Published in final edited form as:

Stat Med. 2019 July 10; 38(15): 2868–2882. doi:10.1002/sim.8157.

Prioritized Concordance Index for Hierarchical Survival Outcomes

Li C. Cheung^{a,*}, Qing Pan^b, Noorie Hyun^c, Hormuzd A. Katki^a

^aDivision of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Rockville, MD, U.S.A.

^bDepartment of Statistics, The George Washington University, Washington, DC

^cInstitute for Health and Equity, Medical College of Wisconsin, Milwaukee, WI, U.S.A.

Abstract

We propose an extension of Harrell's concordance (C) index to evaluate the prognostic utility of biomarkers for diseases with multiple measurable outcomes that can be prioritized. Our prioritized concordance index measures the probability that, given a random subject pair, the subject with the worst disease status as of a time τ has the higher predicted risk. Our prioritized concordance index uses the same approach as the win-ratio, by basing generalized pairwise comparisons on the most severe or clinically important comparable outcome. We use an inverse probability weighting technique to correct for study-specific censoring. Asymptotic properties are derived using U-statistic properties. We apply the prioritized concordance index to two types of disease processes with a rare primary outcome and a more common secondary outcome. Our simulation studies show that when a predictor is predictive of both outcomes, the new concordance index can gain efficiency and power in identifying true prognostic variables compared to using the primary outcome alone. Using the prioritized concordance index, we examine whether novel clinical measures can be useful in predicting risk of type II diabetes in patients with impaired glucose resistance whose disease status can also regress to normal glucose resistance. We also examine the discrimination ability of four published risk models among ever-smokers at risk of lung-cancer incidence and subsequent death.

Keywords

area under the receiver operating curve; evaluating predictions; U-statistics; progressive/regressive disease process; illness-death disease process

1. Introduction

The practice of combining disease outcomes into composite measures is used in clinical trials to better account for all benefits and harms of a disease intervention or to boost power and reduce costs when the primary outcome is rare [1]. Measures such as the widely used

*Correspondence to: 9609 Medical Center Drive, Room 7E612, Rockville, MD 20850. li.cheung@nih.gov. Phone: (240) 276-5812. Fax: (240) 276-7838.

concordance (C) index [2, 3, 4, 5] and related statistics, such as the area under the time-dependent ROC curve [6, 7] and the K index [8], are used to evaluate the predictive ability of predictors for a single survival outcome or in the presence of competing risks [9]. These measures have been extended to recurrent event data [10], but no measure currently exists for diseases with many different outcomes. Event-specific concordance indices can evaluate predictions of each outcome separately; however, such an approach would lose much statistical power, and a single measure would better capture the overall prognostic value of predictors for a multifaceted disease process.

The conventional approach to constructing composite measures emphasizes each patient's first event, but this can lead to ignoring serious outcomes and deaths in favor of less serious, non-fatal events. An alternative approach is to construct composite measures that prioritize the most severe or clinically important comparable outcome, whenever possible. An example of this is the win-ratio [11], which uses generalized pairwise comparisons of the most severe/important comparable outcome [12, 13] to estimate the probability that the treatment group would have better clinical outcomes than the control group. Using a similar pairwise comparison approach, we develop the prioritized concordance (C_{τ}^*) index, an extension of the C index using prioritized outcomes. For a single survival outcome, the C index measures the probability that the subject observed to have the earlier event has the higher predicted risk. Our C_{τ}^* index measures the probability that the subject with the worse disease status as of time τ has the higher predicted risk, where the subject with the worst disease status is the one with the more severe/important disease outcome, or if they are tied in that respect, the one with the earlier event.

We may wish to identify prognostic biomarkers for patients with impaired glucose resistance (IGR) who are at high risk for type II diabetes but can also regress to normal glucose resistance (NGR), which has been shown to reduce long-term risks of diabetes [14]. Diabetes is the most severe outcome but is rarely observed, so that the C index based on diabetes alone would be estimated from only a small proportion of all patient pairs in the study. Instead, a composite measure that includes regression to NGR, yet prioritizes diabetes as the most important outcome, would allow for a more comprehensive evaluation. For patient pairs, we determine a “winner” and a “loser” by first comparing the times to the diabetes event. If both times are tied (eg. if both patients have not developed diabetes by the end of follow-up), then the time to the next most important outcome, regression to NGR, is compared. If the times to NGR are also tied, then a winner cannot be determined and the pair is not usable in the C_{τ}^* index. The definition of winner and loser in a pair of subjects easily adapts to scenarios with competing hierarchical outcomes [9, 15]. If a subject is censored after having one outcome, the competing outcomes never occur, and the winner is the subject with the more favorable outcome or observed event time (if both subjects have the same outcome).

The C_{τ}^* index can be used for any disease processes in which the outcomes can be prioritized, such as the illness-death process [16]. In cancer analyses, cancer death is the preferred gold standard endpoint, as diagnosed cancer can differ in severity by stage and aggressiveness. Due to the rarity of cancer deaths however, cancer diagnosis is often used as

a surrogate endpoint. The C_τ^* index would allow use of both outcomes while prioritizing cancer deaths.

In section 2, we review the C index and the C_τ index proposed by Uno et al. [17], which correct the C index for study-specific censoring up to a time τ . We formally introduce the C_τ^* index, which estimates the probability that the subject with the worse disease status as of time τ has the higher predicted risk. The C_τ^* determines winners/losers of the disease process using observed times and applies inverse probability weighting [17] to correct for censoring bias. To quantify improvement in predictive ability between two sets of biomarkers, we use the difference between two C_τ^* indices [18].

Two types of disease processes are considered in our simulation studies (Section 3) and example applications (Section 4): intermediate disease that can progress or regress and the illness-death disease process. In section 3, we use simulation studies to compare the C_τ^* index against the C_τ index for the primary outcome alone. We show that when the primary outcome is rare, the C_τ^* index, which incorporates information from auxiliary processes, leads to more accurate and efficient identification of biomarkers that are predictive of all disease outcomes. Using data from the Diabetes Prevention Program (DPP) clinical trial [19] and the C_τ^* index, we show that adding certain novel clinical measures (adiponectin blood level, insulin sensitivity, and corrected insulin response) to the usual clinical workup may better identify the African-Americans with impaired glucose resistance who are likely to progress to type II diabetes (Section 4). With the extra information from regression to normal glucose resistance, the C_τ^* index significantly detects the added predictive ability of these new clinical measures, which would have been missed using the C_τ index for type II diabetes alone. We also evaluate the predictive ability of four proposed lung-cancer models in the control arm of the National Lung Screening Trial (NLST) [20] (Section 4). Despite previous studies noting large differences in calibration (number of expected versus observed events) performance of these models among racial/ethnic subgroups in the US [21], we did not detect a significant difference between the discrimination performance of these models in racial/ethnic subgroups within the NLST.

2. Concordance Indices for Prioritized Survival Outcomes

2.1. Notation

We use the following notation for survival data of prioritized outcomes. For an event $k \in \{1, 2, \dots, K\}$ (ordered by decreasing severity or clinical importance), let T_k be the survival time and Z_k be the corresponding predictive score, so that larger predictive scores indicate shorter predicted survival time. If the reverse is true, let Z_k be the negative of the original predictive score. The predictive scores are event-specific scores based on one or more covariates, such as $Z_k = \hat{\beta}_k' W_k$, where $\hat{\beta}_k$ is a vector of event-specific regression coefficients estimated for a vector of prognostic factors, W_k . Let D be the censoring time; because one event type may effectively censor another, let D_k , $k = 1, 2, \dots, K$, be the event-specific censoring times.

For the k th event, let (T_{ki}, D_{ki}, Z_{ki}) be identical and independently distributed for subjects $i = 1, 2, \dots, n$. For the i th subject, we observe (O_{ki}, Z_{ki}) , where $O_{ki} = (X_{ki}, \Delta_{ki})$, $X_{ki} = T_{ki} \wedge D_{ki}$, and Δ_{ki} is 1 if $X_{ki} = T_{ki}$ and 0 otherwise. Define $\mathbf{O}$ and $\mathbf{Z}$ to be K by n matrices of observed times and predictive scores. Let $\mathbf{O}_k, \mathbf{Z}_k, \mathbf{O}_i$ and $\mathbf{Z}_i$ be k th event-specific or i th subject-specific vectors within those matrices.

2.2. A Review of Harrell's C Index

The concordance statistic for a single event can be defined as the probability of agreement between the ordering of event times and predicted scores among random subject pairs,

$$C = P(Z_i > Z_j | T_i < T_j).$$

(1)

The ordering of the observed times is not always known under censoring. For right-censoring, Harrell et al. [2, 3, 4, 5] proposed estimating C using all comparable (usable) pairs, where subject pairs are “comparable” if it is known which subject has the event first. An indicator function for whether the event times for subject pair (i, j) are comparable (usable) is given by the following,

$$\mathbb{U}(O_i, O_j) = I(X_i < X_j) \Delta_i + I(X_j < X_i) \Delta_j,$$

(2)

where $I(\cdot)$ is the indicator function. For subject pair (i, j) , a function for whether the ordering of the predictive scores agree with that of the observed event times is given by the following,

$$\begin{aligned} \mathbb{A}(O_i, Z_i, O_j, Z_j) &= I(Z_i > Z_j) I(X_i < X_j) \Delta_i + I(Z_j > Z_i) I(X_j < X_i) \Delta_j \\ &+ \frac{1}{2} I(Z_i = Z_j) \{ I(X_i < X_j) \Delta_i + I(X_j < X_i) \Delta_j \}. \end{aligned}$$

(3)

This expression for $\mathbb{A}(O_i, Z_i, O_j, Z_j)$ allows for the possibility of ties in the predictive score, but in much of the statistics literature, predictive scores are assumed to be continuous, with no ties. To keep the mathematical expressions simple, we will adopt the convention of no

ties in the rest of the manuscript, with the understanding that ties can occur, especially when evaluating categorical biomarkers, and are accounted for in the calculations.

Averaging over all possible sample pairs in the data set,

$$U(O) = \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n \mathbb{U}(O_i, O_j),$$

(4)

we get a U-statistic type estimator for the probability that the event time is comparable in a subject pair. Similarly,

$$A(O, Z) = \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n \mathbb{A}(O_i, Z_i, O_j, Z_j)$$

(5)

gives us a U-statistic type estimator for the joint probability that the event time is comparable and that the ordering agree with that of the predicted scores.

The concordance (C) index is the ratio of these U-statistics,

$$C(O, Z) = \frac{A(O, Z)}{U(O)},$$

(6)

which converges in probability to the following [17],

$$C(O, Z) \xrightarrow{P} P(Z_i > Z_j | T_i < T_j, T_i \leq D_i \wedge D_j)$$

(7)

Because the estimator is a simple ratio involving comparable pairs, the C index converges to a concordance probability that depends on the study-specific censoring distribution. Uno et al. [17] proposed the C_τ index, a modified version of the C index that does not converge to a

quantity that depends on the study-specific censoring and would instead estimate the concordance probability up to a time τ , which is given by the following,

$$C(\tau) = P(Z_i > Z_j | T_i < T_j, T_i \leq \tau).$$

(8)

In their estimator, equations (2) and (3) are corrected for study-specific censoring using an inverse probability weighting technique as followed,

$$\mathbb{U}(O_i, O_j, \tau) = \hat{G}(X_i)^{-2} I(X_i < X_j, X_i \leq \tau) \Delta_i + \hat{G}(X_j)^{-2} I(X_j < X_i, X_j \leq \tau) \Delta_j,$$

(9)

and

$$\begin{aligned} \mathbb{A}(O_i, Z_i, O_j, Z_j, \tau) &= I(Z_i > Z_j) \hat{G}(X_i)^{-2} I(X_i < X_j, X_i \leq \tau) \Delta_i \\ &+ I(Z_j > Z_i) \hat{G}(X_j)^{-2} I(X_j < X_i, X_j \leq \tau) \Delta_j, \end{aligned}$$

(10)

where $\hat{G}(\cdot)$ is the Kaplan-Meier estimator for the censoring distribution, $G(t) = P(D > t)$, and τ is a prespecified time point such that $P(D > \tau) > 0$. Because the tail part of the Kaplan-Meier survival estimate can be unstable, we found that restricting to values of τ for which at least 10% of subjects are still followed is necessary to avoid large variance estimates for the resulting estimator.

Let $U(O, \tau)$ and $A(O, Z, \tau)$ be U-statistics constructed by averaging equations (9) and (10), respectively, over all possible subject pairs. Then the C_τ index is given by the following,

$$C(O, Z, \tau) = \frac{A(O, Z, \tau)}{U(O, \tau)},$$

(11)

which converges in probability to $C(\tau)$. [17]

2.3. The C_{τ}^* Index

For a disease process with multiple disease outcomes that can be prioritized by greatest severity or clinical importance, the win-ratio [11] assigns the loser of the disease process among a subject pair to be the one experiencing the worse disease outcome. If both subjects have the same worst outcome, then the loser is the subject who experiences that outcome first.

We propose estimating $C(\tau)$ using this definition of pairwise winners and losers of a disease process as of time τ . For a subject pair (i, j) , denote $*$ to be the highest priority (most severe or clinically important) outcome that occurs among either subject as of time τ . The winner and loser of the disease process as of time τ can then be defined purely in terms of their ordering of times to event $*$ (if a subject does not have outcome $*$ as of time τ , then it is assumed to have occurred after the subject with the outcome $*$ as of time τ).

The concordance probability for prioritized outcomes can be expressed as the following,

$$\begin{aligned} C^*(\tau) &= P(Z_{*i} > Z_{*j} | T_{*i} < T_{*j}, T_{*i} \leq \tau) \\ &= \frac{P(Z_{*i} > Z_{*j}, T_{*i} < T_{*j}, T_{*i} \leq \tau)}{P(T_{*i} < T_{*j}, T_{*i} \leq \tau)}. \end{aligned}$$

(12)

Thus $C^*(\tau)$ is the probability that, given one of the subject pairs has the worse overall disease status, he will also have the higher predicted risk.

Due to study-specific censoring, the highest priority comparable outcome for a subject pair is not necessarily the same as if both subjects had been followed until time τ . Therefore estimating the concordance probability simply as a ratio of counts would lead to a similar bias as Harrell's C index. To correct for study-specific censoring bias, we construct an estimator for $C^*(\tau)$ using the inverse probability weighting technique described by Uno et al. [17].

We construct a prioritized concordance (C_{τ}^*) index to estimate $C^*(\tau)$. Define

$$\mathbb{U}^*(O_{ki}, O_{kj}, \tau) = \mathbb{U}(O_{ki}, O_{kj}, \tau) \left\{ 1 - \sum_{l=1}^{k-1} \mathbb{U}_{\tau}(O_{li}, O_{lj}, \tau) \right\},$$

(13)

and

$$\mathbb{A}^*(O_{ki}, Z_{ki}, O_{kj}, Z_{kj}, \tau) = \mathbb{A}(O_{ki}, Z_{ki}, O_{kj}, Z_{kj}, \tau) \left\{ 1 - \sum_{l=1}^{k-1} \mathbb{U}_{\tau}(O_{li}, O_{lj}, \tau) \right\}.$$

(14)

where $\mathbb{U}_{\tau}(O_{ki}, O_{kj})$ and $\mathbb{A}_{\tau}(O_{ki}, Z_{ki}, O_{kj}, Z_{kj})$ is given in equations (9) and (10) respectively.

We can then construct U-statistic estimators as followed,

$$U^*(O, \tau) = \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n \sum_{k=1}^K w_k \mathbb{U}^*(O_{ki}, O_{kj}, \tau)$$

(15)

and

$$A^*(O, Z, \tau) = \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n \sum_{k=1}^K w_k A^*(O_{ki}, Z_{ki}, O_{kj}, Z_{kj}, \tau),$$

(16)

where $0 \leq w_k \leq 1$ are outcome-specific weights determined apriori. Note that these U-statistics estimate $\sum_{k=1}^K w_k P(Z_{ki} > Z_{kj}, T_{ki} < T_{kj}, T_{ki} \leq \tau, k = \star)$ and $\sum_{k=1}^K w_k P(T_{ki} < T_{kj}, T_{ki} \leq \tau, k = \star)$, respectively.

The prioritized concordance (C_{τ}^*) index is the ratio of these U-statistics, as followed,

$$C^*(O, Z, \tau) = \frac{A^*(O, Z, \tau)}{U^*(O, \tau)}.$$

(17)

As $n \rightarrow \infty$, the prioritized concordance index converges in probability to the following,

$$C^*(O, Z, \tau) \xrightarrow{p} \frac{\sum_{k=1}^K w_k P(Z_{ki} > Z_{kj}, T_{ki} < T_{kj}, T_{ki} \leq \tau, k = \star)}{\sum_{k=1}^K w_k P(T_{ki} < T_{kj}, T_{ki} \leq \tau, k = \star)}.$$

(18)

The full derivation is given in Web Appendix A. When all outcome-specific weights are equal to 1, equation (18) is equivalent to equation (12) from the law of total probability. When only the weight of the highest priority outcome is equal to one (and all other weights are less than one), then we have a penalized estimator for $C^*(\tau)$ that puts less emphasis on lower priority events.

Using the properties of U-statistics, we show that the prioritized concordance index is asymptotically normal and has closed form variance, which due to the length of the expressions, is given in Web Appendix B. Confidence intervals and hypothesis testing for the prioritized concordance index can be derived based on the properties of asymptotically normal random variables [22].

2.4. Comparing two correlated C^*_τ indices

Suppose we have two prioritized concordance indices, $C^*(O, Z, \tau)$ and $C^*(O, Y, \tau)$, based on two correlated predictors, Z and Y , respectively, and we wish to evaluate $H_0: C^*(O, Z, \tau) = C^*(O, Y, \tau)$ in the same data set. The null hypothesis can also be evaluated as $H_0: C^*(O, Z, \tau) - C^*(O, Y, \tau) = 0$. $C^*(O, Z, \tau) - C^*(O, Y, \tau)$ can be written as a function of U-statistics, as followed,

$$C^*(O, Z, \tau) - C^*(O, Y, \tau) = \frac{A^*(O, Z, \tau) - A^*(O, Y, \tau)}{U^*(O, \tau)} = \frac{A^*(O, Z - Y, \tau)}{U^*(O)},$$

(19)

where

$$\begin{aligned}
A^*(O, Z - Y, \tau) = & \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n \sum_{k=1}^K w_k \left[\left\{ I(Z_{ki} > Z_{kj}) - I(Y_{ki} > Y_{kj}) \right\} \hat{G}(X_{ki})^{-2} I \right. \\
& \left. (X_{ki} < X_{kj}, X_{ki} \leq \tau) \Delta_{ki} \right. \\
& + \left. \left\{ I(Z_{kj} > Z_{ki}) - I(Y_{kj} > Y_{ki}) \right\} \hat{G}(X_{kj})^{-2} I(X_{kj} < X_{ki}, X_{kj} \leq \tau) \Delta_{kj} \right] \\
& \left\{ 1 - \sum_{l=1}^{k-1} \cup_{\tau}(O_{li}, O_{lj}, \tau) \right\}
\end{aligned}
\tag{20}$$

Using a one-shot non-parametric approach [18], $C^*(O, Z, \tau) - C^*(O, Y, \tau)$ can be shown to be asymptotically normal with closed form variance, which due to the length of the expressions, is given in Web Appendix B. Confidence intervals and hypothesis testing follow that of asymptotically normal random variables.

3. Simulation Studies

We conduct simulation studies with 1000 simulation data sets of 150 subjects to evaluate the C^*_τ index against the C_τ index for the most important outcome alone. We consider two disease processes with multiple disease outcomes that can be prioritized: a progressive-regressive disease process and an illness-death process.

3.1. Simulations for progressive-regressive disease process

When patients are in an intermediate disease state, the events of disease regression (RG) and disease progression (PG) are negatively correlated (Figure 1). For many diseases, progression is an absorbing state that terminates the disease regression process while patients whose disease regress can still eventually reach the progression state (semi-competing risks); however, for some diseases, regression can also be an absorbing state that terminates the disease progression process (competing risks). While identifying prognostic factors for disease progression may be the primary goal, utilizing the events from the regression process may provide additional information and improve the discriminatory accuracy. This is especially true when disease progression is rare relative to disease regression.

We simulate regressive, intermediate, and progressive disease states using a Markov chain, where disease progression is an absorbing state but disease regression is not. All subjects start in an intermediate disease state. Times to disease progression and regression states from an intermediate state are drawn from exponential distributions with rates of $\exp(-3.5 + Y_i + W_{ij})$ and $\exp(-.5 - 2Y_i - W_{ij})$, respectively; the earliest of these times is the state that the subject next transitions to. Time to reach an intermediate state again from a regressed state is drawn with rate $\exp(-1 + Y_i)$. The variable Y_i follow a normal distribution with mean zero and variance three, and W_{ij} is a standard normal frailty term drawn each time, $j = 1, 2, \dots$, the

intermediate state is reached. Censoring times follow independent exponential distributions. Time to disease progression and time to first disease regression are 69% and 34% right-censored, respectively.

All C_τ and C_τ^* indices are evaluated at $\tau = 25$ to ensure each simulation data set had at least one event beyond that time point (for stability of the estimates). For performance measures, such as bias and coverage, the true value of the concordance indices are estimated from the uncensored simulated event time data (by averaging across 100 data sets of 15,000 subjects).

3.1.1. Evaluating discovery rates—One potential use of concordance indices is to discover predictors for disease while minimizing false discovery of predictors independent of the disease process. We evaluate ten candidate predictors separately: five of which are correlated with the (true) predictor, Y_j , with correlation coefficient of 1/6 and five of which are independent of the true predictor. All candidate predictors are normally distributed with means of 0, variances of 3, and are mutually independent.

The bias, empirical standard deviation (ESD), asymptotic standard error (ASE), and coverage probability (CP) for the C_τ and C_τ^* indices are presented in Table 1. All reported performance measures are from averaging over the five correlated or independent predictors, as appropriate for each section. Table 1 also includes the proportion of times that the null hypothesis of the concordance index being equal to 0.5 was rejected at $\alpha = 0.05$ significance level and the rate at which predictors independent of the disease process was discovered. Only the progression outcome is used for the C_τ index. For the C_τ^* index, weights of 1 and .75 are used for disease progression and regression, respectively.

The C_τ^* index also utilizes the information from the disease regression outcome and thus, have smaller variance than the C_τ index that uses disease progression alone. Compared to the C_τ index, the C_τ^* index reject the null hypothesis more often when the predictors are correlated with the true predictor (35.6% versus 23.0% rejection rates) and have lower false discovery rate of independent predictors (12.8% versus 19.9%). For predictors independent of the disease process, both the C_τ and C_τ^* indices appropriately have estimates of approximately 0.5 and reject the null hypothesis at approximately $\alpha = 0.05$ level (6% and 5.2% rates, respectively).

As a sensitivity analysis, we examine the scenario where the evaluated predictor is predictive only for disease progression but not disease regression. In these sets of simulations, Y_j is replaced with independent predictors, Y_{1j} and Y_{2j} , in defining the transition rates for disease progression and disease regression, respectively. We evaluate ten candidate predictors separately: five of which are correlated with Y_{1j} with correlation coefficient of 1/6 and five of which are independent of Y_{1j} . All candidate predictors are normally distributed with means of 0, variances of 3, and are mutually independent. Time to disease progression and time to first disease regression are 78% and 35% right-censored, respectively. While the C_τ^* index has a smaller variance, the estimated concordance probability is also smaller because the risk scores are not predictive for subject pairs for which disease regression is the most

important comparable event. In this case, the C_τ index rejected the null hypothesis more often than the C_τ^* index when the predictors are correlated with the true predictor for disease progression (9.9% versus 6.9% rejection rates) and had lower false discovery rate of independent predictors (38.3% versus 42.8%) (Web Table 1).

3.1.2. Evaluating correlated predictors—Concordance indices should ideally select the variables that are the best predictors for the disease process, while not selecting unnecessary predictors. Let W_i and Z_i be predictors that are correlated with the true predictor used to simulate the disease process, Y_i , with correlation coefficient of .9 and 1/6, respectively. Using the C_τ index and the C_τ^* index, we assess the gain in predictive ability, as measured by the difference in the concordance indices, from using predictors 1) W_i , 2) Y_i , and 3) $\beta_0 Y_i + \beta_1 Z_i$, where β_0 and β_1 are estimated by a Cox model.

The bias, empirical standard deviation (ESD), asymptotic standard error (ASE), and coverage probability (CP) for the difference in concordance indices are given in Table 2. Table 2 also includes the proportion of times the null hypothesis of no difference was rejected. For the C_τ^* index, weights for disease progression and regression are 1 and .75, respectively.

Compared to the C_τ index, the C_τ^* index has greater ability to detect that Y_i is a better predictor than W_i (92.9% versus 59.4% reject rates) without increasing the probability of selecting models that also included the unnecessary predictor, Z_i . Of note, in our test for the difference in C_τ^* for a nested model adding an unnecessary predictor, there was upward bias of < 0.0015 and the null hypothesis was rejected at considerably less than $\alpha = 0.05$ level.

3.2. Simulations for illness-death disease process

The illness-death disease process [16] consist of a three state process where patients begin in a non-exposed/alive state. They can either become ill and subsequently die, or they can die without transitioning through an illness state. Once patients are ill, they cannot regress to a non-exposed/alive state and death is an absorbing state (Figure 2). This has also been referred to as a semi-competing risk framework [23].

When considering biomarkers for an illness-death disease process, we are interested in variables that can predict who will transition from the non-exposed/alive state to illness and who will transition from illness to death. In our simulation studies, time from non-exposed/alive to illness and time from illness to death are drawn from exponential distributions with rates of $\exp(-2.9 + Y_i)$ and $\exp(-3.7 + .5 Y_i)$ respectively, where Y_i is drawn from a normal distribution with mean zero and variance three. Censoring times (which include death from causes other than illness) follow independent exponential distributions. Time to illness and time to death from illness are 50% and 80% right-censored, respectively.

All estimates of C_τ and C_τ^* indices are evaluated at $\tau = 25$ to ensure each simulation data set had at least one event beyond that time point (for stability of the estimates). For performance measures, such as bias and coverage, the true value of the concordance indices were

estimated from the uncensored simulated event time data (averaging across 100 data sets with 15,000 subjects).

3.2.1. Evaluating discovery rates—As with the progressive-regressive disease process, we evaluate the ability of the prioritized concordance index to discover predictors while minimizing discovery of false predictors in an illness-death disease process (Table 3). We evaluate ten candidate predictors separately: five of which are correlated with the (true) predictor, Y_i , with correlation coefficient of 1/6 and five of which are independent of the true predictor. All candidate predictors are normally distributed with means of 0, variances of 3, and are mutually independent. Only the death from illness outcome is used for C_τ . For C_τ^* , weights of .75 and 1 are used for illness and for subsequent death.

The findings were similar to that of the illness-death process. The C_τ^* index has a smaller variance as it also utilizes information from the illness outcomes. Compared to the C_τ index, the C_τ^* index reject the null hypothesis more often when the predictors are correlated with the true predictor (23.1% versus 18.0% rejection rates) and have lower false discovery rate for independent predictors (13.6% versus 25.8%). Predictors independent of the disease process have estimated C_τ and C_τ^* of approximately 0.5 and reject the null hypothesis at approximately $\alpha = 0.05$ level (6.3% and 4.4% rates, respectively).

We also conducted a sensitivity analysis examining the scenario where the predictors are predictive only for death from illness, but not illness. In these sets of simulations, Y_i is replaced with independent predictors Y_{1i} and Y_{2i} for illness and subsequent death. We evaluate ten candidate predictors separately: five of which are correlated with Y_{2i} with correlation coefficient of 1/6 and five of which are independent of Y_{2i} . All candidate predictors are normally distributed with means of 0, variances of 3, and are mutually independent. Time to illness and subsequent death are 50.5% and 85.5% right-censored, respectively. While the C_τ^* index has a smaller variance, the concordance probability is also smaller because the risk scores are not predictive for subject pairs whose most important comparable event is illness. The C_τ index rejected the null hypothesis more often than the C_τ^* index when the predictors are correlated with the true predictor (9.6% versus 4.9% rejection rates) with similar false discovery rate (41.7% versus 40.0%) (Web Table 2).

3.2.2. Evaluating correlated predictors—As with the progressive-regressive disease process, we evaluate whether the difference in C_τ^* indices can select the variables that are the best predictors for the disease process, while not selecting unnecessary predictors. Let W_i and Z_i be predictors that are correlated with the true predictor used to simulate the disease process, Y_i , with correlation coefficient of .9 and 1/6, respectively. Using the C index and the C_τ^* index, we assess the gain in predictive ability, as measured by the difference in the concordance indices, from using predictors 1) W_i , 2) Y_i , and 3) $\beta_0 Y_i + \beta_1 Z_i$ where β_0 and β_1 are estimated by a Cox model.

The bias, empirical standard deviation (ESD), asymptotic standard error (ASE), and coverage probability (CP) for the difference in concordance indices are given in Table 4.

Table 4 also includes the proportion of times the null hypothesis of no difference was rejected. For the C_τ^* index, weights of .75 and 1 are used for illness and subsequent death, respectively.

Compared to the C_τ index, the C_τ^* index has greater ability to detect that Y_i is a better predictor than W_i (47.3% versus 36.2% reject rates) without increasing the probability of selecting models that also included the unnecessary predictor, Z_i . Of note, in our test for the difference in C_τ^* for a nested model adding an unnecessary predictor, there was upward bias of 0.0020 and the null hypothesis was rejected at considerably less than $\alpha = 0.05$ level.

4. Applications of the C^* Index

4.1. Example application in progressive-regressive disease process

Incidence of type 2 diabetes has been found to be 2.4-fold greater in African-American women and 1.5-fold greater in African-American men than in their white counterparts [24]. Diagnosis of type 2 diabetes is often delayed until serious complications of the eyes, nerves, kidneys, and cardiovascular system are present [25]. Identifying African-Americans at greatest risk of developing diabetes would allow targeted interventions, such as lifestyle changes or metformin, that have been proven to reduce diabetes incidence [26]. In addition to conventional clinical measures such as fasting plasma glucose, 2-hour plasma glucose, and glycosylated hemoglobin (HbA_{1C}), the Diabetes Prevention Program (DPP) clinical trial [19] collected novel clinical measures, such as adiponectin blood level, insulin sensitivity, and corrected insulin response (CIR) on 3,234 participants of all races with impaired glucose resistance at enrollment. We evaluate whether these additional clinical measures have improved prognostic ability compared to using conventional clinical measures and demographic covariates alone in 260 African-Americans with impaired glucose resistance assigned to the placebo arm. Tables for the hazard ratios in the fitted models and for describing the covariates for the 260 African-Americans are given in Web Tables 3–5. The coefficients for the NGR and type II diabetes outcomes are different in size and significance level because these two outcomes reflect different aspects of the underlying hyperglycemia process. Nevertheless, the contributions of the additional novel predictors are combined by the proposed concordance index. As the primary outcome of diabetes was observed in only 71 subjects, we use the C_τ^* index to supplement this information with the secondary outcome of regression back to normal glucose resistance (NGR). Patients who reach NGR have been shown to have a 56% reduction in long-term risks of diabetes in the Diabetes Prevention Program Outcomes Study (DPPOS) [14], which is the subsequent long-term observation study of participants from the DPP trial [27].

Using the 645 African-Americans enrolled in the three arms (lifestyle changes, metformin, and placebo) of the DPP trial, we fit Cox proportional hazard models for type II diabetes and NGR, with and without the novel clinical measures, while also including demographic covariates, conventional clinical measures, and study arm as covariates. Risk scores are created using the linear predictors from the models. For each set of risk scores, we estimate the event specific C_τ index and the C_τ^* index in African-Americans enrolled in the placebo

arm of DPP, using τ equal to six year. For the C_τ^* index, we used outcome weights of 1 for type II diabetes and 0.56 for NGR, which corresponds to the long-term diabetes risk reduction from achieving NGR [14]. Further discussion of the choices of outcome weights for a progressive-regressive disease process are given in Web Appendix D. In Table 5, we report the estimated concordance indices, the increase in the concordance indices from the additional clinical measures, the associated standard errors, and the p-values for rejecting the null hypothesis that the increase in the concordance indices are zero.

The C_τ index for type II diabetes based on conventional risk factors is 0.719 and increased to 0.735 when novel risk factors are added. However, the difference in C_τ was not significant ($p=0.100$). The prioritized concordance index increased from 0.713 to 0.731 when novel risk factors are added to conventional risk factors. While the difference in C_τ^* was similar to that of C_τ , the variance was smaller because the underlying number of subject pair comparisons increased by 50% when we supplement type II diabetes with improvement to NGR. The C_τ^* index has greater ability to detect the added predictive value for the additional clinical measures ($p=0.049$).

4.2. Example application in illness-death disease process

The US Preventative Services Task Force (USPSTF) recommend lung-cancer screening with low-dose computed tomography (CT) for ever-smokers, ages 55–80 years, who have smoked in the past 15 years and have at least 30 pack-years of lifetime exposure [28]. Using risk prediction models to select US ever-smokers for CT lung screening is likely to prevent more lung-cancer deaths than the current USPSTF recommendations [29, 30]. Previously published work examined calibration and discrimination of all lung-cancer incidence/mortality models in the existing literature and found that 3 models for lung cancer incidence, the Bach [31], PLCom2012 [32], and LCRAT [29] models, and 1 model for lung cancer death, the LCDRAT [29] model, performed best in US cohorts and may be best suited for risk-based selections [21]. Subgroup analysis found that these models performed less well in racial/ethnic minorities, likely because the data sets used to develop the models were predominantly caucasian [21]. The differences in calibration ability of these models in racial/ethnic minorities have previously been described [21]; we used the C_τ^* index to examine potential differences in discrimination ability of these models in racial/ethnic minorities.

The National Lung Screening Trial randomized 53,454 patients into CT and control arms from 2002 to 2004 and ended in 2010 [20]. We use data from the control arm of the National Lung Screening Trial (NLST) to examine the discrimination ability of the four models in African-Americans, Hispanic-Americans, and Asian-Americans (the data is described in Web Table 6). One advantage of using the control arm of the NLST for evaluating risk models for lung-cancer screening is that detection (annual screens with chest x-ray) and access to health care post-diagnosis is largely consistent across participants. Outcome data is available as both lung-cancer incidence and lung-cancer mortality. As diagnosed lung-cancers can differ by stage and aggressiveness, lung-cancer death is considered the better gold standard endpoint. However, lung-cancer incidence is more common in data sets and is

often used as a surrogate endpoint. We evaluate the four models using C_τ indices, for lung-cancer death alone and for lung-cancer incidence alone, and the C_τ^* index that uses both lung-cancer death and incidence, but prioritizes death. For C_τ^* , outcome weights of 1 for both lung-cancer death and lung-cancer incidence was used (assigning no penalty for comparisons based on lung-cancer incidence). The value of τ for both indices is seven years for African-Americans and Hispanic-Americans and six years for Asian-Americans (as estimates became unstable at τ equal to seven years for Asian-Americans).

The LCDRAT model performed best among African-Americans, with the highest value of C_τ for both lung-cancer incidence and death and for the C_τ^* index that uses both endpoints. Among Hispanic-Americans, PLCOm2012 performed best by the C_τ index for lung-cancer death, but the Bach model performed best by the C_τ index for lung-cancer incidence and by the C_τ^* index that uses both endpoints. Among Asian-Americans, PLCOm2012 performed best by both lung-cancer incidence and death and for the C_τ^* index that uses both endpoints. For the four models, none of the differences in discrimination ability in ethnic subgroups were significant at a $\alpha = 0.05$ level for either C_τ or C_τ^* .

5. Discussion

Harrell's concordance (C) index measures the discriminatory accuracy of biomarkers for a single survival outcome, but many disease processes have multiple survival outcomes that can be prioritized by severity or importance. We proposed the prioritized concordance (C_τ^*) index that can better capture predictive ability for a multifaceted disease process. The C_τ^* index is the probability that among a random subject pair, the subject with the worst disease status as of time τ also has the higher predicted risk or biomarker value. The C_τ^* index compares pairs using the same method as the win ratio, determining a subject has the worst disease status if he experiences the more severe/clinically important event or, if they are tied in that respect, he has the earlier event. Following the methods proposed by Uno et al. [17], the C_τ^* index uses an inverse probability weighting technique to create a concordance index that is free of study-specific censoring distribution.

The utility of the prioritized concordance (C_τ^*) index is demonstrated in simulation studies and example applications for two disease processes with multiple disease outcomes that can be prioritized: a progressive-regressive disease process and an illness-death process. The simulation studies demonstrated that the C_τ^* index, which supplements primary outcomes (eg. disease progression or death) with correlated secondary outcomes (eg. disease regression or illness), can be more efficient and accurate than the C_τ index (a version of Harrell's C index that is corrected for study-specific censoring) that uses only the primary outcome. In addition to having greater power for hypothesis testing, these measures had a lower false discovery rate. One caveat is that when the predictor is predictive of only the primary outcome, then the C_τ^* index was less useful than the C_τ index. However, when a biomarker is only predictive for one disease outcome but not others, some thought must be

given to whether the biomarker is actually a predictor for the disease process (eg. lack of insurance may be predictive of cancer death but may not be predictive of acquiring cancer).

Our example applications show how the C_{τ}^* index can use multiple disease outcomes that can be prioritized to assess predictions for a disease process. We applied the C_{τ}^* index to patients with impaired glucose resistance who could progress to type II diabetes or regress to normal glucose resistance. Using the C_{τ}^* index found that adding novel clinical biomarkers (adiponectin blood level, insulin sensitivity, and corrected insulin response) to the usual pre-diabetic clinical examinations can better predict the risk of type II diabetes among African-Americans and thus allow for better targeted interventions. We also applied the C_{τ}^* index to ever-smokers at risk of lung-cancer incidence and subsequent death. The four best performing lung-cancer models (identified as such in previously published work [21]) had previously reported differences in calibration performance in ethnic/racial subgroups [21]. We examined discrimination ability in ethnic/racial subgroups of the NLST; despite using information from both lung-cancer incidence and death, the C_{τ}^* index was not able to find a statistically significant difference in discrimination ability.

The C^* index allows the use of weights to penalize for less important outcomes. In our application to patients with impaired glucose resistance who can progress to type II diabetes or regress to normal glucose resistance, we suggest outcome-specific weights for use in intermediate disease that can progress or regression. It is less clear what weights to use for other disease processes with prioritized outcomes, such as the illness-death process. Choice of weights could be based on the increased probability of death, the change in quality of life scores [33], or on costs of treatment/care. The use of weights allows for valuable flexibility based on the aims of the predictions but are also subjective.

We suggest that weights should be determined apriori to data analysis using external information. If existing information to inform weights are scant or non-existent and users wish to estimate weights from the same data set that they are estimating the concordance index, then the variance methods would need to account for the extra variation from the estimated weights. In such cases, the impact on efficiency of the C_{τ}^* index would depend largely on the efficiency of the weight estimators. A general approach that accounts for the extra variation is given in Web Appendix F.

Much of the same caveats that apply to the C_{τ} index also apply to the C_{τ}^* index. These concordance indices are based on the pairwise rankings of predicted and actual event times and does not differentiate between pairs whose times are close together versus those who are far apart. The use of inverse probability weighting to estimate concordance probabilities up to a time τ removes much of the bias that occurs in the C index due to study-specific censoring; however, to avoid C_{τ}^* estimates with large variances, we found that the choice of τ should be such that at least 10% of subjects are followed until that time. This means that the trade-off in bias reduction from using inverse probability weighting is that the underlying concordance index is constructed from fewer pairwise comparisons than traditional concordance indices. Even after correction for study-specific censoring, concordance indices

estimated from one population may not be transportable among populations, especially if the distribution of disease risks in those populations are dissimilar.

We used the difference in the C_τ^* indices to measure the added predictive value of new biomarkers or to compare predictive ability of competing models. In conducting inference, we assumed asymptotic normality for the difference in the C_τ^* indices and estimated variance using methods similar to the one-shot non-parametric method proposed by Kang et al. [18]. Our methods account for correlation between the sets of predictive scores. However, recent work on the difference in AUC for nested models for binary outcomes has shown that comparing concordance probabilities is not trivial and many of the same issues may apply to using the difference in C_τ^* for nested models. The difference in concordance probabilities for nested models is biased upward as regression models will typically add to the concordance probability even if the predictor has no added predictive value [34]. Some variance formulas are based on the assumption that predictions (with and without new variables) are mutually independent among patients, but this property may be violated among nested models, leading to conservative variance estimates [34]. In addition, asymptotical normality may hold only when the new predictors are predictive of the outcome [35]. These issues may explain the simulation results in which the added discrimination ability of an unnecessary predictor failed to reject at $\alpha = 0.05$ level (Tables 2 and 4). More work is needed in the proper comparison of concordance probabilities for nested models for time-to-event outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The Diabetes Prevention Program (DPP) was conducted by the DPP Research Group and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the General Clinical Research Center Program, the National Institute of Child Health and Human Development (NICHD), the National Institute on Aging (NIA), the Office of Research on Women's Health, the Office of Research on Minority Health, the Centers for Disease Control and Prevention (CDC), and the American Diabetes Association. The data from the DPP were supplied by the NIDDK Central Repositories. This manuscript was not prepared under the auspices of the DPP and does not represent analyses or conclusions of the DPP Research Group, the NIDDK Central Repositories, or the NIH.

The authors thank the referees and editors for many useful suggestions that improved the article.

References

1. Williams SCP. Composite endpoints in clinical trials. *The Scientist* 2016; 30(7). Available from: <http://www.the-scientist.com>.
2. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA: The Journal of the Medical Association* 1982; 247 2543–2546.
3. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine* 1984;3:143–152. [PubMed: 6463451]
4. Harrell FE Jr, Lee KL, Mark DB. Tutorial in biostatistics: multivariate prognosis models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; 15:361–387. [PubMed: 8668867]

5. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in statistical analysis: model specific population value and confidence interval estimation. *Statistics in Medicine* 2004; 23 2109–2123. DOI: 10.1002/sim.1802. [PubMed: 15211606]
6. Heagerty PJ, Zheng Y. Survival model predictive accuracy and roc curves. *Biometrics* 2005; 61(1) 92–105. DOI: 10.1111/j.0006-341X.2005.030814.x. [PubMed: 15737082]
7. Saha-Chaudhuri P, Heagerty PJ. Non-parametric estimation of a time-dependent predictive accuracy curve. *Biostatistics* 2013; 14:42–59. DOI: 10.1093/biostatistics/kxs021. [PubMed: 22734044]
8. Gnen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005; 92(4) 965–970. DOI: 10.1093/biomet/92.4.965.
9. Saha P, Heagerty PJ. Time-dependent predictive accuracy in the presence of competing risk. *Biometrics* 2010; 66(4) 999–1011. DOI: 10.1111/j.1541-0420.2009.01375.x. [PubMed: 20070296]
10. Kim S, Schaubel DE, McCullough KP. A C-index for recurrent event data: application to hospitalizations among dialysis patients. *Biometrics* 2017 DOI: 10.1111/biom.12761.
11. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal* 2012; 33(2) 176–182. DOI: 10.1093/eurheartj/ehr352. [PubMed: 21900289]
12. Finkelstein DM and Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine* 1999; 18(11) 1341–1354. DOI: 10.1002/(SICI)1097-0258(19990615)18:11<1341::AID-SIM129>3.0.CO;2-7. [PubMed: 10399200]
13. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine* 2010; 29(30) 3245–3257. DOI: 10.1002/sim.3923. [PubMed: 21170918]
14. Perreault L, Pan Q, Mather KJ, Watson KE, Hamman RF, Kahn SE, Diabetes Prevention Program Research Group. Effect of regression from prediabetes to normal glucose regulation on long-term reduction in diabetes risk: results from the Diabetes Prevention Program Outcomes Study. *Lancet* 2012; 379(9833):2243–51. DOI: 10.1016/S0140-6736(12)60525-X. [PubMed: 22683134]
15. Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA. Concordance for prognostic models with competing risks. *Biostatistics* 2014; 15(3):526–539. DOI: 10.1093/biostatistics/kxt059. [PubMed: 24493091]
16. Xu J, Kalbfleisch JD, Tai B. Statistical analysis of illness-death processes and semicompeting risk data. *Biometrics* 2010; 66(3) 716–725. DOI: 10.1111/j.1541-0420.2009.01340.x. [PubMed: 19912171]
17. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* 2011; 30(10) 1105–1117. DOI: 10.1002/sim.4154. [PubMed: 21484848]
18. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. *Statistics in Medicine* 2015; 34(4) 685–703. DOI: 10.1002/sim.6370.
19. The Diabetes Prevention Program (DPP) Research Group. The Diabetes Prevention Program: design and methods for a clinical trial in the prevention of type 2 diabetes. *Diabetes Care* 1999; 22 623–34. DOI: 10.2337/diacare.22.4.623. [PubMed: 10189543]
20. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine* 2011; 365(5) 395–409. DOI: 10.1056/NEJMoa1102873. [PubMed: 21714641]
21. Katki HA, Kovalchik SA, Petito LC, Cheung LC, Jacobs E, Jemal A, Berg CD, Chaturvedi AK. Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening. *Annals of Internal Medicine* 2018; 169(1) 10–19. DOI: 10.7326/M17-2701. [PubMed: 29800127]
22. Casella G, Berger RL. *Statistical Inference*: 2nd edition Duxbury Press; CA: 2001.
23. Fine JP, Jiang H, Chappell R. On semi-competing risk data. *Biometrika* 2001; 88(4) 907–919. DOI: 10.1093/biomet/88.4.907.
24. Brancati FL, Kao L, Folsom AR, Watson RL, Szklo M. Incident Type 2 Diabetes Mellitus in African American and White Adults: The Atherosclerosis Risk in Communities Study. *JAMA* 2000; 283(17) 2253–2259. DOI:10.1001/jama.283.17.2253. [PubMed: 10807384]

25. Harris MI, Eastman RC. Early detection of undiagnosed diabetes mellitus: a US perspective. *Diabetes/Metabolism Research and Reviews* 2000; 16(4) 230–236. DOI: 10.1002/1520-7560(2000)9999:9999::AID-DMRR122;3.0.CO;2-W. [PubMed: 10934451]
26. The Diabetes Prevention Program (DPP) Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *The New England Journal of Medicine* 2002; 346(6) 393–403. DOI: 10.1056/NEJMoa012512. [PubMed: 11832527]
27. The Diabetes Prevention Program (DPP) Research Group. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *Lancet* 2009; 374 1677–1686. DOI: 10.1016/S0140-6736(09)61457-4.
28. Moyer VA, U. S. Preventive Services Task Force. Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine* 2014; 160(5) 330–338. DOI: 10.7326/M13-2771. [PubMed: 24378917]
29. Katki HA, Kovalchik SA, Berg CD, Cheung LC, Chaturvedi AK. Development and validation of risk models to select ever-smokers for CT lung cancer screening. *JAMA* 2016; 315 2300–11. DOI: 10.1001/jama.2016.6255. [PubMed: 27179989]
30. Cheung LC, Katki HA, Chaturvedi AK, Jemal A, Berg CD. Preventing lung cancer mortality by computed tomography screening; the effect of risk-based versus U.S. Preventive Services Task Force eligibility criteria. *Annals of Internal Medicine* 2018; 168(3) 229–232. DOI: 10.7326/M17-2067. [PubMed: 29297008]
31. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, Hsieh LJ, Begg CB. Variations in lung cancer risk among smokers. *Journal of the National Cancer Institute* 2003; 95(6) 470–478. DOI: 10.1093/jnci/95.6.470. [PubMed: 12644540]
32. Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, Chaturvedi AK, Silvestri GA, Riley TL, Commins J, Berg CD. Selection criteria for lung-cancer screening. *New England Journal of Medicine* 2013; 368 728–736. DOI: 10.1056/NEJMoa1211776. [PubMed: 23425165]
33. Gold MR, Franks P, McCoy KI, Fryback DG. Toward Consistency in Cost-Utility Analyses: Using National Measures to Create Condition-Specific Values. *Medical Care* 1998; 36(6) 778–792. [PubMed: 9630120]
34. Seshan VE, Gnen M, Begg CB. Comparing ROC curves derived from regression models. *Statistics in Medicine* 2012; 32(9) 1483–1493. DOI: 10.1002/sim.5648. [PubMed: 23034816]
35. Heller G, Seshan VE, Moskowitz CS, Gnen M. Inference for the difference in the area under the ROC curve derived from nested binary regression models. *Biostatistics* 2017; 18(2) 260–274. [PubMed: 27655817]

Figure 1. Intermediate disease that can progress or regress. The dashed lines indicate that, in some cases, transition from regressed disease back to intermediate disease is possible.

Figure 2.
Illness-Death Process

Table 1.

Evaluating discovery rate of C_τ and C_τ^* indices for intermediate disease that can progress or regress. 1000 simulation data sets of 150 subjects were used. 10 candidate predictors are evaluated: 5 correlated and 5 independent of the disease process. Time to progression and first regression are 69% and 34% right-censored, respectively. For C_τ , only disease progression was used. The C_τ^* index uses both disease progression and regression events with outcome-specific weights of 1 and 0.75, respectively. Both indices were evaluated at $\tau = 25$. Negative and positive signs for the bias indicates underestimation and overestimation, respectively. ESD is the empirical standard deviation; ASE is the asymptotic standard error; CP is the coverage probability. False discovery rate (FDR) is defined as the number of independent predictors that reject H_0 divided by the number of all predictors that reject H_0 , averaged over 1000 simulation runs. The null hypothesis is $H_0 : C_\tau = 0.5$ or $H_0 : C_\tau^* = 0.5$.

		C_τ	C_τ^*
True Predictor	Estimate	0.841	0.834
	Bias	0.000	0.000
	ESD	0.026	0.018
	ASE	0.025	0.018
	CP	0.928	0.929
	Reject H_0	1	1
Correlated Factors	Estimate	0.549	0.547
	Bias	0.003	0.000
	ESD	0.045	0.030
	ASE	0.043	0.030
	CP	0.940	0.947
	Reject H_0	0.230	0.356
Independent Factors	Estimate	0.500	0.500
	Bias	0.000	0.000
	ESD	0.045	0.030
	ASE	0.044	0.030
	Cov	0.940	0.948
	Reject H_0	0.060	0.052
FDR		0.199	0.126

Table 2.

Performance of the C_τ and C_τ^* indices for intermediate disease that can progress or regress. 1000 simulation data sets of 150 subjects were used. Predictors W_i and Z_i are correlated with Y_i , the true predictor used to simulate the disease process, with correlation coefficient of .9 and 1/6, respectively. C represents the event-specific C index for disease progression. The C_τ^* index uses both disease progression and regression events with outcome-specific weights of 1 and 0.75, respectively. Both indices were evaluated at $\tau = 25$. δ_1 is the difference in concordance indices for Cox models using Y_i versus W_i , respectively. δ_2 is the difference in concordance indices from Cox models using Y_i and Z_i versus Y_i alone. Negative and positive signs for the bias indicates underestimation and overestimation, respectively. ESD is the empirical standard deviation; ASE is the asymptotic standard error; CP is the coverage probability. Hypothesis tests are one-sided test for $H_0 : \delta_1 = 0$ or $H_0 : \delta_2 = 0$.

		C_τ	C_τ^*
1	Estimate	0.0402	0.0483
	Bias	0.0005	0.0007
	ESD	0.0190	0.0141
	ASE	0.0188	0.0140
	CP	0.949	0.945
	Reject H_0	0.594	0.929
2	Estimate	0	0
	Bias	0.0014	0.0012
	ESD	0.0037	0.0028
	ASE	0.0037	0.0029
	CP	0.978	0.973
	Reject H_0	0.022	0.027

Table 3.

Evaluating discovery rate of C_τ and C_τ^* indices for an illness-death process. 1000 simulation data sets of 150 subjects were used. 10 candidate predictors are evaluated: 5 correlated and 5 independent of the disease process. Time to illness and death from illness are 50% and 80% right-censored, respectively. For C_τ , only disease from illness was used. The C_τ^* index uses both illness and subsequent death with outcome-specific weights of 0.75 and 1, respectively. Both indices were evaluated at $\tau = 25$. Negative and positive signs for the bias indicates underestimation and overestimation, respectively. ESD is the empirical standard deviation; ASE is the asymptotic standard error; CP is the coverage probability. False discovery rate (FDR) is defined as the number of independent predictors that reject H_0 divided by the number of all predictors that reject H_0 , averaged over 1000 simulation runs. The null hypothesis is $H_0 : C_\tau = 0.5$ or $H_0 : C_\tau^* = 0.5$.

		C_τ	C_τ^*
True Predictor	Estimate	0.835	0.819
	Bias	0.002	0.002
	ESD	0.037	0.036
	ASE	0.034	0.030
	CP	0.906	0.936
	Reject H_0	1	1
Correlated Factors	Estimate	0.550	0.547
	Bias	0.002	0.002
	ESD	0.057	0.046
	ASE	0.055	0.044
	CP	0.935	0.953
	Reject H_0	0.180	0.231
Independent Factors	Estimate	0.500	0.500
	Bias	-0.001	0.000
	ESD	0.058	0.047
	ASE	0.056	0.045
	Cov	0.937	0.956
	Reject H_0	0.063	0.044
FDR		0.258	0.136

Table 4.

Performance of the C_τ and C_τ^* indices for an illness-death process. 1000 simulation data sets of 150 subjects were used. Predictors W_i and Z_i are correlated with Y_i , the true predictor used to simulate the disease process, with correlation coefficient of .9 and 1/6, respectively. C_τ represents the event-specific C_τ index for death from illness. The C_τ^* index uses both illness and subsequent death with outcome-specific weights of 0.75 and 1, respectively. Both indices were evaluated at $\tau = 25$. δ_1 is the difference in concordance indices for Cox models using Y_i versus W_i respectively. δ_2 is the difference in concordance indices from Cox models using Y_i and Z_i versus Y_i alone. Negative and positive signs for the bias indicates underestimation and overestimation, respectively. ESD is the empirical standard deviation; ASE is the asymptotic standard error; CP is the coverage probability. Hypothesis tests are one-sided test for $H_0: \delta_1 = 0$ or $H_0: \delta_2 = 0$.

		C	C^*
1	Estimate	0.0384	0.0487
	Bias	0	0.0107
	ESD	0.0256	0.0283
	ASE	0.0245	0.0263
	CP	0.942	0.964
	Reject H_0	0.362	0.473
2	Estimate	0	0
	Bias	0.0020	0.0020
	ESD	0.0070	0.0107
	ASE	0.0069	0.0091
	CP	0.974	0.981
	Reject H_0	0.026	0.019

Table 5.

Comparison of concordance indices evaluating model discrimination for progression to type II diabetes (T2D) among African Americans with impaired glucose resistance enrolled in the placebo arm of the Diabetes Prevention Program. Predictions used are based on models for demographic covariates, treatment assignment, and conventional clinical measurements (standard model) and models for demographic covariates, treatment assignment, and all clinical measurements (new model). The added predictive value () is estimated as the difference between the concordance indices for the two sets of predictions. $SE()$ is the standard error for the estimated added predictive value. The reported p-value is for the null hypothesis that the concordance indices for the two sets of predictions are the same. Harrell's concordance (C_τ) index is used to evaluate the progression to diabetes alone; the prioritized concordance (C_τ^*) index is used to evaluate progression to diabetes and regression to normal glucose resistance (NGR) together, with outcome specific weights of 1 and 0.56, respectively. The value of τ for both indices is 6 years.

Events	Standard Model	New Model	Added predictive value		
				$SE()$	p-value
C_τ : T2D	0.719	0.735	0.016	0.012	0.100
C_τ^* : T2D+NGR	0.713	0.731	0.018	0.011	0.049

Table 6.

Evaluation of discrimination ability of lung-cancer models in African-Americans, Hispanic-Americans, and Asian-Americans in the control arm of the National Lung Screening Trial. We evaluate three models for lung-cancer incidence (Bach, PLCOm2012, and LCRAT) and one model for lung-cancer death (LCDRAT). Outcomes are lung-cancer incidence and death. The C_τ index is used to evaluate lung-cancer death ($C_{1\tau}$) and lung-cancer incidence ($C_{2\tau}$) separately. The C_τ^* index is used to evaluate lung-cancer incidence and death together, with death considered the higher priority outcome (weights set to 1 for both outcomes). The value of τ for both indices is 7 years for African-Americans and Hispanic-Americans and six years for Asian-Americans (as estimates became unstable at τ equal to seven years for Asian-Americans).

Race/Ethnicity	Model	$C_{1\tau}$	$SE(C_{1\tau})$	$C_{2\tau}$	$SE(C_{2\tau})$	C_τ^*	$SE(C_\tau^*)$
non-Hispanic African-Americans	Bach	0.603	0.040	0.623	0.035	0.620	0.036
	PLCOm2012	0.610	0.041	0.664	0.032	0.664	0.033
	LCRAT	0.625	0.041	0.662	0.033	0.661	0.033
	LCDRAT	0.633	0.042	0.671	0.033	0.670	0.034
Hispanic-Americans	Bach	0.874	0.036	0.790	0.060	0.809	0.047
	PLCOm2012	0.884	0.054	0.749	0.065	0.805	0.067
	LCRAT	0.880	0.057	0.733	0.066	0.795	0.070
	LCDRAT	0.882	0.053	0.734	0.065	0.795	0.070
Asian-Americans	Bach	0.653	0.086	0.700	0.054	0.700	0.055
	PLCOm2012	0.671	0.080	0.731	0.048	0.730	0.049
	LCRAT	0.657	0.082	0.727	0.049	0.727	0.049
	LCDRAT	0.643	0.083	0.721	0.049	0.721	0.049