) = 1) \\
&= (1 - S(t|z))g(z) \Big/ \left(\int (1 - S(t|u))g(u)\,\mathrm{d}u \right) \\
&= (1 - S(t|z))g(z)/E(1 - S(t|Z))
\end{aligned}$$

and $g(z|D(t)=0)=S(t|z)g(z)/E(S(t|Z))$, or more succinctly,

$$g(z|D(t)=k)=(k-S(t|z))g(z)/E(k-S(t|Z)) \quad \text{for } k=0,1$$

Then

$$\begin{aligned}
\mathrm{AUC}(t) &= \int_{(-\infty,\infty)} \left(\int_{(v,\infty)} g(u|D(t)=1)g(v|D(t)=0)\,\mathrm{d}u \right) \mathrm{d}v \\
&= (E(1-S(t|Z))E(S(t|Z)))^{-1} \int_{(-\infty,\infty)} \left(\int_{(v,\infty)} (1-S(t|u))S(t|v)g(u)g(v)\,\mathrm{d}u \right) \mathrm{d}v \\
&= (E(1-S(t|Z))E(S(t|Z)))^{-1} E((1-S(t|U))S(t|V)I(V<U)) \qquad (9)
\end{aligned}$$

where U and V are independent observations of Z and $I(V<U)$ is the indicator function for $V<U$ (1 when true, 0 otherwise). The bivariate expectation can be estimated as the mean over all (U,V) pairs of distinct observations. These expected values can be estimated using estimated values of Z from a regression survival model and predicted values of $S(t|z)$ also from that model, noting that these vary by t and are not the values output by the software at the individuals' own follow-up times. For example, from a proportional hazards model $S(t|z)=S_0(t)^{\exp(ZB)}$ for B the estimated coefficients of the variables Z and $S_0(t)$ is the estimated survival function for $Z=0$ [16, 19]. Similar estimates are available for sens and spec:

$$\begin{aligned}
\mathrm{sens}(t,K) &= \int_{(K,\infty)} g(u|D(t)=1)\,\mathrm{d}u \\
&= (E(1-S(t|Z)))^{-1} E((1-S(t|Z))I(K<Z)) \qquad (10) \\
\mathrm{spec}(t,K) &= \int_{(-\infty,K)} g(u|D(t)=0)\,\mathrm{d}u \\
&= (E(S(t|Z)))^{-1} E((S(t|Z))I(Z<K)) \qquad (11)
\end{aligned}$$

replacing the expected values by sample means.

Note that these alternative estimators require the use of a survival function of the score Z. If survival analysis has been used to produce the score $Z=X\mathrm{BETA}$, then this alternative estimate is immediately available. On the other hand, if the score Z is taken from an external source, then fitting a survival model with the single risk factor Z produces an equivalent score (Z multiplied by a beta coefficient) as well as a survival function of Z. Thus, in effect, this alternative estimator is always available. Note also that the estimators for $\mathrm{sens}(t,K)$ and $\mathrm{spec}(t,K)$ are monotonic in K if the survival function has that property.

4. EXAMPLE

The ARIC study is an NIH-supported study designed to study trends in CHD and stroke and the relation of CHD and stroke to potential risk factors [13, 14]. The 15 792 participants in the study had a baseline clinical examination in 1987–1989, while 45–64 years old. Ischemic stroke risk prediction using a set of traditional risk factors and potential improvement in AUC(10) by newer risk factors has been considered for follow-up through 2000 [13]. We will focus here only on women ($n = 7274, 155$ strokes). The traditional risk factors were age, current smoking, diabetes, left ventricular hypertrophy by ECG(LVH), previous CHD, hypertension medications, and systolic blood pressure (SBP), to which we added race (black, white). The newer risk factors included body mass index (BMI), waist–hip ratio, HDL-cholesterol, albumin, von Willebrand's factor (vWF), ethanol consumption, carotid artery wall thickness (IMT), and peripheral arterial disease (PAD). AUC(t) estimated by (1) using a Cox model is plotted *versus* follow-up years in Figure 1, showing some variation over time. Because of this variation it is clear that an approach to AUC that considers time in the estimation cannot agree with an approach that ignores time, except at one or a few particular times. For example, one might use the risk score *X*BETA from the Cox model to form 2×2 tables of disease status (ignoring time of onset) *versus* *X*BETA above a given cutpoint and then estimate sensitivity and specificity from the table and integrate to get the AUC. Alternatively one could apply $\text{AUC} = P(Z_i > Z_j | D_i = 1,\ D_j = 0)$ directly, ignoring time, or could use logistic regression software with *X*BETA as the independent variable. All of these lead to the same results. The estimated sensitivity and specificity at the median risk score and AUC, from the traditional risk factors, were $\widehat{\text{sen}} = 0.918$, $\widehat{\text{spe}} = 0.613$, and AUC ($c$-statistic) $= 0.792$, respectively. Even more simply the model itself could ignore time and be fit with logistic regression, and the results were $0.918, 0.607$, and 0.791. Here, and in the simulations to follow,

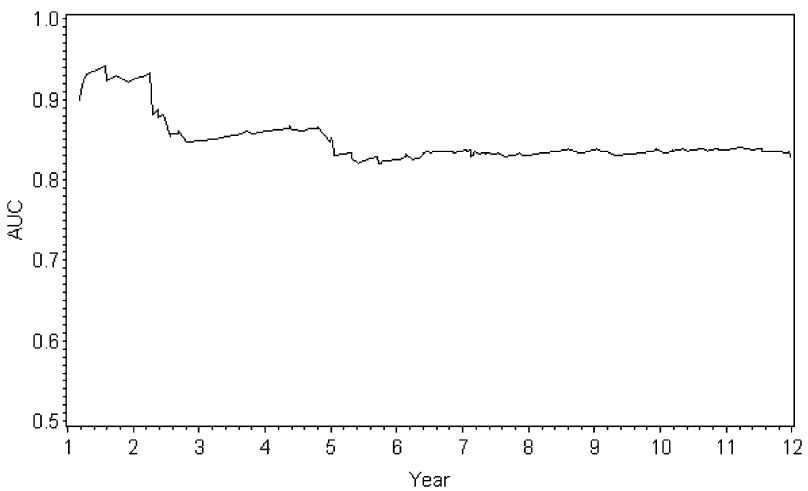

Figure 1. Area under the ROC curve for prediction of ischemic stroke by traditional stroke risk factors, by years of follow-up. Women, the Atherosclerosis Risk in Communities Study 1987–2000.

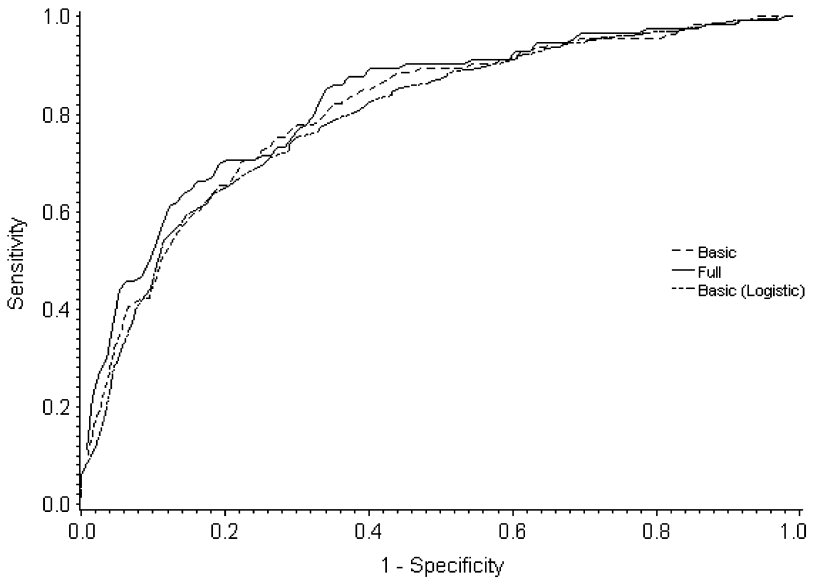

Figure 2. ROC curves for prediction of ischemic stroke within 10 years by traditional stroke risk factors (-- Basic, using (1), or ---Basic, using logistic regression) or including additional novel risk factors (full, using (1)). Women, the Atherosclerosis Risk in Communities Study 1987–2000.

all persons are included in the logistic regression, regardless of time of censoring, and the disease onset status is as at the end of the individual's observation time. From time-dependent method (1), at 10 years, the similar estimates were 0.894, 0.506, and 0.813, and from the alternative time-dependent method (9), at 10 years, 0.861, 0.506, and 0.796.

The evaluation of AUC, sensitivity, and specificity above are subject to being overly optimistic due to being estimated on the same data set used to fit the models to derive the risk scores. We now remedy this in an actual application of AUC, to compare risk prediction from a set of traditional risk factors to that of an expanded set that includes newly discovered risk factors. To avoid evaluating the difference in AUC between the two models on the same data used to fit the models, we randomly split the data set into two samples, here of equal size, the training set and the test set. The two models were fit on the training set and AUC from each was calculated on the test set and the AUC difference calculated. To get an estimate of the variability of our estimate of the difference in AUC, we drew 500 bootstrap samples [17], and from each sample repeated the training set/testing set approach. Then the mean AUC difference over the 500 bootstrap samples was calculated, and the central 95 per cent of the 500 differences used as a confidence interval. The difference in AUC (at 10 years) was estimated as 0.014 and 0.027 by our two methods (1) and (9), respectively, on the test set. The similar differences estimated on the training set, which we would not use, were 0.023 and 0.041. By the bootstrap confidence interval the AUC from the extended model was not statistically significantly greater than that from the traditional factors model.

To illustrate ROC curves in one example, without the training set/test set approach for simplicity, we plot them in Figure 2 at 10 years, for the above application, for a risk score

using the traditional risk factors and either computed with logistic regression ignoring time and censoring, or by method (1), or using traditional risk factors plus the expanded list and applying method (1).

5. SIMULATIONS

For the simulations we took $S_0(t)$ as exponential or Weibull survival and risk variable X as a standard univariate Gaussian variable or $X=(X_1,X_2)$ with X_2 binomial and X_1 Gaussian with mean varying by X_2 value, and then set $S(t|X)=S_0(t)^{\exp(XB)}$ with known B (proportional hazards, $XB=X\text{BETA}$). We simulated event time t_e from $S(t|X)$ and censoring time t_c from an exponential distribution independent of $S_0(t)$ and X, and took observed follow-up time as $t=\min(t_e,t_c)$. We let $D(c)=1$ if person had an event by time c, else $D(c)=0$. (We omit subscripts for persons.) Then (9) can be integrated numerically to give the 'known' values for AUC(t) under the assumptions used to generate the data, and these can be compared with average estimated values over many simulated data sets. For each generated data set, we estimated AUC values at two different time points from both (1) and the alternative method (9), using a risk score from the coefficients of a Cox regression model. For comparison, we also estimated AUC values based on the c-statistic in logistic regression. Table I summarizes the results based on 500 simulated data sets. For each data set we simulated 1000 observations. The proposed estimators (1) and (9) appear to be virtually unbiased, while the c-statistics seem to be always underestimated. The c-statistics have more bias at longer time points. The standard deviations of the 1000 alternative estimates based on (9) appear to be always smaller than that of the estimator based on (1). Thus, the alternative estimator might be more efficient than the first estimator presented. In all situations, the c-statistics have much larger standard deviations, compared with the proposed two estimators. We also estimated the sensitivity and specificity based on (2) and (3), and (10) and (11), respectively, and the AUC values estimated from integrating numerically the estimated ROC curves are virtually the same as for the proposed AUC estimators based on (1) and (9), respectively (relative differences are <0.1 per cent, results not shown). Simulations with sample size 500 had standard deviations quite

Table I. True values of AUC and means and standard deviations (SD) from simulations for three estimators, based on (1) (Method 1), on (9) (Method 2), or on the c-statistic in logistic regression (Method 3), evaluated at time points $t=5$ or 10. In both models a and c, the survival times are generated from a baseline exponential distribution with mean 50, while in model b, the baseline distribution is Weibull. In both models a and b, the covariate X is standard normally distributed; in model c, X_2 is binomial with $p=0.3$, X_1 given X_2 is normally distributed with mean 0.5 X_2 and variance 1.

			Method 1		Method 2		Method 3	
Model	Time point	True AUC	Mean	SD	Mean	SD	Mean	SD
a	5	0.6165	0.6191	0.0294	0.6170	0.0199	0.6132	0.0552
b	5	0.6159	0.6188	0.0314	0.6170	0.0171	0.6086	0.0519
c	5	0.6536	0.6532	0.0269	0.6538	0.0176	0.6487	0.0499
a	10	0.6225	0.6235	0.0266	0.6220	0.0228	0.6079	0.0436
b	10	0.6251	0.6261	0.0226	0.6261	0.0180	0.6035	0.0390
c	10	0.6617	0.6612	0.0212	0.6620	0.0181	0.6448	0.0379

close to $\sqrt{2}$ times the standard deviations for sample size 1000 (data not shown) as expected, indicating that these standard deviations go to 0 as $1/\sqrt{n}$.

6. DISCUSSION

In the context of survival models, where time to event is modelled as a function of baseline exposure factors, AUC is a function of time and should be estimated with proper consideration of censoring. We have presented several estimators for AUC(t) that meet these requirements. Hypothesis testing or confidence intervals for AUC may be estimated by bootstrapping. Simulations show that the proposed estimators all have little bias, when compared with an estimator (c-statistic) that ignores time.

Heagerty *et al.* [20] have also addressed estimation of ROC curves in the setting of a time-dependent disease variable and as related to one factor Z. They present two estimators of sensitivity and specificity. The first estimator uses Bayes' theorem to write

$$\mathrm{SENS}(K,t) = (1 - S(t|Z > K))P(Z > K)/(1 - S(t))$$

and then estimates both the conditional and unconditional survival functions by the Kaplan–Meier method and $P(Z > K)$ by the empirical distribution function. For large values of K the sample size for $Z > K$ may be small for getting the conditional K–M estimate. This method avoids estimation of the survival function $S(t|z)$, but in our applications this survival function is often of interest, for example in presenting 10 year predicted risk of heart disease. The second method presented is similar to our method using (10) and (11), except that we get at the survival curve through estimating the risk factor coefficients from a Cox proportional hazards model and then taking these coefficients as known to estimate the survival function (see SAS PHREG [19] and references therein), while Heagerty *et al.* use a non-parametric kernel estimator ('nearest neighbour') of $S(t|Z > K)$, smoothing by using nearby values of Z. In our application where Z is a linear combination of risk factors, this method would not consider the individual factor coefficients, which are often of interest. Hagerty and Zheng [21] contrast the above sensitivity/specificity parameters, called 'Cumulative/Dynamic', with alternative time-dependent parameters called 'Incident/Static', where now, in our notation, at time t_m $\mathrm{sens}(t_m, K) = P(Z_i > K|D_i(t_m) = 1, D_i(t_{m-1}) = 0)$ and $\mathrm{spec}(t_m, K) = P(Z_i \leqslant K|D_i(t^*) = 0)$ for some fixed follow-up period through t^*, but their main focus is on the combination of the Incident sensitivity from the 'Incident/Static' pair and the Dynamic specificity from the 'Cumulative/Dynamic' pair. We find, however, that the first formulation (Cumulative/Dynamic, as in the present paper) more appropriate to the setting of evaluating predictivity of risk of disease onset over some fixed period, as is often considered in manuscripts referenced herein.

Dodd and Pepe [22] proposed a semi-parametric regression method for the AUC in the context of binary response (disease or no disease). It would be interesting to extend this regression method to survival data and compare the performance with our proposed methods, although for the applications presented here we would nevertheless still be interested in the survival curve, to estimate 10 year risk.

The methods presented here would often be used when deriving a risk score on a random test subset of the data and estimating AUC on the rest of the data, the training set [23]. This strategy avoids the overestimation of the performance of a prediction rule when the

performance evaluation is done on the same data set used to fit the model leading to the prediction rule. The authors have done this [13] in a study of stroke risk prediction, in which the ARIC data were used on a randomly selected three-fourths of the data to get one risk score $Z = X\text{BETA}$, and then that risk score and one from the Framingham Study [5] were compared for AUC on the rest of the data. In fact, this was repeated for 1000 random reselections of the three-fourths of the data and the differences in AUCs averaged over all. Further, our application in this paper showed that the improvement in adding variables to a traditional risk score was overstated when the AUC calculations were done on the same data on which the risk score model was fit.

Our methods to estimate time-dependent sensitivity and specificity could also be used to compute time-dependent partial AUC, generally for specified ranges of specificity [24], since in clinical practice certain ranges of specificity would be of no practical interest though they contribute to overall AUC. This partial area is of interest in clinical applications, and can be interpreted as the average sensitivity over a specified range of specificity. If one wanted to consider also a limited range for sensitivity, the ROC curve could be used to find the corresponding specificity for partial AUC calculations. And in some cases it may only be the possibly limited-to-a-subrange ROC curve which is of interest, in which case the methods presented here are still of use for estimation of sensitivity and specificity in a manner that accounts for time and censoring.

Either of our proposed methods to estimate AUC or sensitivity and specificity would appear to be preferable to a method that ignores time, but more investigation is needed to choose between the two. The estimator based on expected values of functions of the estimated survival function had smaller variance in the simulations than the iterative estimator, though it is computationally more intensive. At the cost of higher variances the former estimator may be made less computationally intensive by being based on a sample of the cohort. When comparisons are to be made to a risk score for which the survival function is not available, our first method is straightforward to apply. SAS macros for these estimators of AUC are available from the authors at www.aricnews.net.

APPENDIX A

Proof of (1): The proof is by induction, with the proof for $m = 1$ left to the reader. It is useful to note that $D_i(t_m) = 1$ implies $D_i(t_{m+1}) = 1$, and that $D_j(t_{m+1}) = 0$ implies $D_j(t_m) = 0$. If we assume the truth of (1) for m, then consider the case for $m + 1$:

$$
\begin{aligned}
&P(Z_i > Z_j,\ D_i(t_{m+1}) = 1,\ D_j(t_{m+1}) = 0) \\
&\quad = P(Z_i > Z_j,\ D_i(t_m) = 1,\ D_j(t_{m+1}) = 0) + P(Z_i > Z_j,\ D_i(t_{m+1}) = 1,\ D_i(t_m) = 0,\ D_j(t_{m+1}) = 0) \\
&\quad = P(Z_i > Z_j,\ D_i(t_m) = 1,\ D_j(t_m) = 0) - P(Z_i > Z_j,\ D_i(t_m) = 1,\ D_j(t_{m+1}) = 1,\ D_j(t_m) = 0) \\
&\qquad + P(Z_i > Z_j,\ D_i(t_{m+1}) = 1,\ D_i(t_m) = 0,\ D_j(t_{m+1}) = 0) \\
&\quad = P(Z_i > Z_j,\ D_i(t_m) = 1,\ D_j(t_m) = 0)
\end{aligned}
$$

$$-P(Z_i > Z_j | D_i(t_m)=1,\ D_j(t_{m+1})=1,\ D_j(t_m)=0)P(D_i(t_m)=1,\ D_j(t_{m+1})=1,\ D_j(t_m)=0)$$

$$+P(Z_i > Z_j | D_i(t_{m+1})=1,\ D_i(t_m)=0,\ D_j(t_{m+1})=0)P(D_i(t_{m+1})=1,\ D_i(t_m)=0,\ D_j(t_{m+1})=0)$$

$$= \left(\sum_{(k \leqslant m)} \gamma_k \lambda(t_k)(1-\lambda(t_k))S(t_{k-1})^2 - \sum_{(k \leqslant m)} \tau_k \lambda(t_k)(1-S(t_{k-1}))S(t_{k-1}) \right)$$

$$-\tau_{m+1} P(D_i(t_m)=1)P(D_j(t_{m+1})=1,\ D_j(t_m)=0)$$

$$+\gamma_{m+1} P(D_i(t_{m+1})=1,\ D_i(t_m)=0)P(D_j(t_{m+1})=0)$$

$$= \left(\sum_{(k \leqslant m)} \gamma_k \lambda(t_k)(1-\lambda(t_k))S(t_{k-1})^2 - \sum_{(k \leqslant m)} \tau_k \lambda(t_k)(1-S(t_{k-1}))S(t_{k-1}) \right)$$

$$-\tau_{m+1}(1-S(t_m))P(D_j(t_{m+1})=1 | D_j(t_m)=0)P(D_j(t_m)=0)$$

$$+\gamma_{m+1} P(D_i(t_{m+1})=1 | D_i(t_m)=0)P(D_i(t_m)=0)S(t_{m+1})$$

$$= \left(\sum_{(k \leqslant m)} \gamma_k \lambda(t_k)(1-\lambda(t_k))S(t_{k-1})^2 - \sum_{(k \leqslant m)} \tau_k \lambda(t_k)(1-S(t_{k-1}))S(t_{k-1}) \right)$$

$$-\tau_{m+1}(1-S(t_m))\lambda(t_{m+1})S(t_m) + \gamma_{m+1}\lambda(t_{m+1})S(t_m)S(t_m)(1-\lambda(t_{m+1}))$$

$$= \left(\sum_{(k \leqslant m+1)} \gamma_k \lambda(t_k)(1-\lambda(t_k))S(t_{k-1})^2 - \sum_{(k \leqslant m+1)} \tau_k \lambda(t_k)(1-S(t_{k-1}))S(t_{k-1}) \right) \quad \square$$

ACKNOWLEDGEMENTS

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, and N01-HC-55022. The authors thank the staff and participants of the ARIC study for their efforts, and thank Mary Jo Earp and Ding-yi Zhao for their programming.

REFERENCES

1. Campbell G. General methodology I: advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine* 1994; **13**:499–508.
2. Marwick TH, Case CC, Siskind V, Woodhouse SP. Prediction of survival from resuscitation: a prognostic index derived from multivariate logistic model analysis. *Resuscitation* 1991; **22**:129–137.
3. Iglesias del Sol A, Moons KGM, Hollander M, Hofman A, Koudstaal PJ, Grobbee DE, Breteler MMB, Witteman JCM, Bots ML. Is carotid intima-media thickness useful in cardiovascular disease risk assessment? The Rotterdam Study. *Stroke* 2001; **32**:1532–1538.
4. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease risk using risk factor categories. *Circulation* 1998; **97**:1837–1847.
5. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham study. *Stroke* 1991; **22**:312–318.

6. Moons KGM, Bots ML, Elwood PC *et al*. Prediction of stroke in the general population in Europe (EUROSTROKE): is there a role for fibrinogen and electrocardiography? *Journal of Epidemiology and Community Health* 2002; **56**(Suppl I):i30–i36.
7. Cooper JA, Miller GJ, Bauer KA, Morrissey JH, Meade TW, Howarth DJ, Barzegar S, Mitchell JP, Rosenberg RD. Comparison of novel hemostatic factors and conventional risk factors for prediction of coronary heart disease. *Circulation* 2000; **102**:2816–2833.
8. Liao Y, McGee DL, Cooper RS. Prediction of coronary heart disease mortality in blacks and whites: polled data from two national cohorts. *American Journal of Cardiology* 1999; **84**:31–36.
9. Assman G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the Prospective Cardiovascular Muenster (PROCAM) Study. *Circulation* 2002; **105**:310–315.
10. Chen MK, Chen THH, Liu JP, Chang CC, Chie WC. Better prediction of prognosis for patients with nasopharyngeal carcinoma using primary tumor volume. *Cancer* 2004; **100**:2160–2166.
11. Navarro-Cano G, Del Rincon I, Pogosian S, Roldan JF, Escalante A. Association of mortality with disease severity in rheumatoid arthritis, independent of comorbidity. *Arthritis and Rheumatism* 2003; **48**(9):2425–2433.
12. Palisar J, Graefen M, Karakiewicz PI, Hammerer PG, Huland E, Haese A, Fernandez S, Erbersdobler A, Henke RP, Huland H. Assessment of clinical and pathologic characteristics predisposing to disease recurrence following radical prostatectomy in men with pathologically organ-confined prostate cancer. *European Urology* 2002; **41**:155–161.
13. Chambless LE, Heiss G, Shahar E, Earp MJ, Toole J. Prediction of ischemic stroke risk in the Atherosclerosis Risk in Communities (ARIC) Study. *American Journal of Epidemiology* 2004; **160**:259–269.
14. Chambless LE, Folsom AR, Sharrett AR, Sorlie P, Couper D, Szklo M, Nieto FJ. Coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC) Study. *Journal of Clinical Epidemiology* 2003; **56**(9):880–890.
15. Harrell FE, Lee LL, Marks DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**:361–387.
16. Kalbfleish J, Prentice R. *The Statistical Analysis of Failure Time Data*. Wiley: New York, NY, 1980.
17. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Technical Report No. 175*, Division of Biostatistics, Stanford University, 1995.
18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**:837–845.
19. SAS Institute Inc. *SAS/STAT User's Guide*, *Version 8*. SAS Institute Inc., Cary, NC, 1999.
20. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**:337–344.
21. Heagerty P, Zheng Y. Survival model predictive accuracy and ROC curves. *UW Biostatistics Working Group Paper Series* 2003; 219.
22. Dodd LE, Pepe MS. Semiparametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association* 2003; **98**:409–417.
23. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician* 1983; **37**:36–48.
24. Zhang DD, Zhou XH, Freeman Jr DH, Freeman JL. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Statistics in Medicine* 2002; **21**:701–715.

