# Application of concordance probability estimate to predict conversion from mild cognitive impairment to Alzheimer's disease

Xiaoxia Han, Yilong Zhang, Yongzhao Shao & the Alzheimer's Disease Neuroimaging Initiative

Taylor & Francis
Taylor & Francis Group

Check for updates

# Application of concordance probability estimate to predict conversion from mild cognitive impairment to Alzheimer's disease

Xiaoxia Han[a], Yilong Zhang[b], Yongzhao Shao[a] and
the Alzheimer's Disease Neuroimaging Initiative[a*]

[a]Department of Population Health, New York University School of Medicine, New York, USA; [b]Merck Research Laboratories, Rahway, NJ, USA

**ABSTRACT**

Subjects with mild cognitive impairment (MCI) have a substantially increased risk of developing dementia due to Alzheimer's disease (AD). Identifying MCI subjects who have high progression risk to AD is important in clinical management. Existing risk prediction models of AD among MCI subjects generally use either the AUC or Harrell's C-statistic to evaluate predictive accuracy. AUC is aimed at binary outcome and Harrell's C-statistic depends on the unknown censoring distribution. Gönen and Heller's K-index, also known as concordance probability estimate (CPE), is another measure of overall predictive accuracy for Cox proportional hazards (PH) models, which does not depend on censoring distribution. As a comprehensive example, using Alzheimer's Disease Neuroimaging Initiative (ADNI) data-set, we built a Cox PH model to predict the conversion from MCI to AD where the prognostic accuracy was evaluated using K-index.

## 1. Introduction

Alzheimer's disease (AD) is the most common cause of dementia. It is projected that the number of Americans with AD will increase from 5.4 million in 2016 to 13.8 million in 2050 [1]. Subjects with mild cognitive impairment (MCI) have a substantially increased risk of developing dementia due to AD. Several studies have suggested that patients with MCI will convert to AD at an annual conversion rate of 10%–15% [2–4]. Early identification of subjects with MCI who are at risk of progression to AD is of great clinical importance in delaying or preventing the transition from MCI to AD.

A major focus of MCI research has been to distinguish individuals who will progress to AD from those who will not [5,6]. Several studies have developed prediction models for MCI to AD conversion using positron emission tomography (PET) images [7, 8],

magnetic resonance images (MRI) [9–12], cerebrospinal fluid (CSF) biomarkers [13,14], or combining MRI and CSF measures [15,16]. Most of these studies employed logistic regression or other classification methods to investigate AD progression at certain year (e.g. 5-year or 10-year), where area under the receiver operating characteristic curve (AUC) was used to evaluate overall predictive accuracy. Yet, AUC depends on the year selected for evaluation. An overall evaluation of the predictive accuracy may be more informative for time to AD conversion from MCI.

There are existing studies that focused on developing risk scores for AD progression using Cox proportional hazards (PH) regression to accommodate time-to-progression from MCI to AD, and the predictive accuracy of the risk scores were evaluated using Harrell's C-statistic [17–19]. It is known that the Harrell's C-statistic approach has a major drawback as it is generally biased and the magnitude of the bias depends on the unknown censoring distribution even asymptotically [20,21]. Inverse probability-of-censoring weighted (IPCW) C-statistic has been proposed by Uno et al. [21] and Liu, Jin [22] to overcome the bias of the C-statistic. The IPCW C-statistic is consistent to C-index when censoring is 'noninformative' (i.e. the random censoring time and AD-free survival time are independent). However, this 'noninformative censoring' assumption is usually not satisfied in practice as both random censoring time and AD-free survival time can depend on certain covariates in many applications. Another approach to overcome the dependence of the Harrell's C-statistic on censoring distribution has been proposed by Gönen and Heller [23], which is particularly useful when the widely used Cox PH model is assumed.

In short, when the commonly used Cox PH models are suitable, K-index can be useful to evaluate the overall predicative accuracy for risk prediction models of AD among MCI subjects. The rest of this paper is organized as follows. In Section 2, we briefly review Harrell's C-statistic, the IPCW C-statistic and K-index. In Section 3, we report simulation studies to compare the performance of these three methods for Cox PH models and provide more insights. In Section 4, we provide a comprehensive example, i.e. we develop an AD risk prediction model among MCI subjects based on Cox PH model using Alzheimer's Disease Neuroimaging Initiative (ADNI) data-set. We evaluate the risk score performance using K-index. The paper is concluded with discussion in Section 5.

## 2. Method

For a randomly selected individual, let $T$, $D$, and $R$ denote event time, censoring time, and risk score, respectively. We assume a high risk score generally corresponds to short survival. Let $\tilde{T} = \min(T, D)$ denote the observed time and $\delta = I(T < D)$ denote the event indicator under right censoring. For a pair of randomly selected independent subjects with $(T_1, R_1)$ and $(T_2, R_2)$, the C-index can be denoted as

$$C = P(R_1 > R_2 \mid T_1 < T_2),$$

and K-index can be defined as

$$K = P(T_1 < T_2 \mid R_1 > R_2).$$

C-index and K-index may take value ranging from 0.5 to 1. When K or C equals to 1, $R_1 > R_2$ implies $T_1 < T_2$ with probability 1 and vice versa. When K or C equals to 0.5, $R_1 > R_2$ implies $T_1 < T_2$ with probability 0.5 (i.e. similar to predicting an outcome via flipping a fair coin) and vice versa.

In the special case where $T$ and $R$ are continuous variables (i.e. the probability of tied observations can be ignored), and two randomly selected subjects with $(T_1, R_1)$ and $(T_2, R_2)$ are independent and identically distributed (iid), we may have

$$P\ (T_1 < T_2) = P\ (T_1 > T_2) = 1/2 \text{ and } P\ (R_1 < R_2) = P\ (R_1 > R_2) = 1/2.$$

Then, by the basic rule of conditional probability, we have

$$K = P\ (T_1 < T_2 \,|\, R_1 > R_2) = \frac{P(T_1 < T_2,\ R_1 > R_2)}{P(R_1 > R_2)} = 2P(T_1 < T_2, R_1 > R_2)$$

$$= \frac{P(T_1 < T_2,\ R_1 > R_2)}{P(T_1 < T_2)} = P\ (R_1 > R_2 \,|\, T_1 < T_2) = C \tag{1}$$

That is, the true value of K-index and C-index are equal under the above condition.

For independently observed data $(\tilde{T}_i,\ \delta_i,\ R_i),\ i = 1,\ \ldots,\ n$, Harrell's C-statistic [24] is defined as

$$\hat{C}_H = \frac{\sum_{i \neq j} I(\tilde{T}_i < \tilde{T}_j) I(R_i > R_j)}{\sum_{i \neq j} \delta_i I(\tilde{T}_i < \tilde{T}_j)} \tag{2}$$

Harrell's C-statistic is known to be a biased estimator for C-index when there is censoring, since it depends on censoring distribution [20,21]. Therefore, even for the same population of subjects, when the censoring distributions differ in different studies, the values of the Harrell's C-statistics are not comparable. The Harrell's C-statistic and its standard error estimate can be directly obtained using the `coxph()` function from R package `survival`.

To overcome the shortcoming of Harrell's C-statistic, some new procedures have been proposed. Notably, Uno et al. [21] and Liu, Jin [22] independently investigated the so-called inverse-probability-of-censoring-weighted (IPCW) C-statistic. The IPCW C-statistic is defined as

$$\hat{C}_W = \frac{\sum_{i \neq j} \delta_i \{\hat{G}(\tilde{T}_i)\}^{-2} I(\tilde{T}_i < \tilde{T}_j) I(R_i > R_j)}{\sum_{i \neq j} \delta_i \{\hat{G}(\tilde{T}_i)\}^{-2} I(\tilde{T}_i < \tilde{T}_j)}, \tag{3}$$

where $\hat{G}(\cdot)$ is the Kaplan–Meier estimator of the censoring distribution. These weighted approaches are aimed at making the C-statistic independent of the underlying unknown censoring distribution. However, the consistency of such weighted version of the C-statistic $\hat{C}_W$ depends on quite restrictive assumptions that event time $T$ (i.e. the conversion time to AD from MCI) is independent of censoring time $D$. Such assumptions are usually not met in practice because the censoring time often depends on covariates that correlate to the

survival endpoint. The IPCW C-statistic and its standard error estimate can be calculated using `survC1()` function in R package `survC1`.

Gönen and Heller [23] proposed K-index as a measure of discriminatory power under the commonly used Cox PH model. As is well known, the Cox PH models are among the most widely used survival models and statistical tests for the PH assumptions have been well developed. The Cox PH model can be written as

$$\lambda\ (t|z) = \lambda_0(t)\exp\left(\beta_0^T z\right), \tag{4}$$

where $\lambda(t|z)$ is the hazard function conditional on a $p$-dimensional covariate vector $z$, $\lambda_0(t)$ is the baseline hazard function independent of covariate, and $\beta_0$ is the true regression parameter. The relationship between the covariate vector $z$ and the survival time $t$ is determined through the survival function,

$$S(t; z, \beta) = \exp\left\{-\exp\left(\beta^T z\right)\int_0^t \lambda_0(u)du\right\} \tag{5}$$

The Gönen and Heller's K-index is also known as concordance probability estimate (CPE). Under Cox PH moels, the CPE is defined as

$$K_n\ (\hat{\beta}) = \frac{2}{n(n-1)}\ \sum_i\sum_{<j}\left\{\frac{I\left(\hat{\beta}^T z_{ji} < 0\right)}{1 + \exp\left(\hat{\beta}^T z_{ji}\right)} + \frac{I\left(\hat{\beta}^T z_{ij} < 0\right)}{1 + \exp\left(\hat{\beta}^T z_{ij}\right)}\right\}, \tag{6}$$

where $z_{ij}$ is the pairwise difference $z_i - z_j$ and $\hat{\beta}$ is the partial likelihood estimator for $\beta$ in Cox PH models. Gönen and Heller also proposed a smooth version of the K-index, and more details can be found in [23]. K-index is a function of the estimated regression parameter in Cox PH models, and therefore it is asymptotically consistent. Thus, even for finite sample, K-index is generally not sensitive to censoring distribution. The K-index and its standard error estimate can be estimated using the `phcpe()` function from the R package `CPE` by directly inputting the coxph object from `coxph()` function in argument `coxfit` and setting argument `CPE.SE = TRUE`. Another function `phcpe2()` in R package `CPE` allows to estimate K-index and the standard error by inputting the coefficients and covariance matrix of the coefficients from the fitted Cox PH model and a design matrix for covariates. Since K-index is suitable for Cox PH models, it is important to test the validity of the PH assumption. The PH assumption can be checked with scaled Schoenfeld residuals for each parameter using the `cox.zph()` function in the R package `survival`. All analyses in this paper were done using R v3.2.2.

## 3. Simulation studies

To provide insights on the behaviours of the three concordance indices under Cox PH models, we conducted extensive simulation studies to investigate their variability across different censoring proportions ranging from 0% to 80%. We consider non-informative random right censoring, that is the censoring, time is independent of the time-to-event.

This is a strong assumption, however, even under this restrictive assumption the Harrell's C-statistic and its weighted versions using inverse probabilities are still quite sensitive to the censoring distribution, which is unknown in practical applications. Following the Cox PH model in Equation (4), the covariate vector was $z = (Z_1, Z_2)$, where $Z_1$ and $Z_2$ were independently generated from Normal (0, 1) and Bernoulli (0.5), respectively. The corresponding log(HR) was $\beta = (\beta_1, \beta_2)$, where $\beta_1 = 1$ and $\beta_2 = -1$. We assumed the baseline hazard function follows an exponential distribution with a constant hazard rate $\lambda_0(t) = 1$. Given covariates, the event time $T$ was generated from exponential distribution ($\lambda$) with parameter $\lambda = \exp(\beta_1 Z_1 + \beta_2 Z_2)$. In our simulations, censoring time $D$ was independent of $T$ and generated from Uniform $[0, \tau]$, where $\tau$ were chosen to achieve the desired censoring proportion ranging from 0% to 80%.

Under the above set up, we obtained iid copies of $(\tilde{T}, \delta, z)$, where $\tilde{T} = \min(T, D)$ and $\delta = I(T < D)$. The corresponding risk score is $R = \beta_1 Z_1 + \beta_2 Z_2$. By the iid assumption and the fact that $T$ and $R$ are continuous variables (thus the probability of tied observations is negligible), for a pair of randomly selected individuals with bivariate $(T_1, R_1)$ and $(T_2, R_2)$, Equation (1) holds, i.e. the true value of K-index and C-index are equal. In our simulation set up, the true value is K-index = C-index = 0.744. It is well known that the Gönen and Heller's K-index is a consistent estimator of the true K regardless of censoring proportion under the Cox PH models, and Harrell's C-statistic is consistent estimate of the true value only in the absence of censoring. Theoretically, under the non-informative censoring assumption, the IPCW C-statistic can consistently estimate the true C-index value, but its finite sample performance in comparison with K-index and Harrell's C-statistic has not been systematically investigated and reported. The aim of this simulation study is to investigate the relative stability of the K-index and relative instability of Harrell's C-statistic, IPCW C-statistic. We computed the Harrell's C-statistic, IPCW C-statistic and K-index using the R functions mentioned in the method section. We used various sample size ranging from $n = 200$ to $n = 400$. The key results and observed patterns were not sensitive to sample size and therefore we only reported results for $n = 200$ and $n = 400$ in the following.

As shown in Figure 1, in the absence of censoring, the three estimators all performed well as expected. For sample size $n = 200$, as censoring proportion increased, the Harrell's C-statistic increased noticeably which reflected its increased biases. The IPCW C-statistic remained stable for relative light censoring, but it started to increase when censoring became heavier. On the other hand, K-index remained stable when the censoring proportion was as high as 80%. When sample size increased to $n = 400$, the behaviour and bias of Harrell's C-statistic and IPCW C-statistic did not improve. More numerical details can be found in Table 1.

## 4. Application to the ADNI data

The data source of this article was the Alzheimer's Disease Neuroimaging Initiative database (https://ida.loni.usc.edu/). ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations, as a $60 million, five-year public-private partnership. The principal investigator of the initiative is Michael W. Weiner, M.D., VA medical Center and
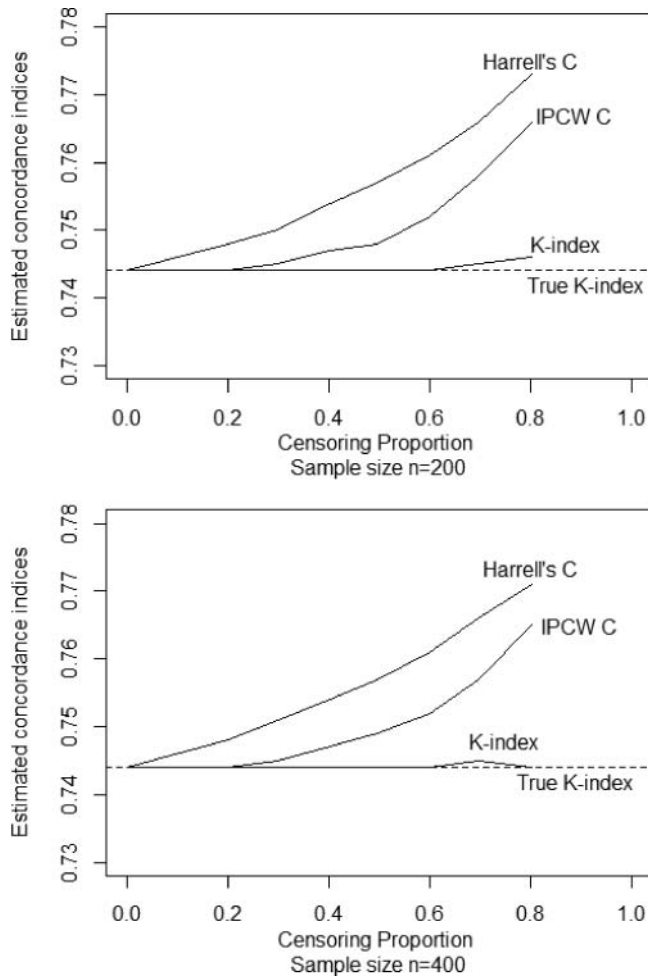
**Figure 1.** Estimated concordance indices Harrell's C-statistic, IPCW C-statistic and K-index for Cox PH model under different censoring proportions with 1000 simulation replicates. The solid curves are the mean of 1000 simulation replicated for each of the estimated concordance indices. The dash line is the true value 0.744.

University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions as well as private corporations. The study subjects have been recruited from over 50 sites across the USA and Canada. The initial goal of ADNI was to recruit 800 subjects and ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal (NL) older individuals, individuals with MCI, and individual with AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. Up-to-date information can be found at http://www.adni-info.org/.

In this real data example, we only consider subjects with baseline diagnosis as MCI. The diagnosis results were downloaded from ADNI website on April 28, 2016. We focus

**Table 1.** Simulation study. Under the Cox PH model in Equation (5), the covariate was $z = (Z_1, Z_2)$, where $Z_1$ and $Z_2$ were independently generated from Normal (0, 1) and Bernoulli (0.5), respectively. The corresponding log(HR) was $\beta = (\beta_1, \beta_2)$, where $\beta_1 = 1$ and $\beta_2 = -1$. We assumed the baseline hazard function has a constant hazard rate $\lambda_0(t) = 1$. Thus, given covariates, the event time $T$ was generated from exponential distribution ($\lambda$) with parameter $\lambda = \exp(\beta_1 Z_1 + \beta_2 Z_2)$. Censoring time $D$ was independent of $T$ and was generated from Uniform $[0, \tau]$, where $\tau$ were chosen to achieve the desired censoring proportion ranging from 0% to 80%.

| Sample size | Censoring proportion | Harrell's C | Harrell's C simulation SE | IPCW C | IPCW C simulation SE | K-index | K-index simulation SE |
|---|---|---|---|---|---|---|---|
| $n = 200$ | 80% | 0.773 | 0.039 | 0.766 | 0.045 | 0.746 | 0.027 |
| | 70% | 0.766 | 0.031 | 0.758 | 0.035 | 0.745 | 0.023 |
| | 60% | 0.761 | 0.027 | 0.752 | 0.028 | 0.744 | 0.020 |
| | 50% | 0.757 | 0.024 | 0.748 | 0.024 | 0.744 | 0.019 |
| | 40% | 0.754 | 0.023 | 0.747 | 0.022 | 0.744 | 0.018 |
| | 30% | 0.750 | 0.021 | 0.745 | 0.020 | 0.744 | 0.017 |
| | 20% | 0.748 | 0.020 | 0.744 | 0.019 | 0.744 | 0.017 |
| | 0% | 0.744 | 0.018 | 0.744 | 0.018 | 0.744 | 0.016 |
| $n = 400$ | 80% | 0.771 | 0.028 | 0.765 | 0.033 | 0.744 | 0.019 |
| | 70% | 0.766 | 0.022 | 0.757 | 0.026 | 0.745 | 0.016 |
| | 60% | 0.761 | 0.019 | 0.752 | 0.021 | 0.744 | 0.014 |
| | 50% | 0.757 | 0.017 | 0.749 | 0.017 | 0.744 | 0.013 |
| | 40% | 0.754 | 0.016 | 0.747 | 0.015 | 0.744 | 0.013 |
| | 30% | 0.751 | 0.015 | 0.745 | 0.014 | 0.744 | 0.012 |
| | 20% | 0.748 | 0.014 | 0.744 | 0.013 | 0.744 | 0.011 |
| | 0% | 0.744 | 0.013 | 0.744 | 0.013 | 0.744 | 0.011 |

on late onset AD, thus individuals with age at AD onset younger than 60 years (indicative of possible familial AD) were excluded. Clinical data from only non-Hispanic Caucasian subjects were used in this investigation. Conversion was defined as incident Alzheimer's disease and time to conversion was measured in years.

Table 2 summarizes the baseline characteristics of the ADNI cohort. The baseline age is significantly greater in converters than non-converters, while there is no significant difference in years of education. Converters have a higher percentage of APOE $\varepsilon 4$ carriers. Neurological disorder other than AD is significantly higher among non-converters. All the neuropsychological test scores are significantly different between converters and non-converters, except for Geriatric Depression Scale.

Table 3 summarizes Pearson correlation among neuropsychological test scores. Several pairs of neuropsychological test are highly correlated. Apolipoprotein E (APOE) $\varepsilon 4$ variant has been confirmed as a risk factor for AD [25,26]. Age, gender, education and medical history have been reported to be associated with the risk of AD in other studies [17,19,27]. We used Adaptive Lasso to select variables [29]. The candidate variables for Adaptive Lasso [28,29] selection include age, gender, education, APOE allele, medical history and neuropsychological test performance. As commonly used in applying Adaptive Lasso, ridge regression was used to obtain an initial $\hat{\beta}(\text{ridge})$ to construct the adaptive weights vector, and then we applied the adaptive weights vector on the `cv.glmnet()` function using the argument `penalty.factor`. We used 10-fold cross validation to find the optimal shrinkage factor $\lambda$, which is the largest value of $\lambda$ within one standard error of the minimum of the partial likelihood deviance. The variables selected by Adaptive Lasso include APOE $\varepsilon 4$, neurological disorder other than AD, cognitive dementia rating scale sum of boxes (CDR-SB), Alzheimer's Disease Assessment Scale cognitive subscale (ADAS-cog) 13 items, mini-mental state examination score (MMSE), and

**Table 2.** Summary of characteristics at baseline.

| Characteristic | Non-converters<br>$n = 457$ | Converters<br>$n = 275$ | $p$-value |
|---|---|---|---|
| Male, No. (%) | 274 (60) | 170 (61.8) | 0.674 |
| Age, mean (sd), y | 72.69 (7.75) | 74.56 (6.35) | <0.001 |
| Education, mean (sd), y | 16.02 (2.84) | 15.85 (2.79) | 0.41 |
| APOE $\varepsilon$4, No. (%) | | | <0.001 |
| 0 | 266 (58.2) | 89 (32.4) | |
| 1 | 153 (33.5) | 139 (50.5) | |
| 2 | 38 (8.3) | 47 (17.1) | |
| Medical history, No. (%) | | | |
| Psychiatric | 181 (39.6) | 103 (37.5) | 0.617 |
| Neurological disorder (other than AD) | 170 (37.2) | 78 (28.4) | 0.018 |
| Alcohol abuse | 17 (3.7) | 12 (4.4) | 0.813 |
| Drug abuse | 4 (0.9) | 1 (0.4) | 0.655 |
| Smoking | 182 (39.8) | 108 (39.3) | 0.944 |
| Malignancy | 108 (23.6) | 64 (23.3) | 0.983 |
| Neuropsychological test, mean (sd) | | | |
| GDS* | 1.63 (1.41) | 1.69 (1.43) | 0.627 |
| CDR-SB | 1.27 (0.73) | 1.88 (0.92) | <0.001 |
| ADAS-cog 11 items | 8.82 (3.86) | 12.8 (4.41) | <0.001 |
| ADAS-cog 13 items | 14.17 (5.86) | 20.73 (6.08) | <0.001 |
| MMSE | 28.01 (1.7) | 27.04 (1.76) | <0.001 |
| RAVLT immediate | 37.34 (10.88) | 29.08 (7.98) | <0.001 |
| RAVLT learning | 4.67 (2.51) | 3.16 (2.46) | <0.001 |
| RAVLT forgetting | 4.48 (2.54) | 4.93 (2.26) | 0.013 |
| RAVLT %forgetting | 52.96 (31.21) | 73.92 (29.6) | <0.001 |
| FAQ | 1.95 (3.08) | 5.27 (4.68) | <0.001 |

Note: Abbreviations: sd, standard deviation; GDS, Geriatric Depression Scale; CDR-SB, cognitive dementia rating scale sum of boxes; ADAS, Alzheimer's Disease Assessment Scale; MMSE, mini-mental state examination score; RAVLT, Rey Auditory Verbal Learning Test; FAQ, Functional Activities Questionnaire.
*GDS was measured during screen visit.

**Table 3.** Pearson correlation between neuropsychological test scores.

| | GDS | CDR-SB | ADAS-cog<br>11 items | ADAS-cog<br>13 items | MMSE | RAVLT<br>immediate | RAVLT<br>learning | RAVLT<br>forgetting | RAVLT %<br>forgetting | FAQ |
|---|---|---|---|---|---|---|---|---|---|---|
| GDS | 1 | 0.09 | 0 | 0 | 0.03 | 0 | 0.02 | −0.01 | −0.02 | 0.09 |
| CDR-SB | | 1 | 0.26 | 0.29 | −0.19 | −0.22 | −0.16 | 0.06 | 0.21 | 0.55 |
| ADAS-cog 11 items | | | 1 | 0.95 | −0.4 | −0.61 | −0.46 | 0.13 | 0.45 | 0.3 |
| ADAS-cog 13 items | | | | 1 | −0.45 | −0.67 | −0.53 | 0.18 | 0.53 | 0.33 |
| MMSE | | | | | 1 | 0.39 | 0.32 | −0.03 | −0.28 | −0.16 |
| RAVLT immediate | | | | | | 1 | 0.6 | −0.12 | −0.59 | −0.27 |
| RAVLT learning | | | | | | | 1 | 0.03 | −0.47 | −0.26 |
| RAVLT forgetting | | | | | | | | 1 | 0.77 | 0.09 |
| RAVLT % forgetting | | | | | | | | | 1 | 0.26 |
| FAQ | | | | | | | | | | 1 |

Functional Activities Questionnaire (FAQ). We then developed a multivariable Cox proportional hazards regression to model the time to AD conversion among MCI patients.

Table 4 summarizes the predictors and corresponding hazard ratios for the final Cox PH model. APOE $\varepsilon$4 allele is the strongest predictor (HR = 1.58, $p$-value < 0.001). All the other predictors have significant influence on the hazard ratio for AD (neurological disorder other than AD: HR = 0.638, $p$-value = 0.001, CDR-SB: HR = 1.503, $p$-value < 0.001; ADAS-cog 13 items: HR = 1.121, $p$-value < 0.001; MMSE: HR = 0.908, $p$-value = 0.01; FAQ: HR = 1.086, $p$-value < 0.001). It is important to verify the proportional hazards assumption when using Cox PH models. Table 5 summarizes the test results for PH assumption for the fitted Cox PH model. We confirm that there is no strong evidence to

**Table 4.** Multivariate Cox proportional hazard model for conversion from MCI to AD.

| Variable | HR (95% CI) | *p*-value |
|---|---|---|
| APOE $\varepsilon 4$ | 1.58 (1.33, 1.877) | <0.001 |
| Neurological disorder (other than AD) | 0.638 (0.487, 0.835) | 0.001 |
| CDR-SB | 1.503 (1.291, 1.748) | <0.001 |
| ADAS-cog 13 | 1.121 (1.098, 1.145) | <0.001 |
| MMSE | 0.908 (0.842, 0.979) | 0.011 |
| FAQ | 1.086 (1.056, 1.117) | <0.001 |

**Table 5.** Test results of the proportional hazards assumption for the fitted multivariable Cox regression model.

| Variable | rho | *p*-value |
|---|---|---|
| APOE $\varepsilon 4$ | 0.112 | 0.063 |
| Neurological disorder other than AD | 0.08 | 0.188 |
| CDR-SB | −0.028 | 0.656 |
| ADAS-cog 13 | −0.017 | 0.788 |
| MMSE | −0.031 | 0.606 |
| FAQ | 0.001 | 0.992 |
| GLOBAL | | 0.496 |

reject the null hypothesis of a correlation coefficient $rho = 0$ for any parameter in the model with time. Furthermore, the *p*-value for the global chi-square test for PH assumption is 0.496. With the relatively small number of baseline predictors in the Cox PH model, the prognostic accuracy of the model is very good with a K-index 0.77 (95% CI 0.75–0.79). Using training and evaluation sets by random splitting the data-set yields similar results, e.g. using 90% data to develop the Cox PH model and the remaining 10% as evaluation set, based on 1000 replications, the estimated K-index has mean 0.77 with standard error 0.01. On the other hand, the Harrell's C-statistic is 0.84 (95% CI 0.80– 0.87). The inflated C-statistic value under heavy censoring is in agreement with our simulation results presented in Figure 1, and also similar to the real data example presented in Gönen and Heller [23].

## 5. Discussion

Several risk prediction tools have been developed to predict the conversion from MCI to AD recently. Many of them utilized MRI image, PET image, CSF biomarkers, or a combination of them in the prediction model [7–16]. Most of these studies employed logistic regression or other classification methods and used AUC to evaluate predictive accuracy at a certain year after MCI. Yet, MCI to AD conversion is a time to event data, and an overall evaluation of the predictive accuracy is of interest and may be more informative. Some of the biomarkers and radiographic imaging evidence can be quite expensive and often invasive. A few studies have developed non-invasive risk scores for AD progression using Cox PH regression, and the accuracy was evaluated using Harrell's C statistic [17– 19]. It is well known Harrell's C statistic depends on the unknown censoring distribution, which diminishes its interpretability and practical utility. On the other hand, Cox PH models are widely applicable, and Gönen and Heller's K-index can be used when Cox PH assumption is satisfied.

As a semi-parametric method, Cox PH model can be fitted without estimating the baseline hazard function. The coefficients fitted by Cox PH model (i.e. $\hat{\beta}$ ) reflect relative risks. Based on such estimated coefficients from the fitted Cox PH model, we can estimate concordance probabilities as overall measures of discriminative power. On the other hand, to obtain the absolute risk using Cox PH models, we would need to estimate the baseline hazard function. In general, the baseline hazard functions cannot be consistently estimated in outcome dependent sampling studies or non-prospective studies. ADNI is an outcome-dependent sampling study, and thus the baseline hazard function may not be reliable estimated. If survival data based on well-designed prospective studies is available, the baseline hazard function $\lambda_0(t)$ and baseline cumulative hazard function $\Lambda_0(t)$ can be estimated using `basehaz()` function in R package `survival`. When a new MCI patient comes in with relevant covariates in the Cox PH model, we can use the fitted Cox PH model to estimate the hazard function $\lambda\,(t|z,\hat{\beta},\hat{\lambda}_0(t)) = \hat{\lambda}_0\,(t)\exp{(\hat{\beta}^T z)}$. The cumulative hazard function $\Lambda\,(t|z,\hat{\beta},\,\hat{\lambda}_0(t)) = \hat{\Lambda}_0\,(t)\exp{(\hat{\beta}^T z)}$ can then be estimated by integrating $\lambda\,(t|z,\hat{\beta},\,\hat{\lambda}_0(t)) = \hat{\lambda}_0\,(t)\exp{(\hat{\beta}^T z)}$. Then one can estimate the probability from MCI to conversion to AD at time $t$ by

$$S\,(t|z,\hat{\beta},\,\hat{\lambda}_0(t)) = \exp{[-\Lambda(t|z,\hat{\beta},\,\hat{\lambda}_0(t))]} = \exp{[-\hat{\Lambda}_0(t)\exp{(\hat{\beta}^T z)}]}.$$

Although $\beta$ is a vector of finite dimension and easy to estimate, the baseline hazard function $\lambda_0(t)$ generally has infinite dimension. Thus, a relative large sample size is required to obtain a reliable estimate of $\lambda_0(t)$. In addition, careful model calibration is usually required.

K-index for Cox PH models is analogous to the area under the ROC curve (AUC) for logistic regression to classify a binary outcome such as disease and no disease. The AUC for logistic regression can be obtained using either retrospective case-control data or prospective data. On the other hand, the positive predictive value and negative predictive value depend on disease prevalence in the underlying study population which cannot be consistently estimated based only on retrospective case-control data. While absolute risk is useful, it is quite expensive to estimate, and in many applications, we do not necessarily need to know the absolute risk for patients. The relative risks estimated from Cox PH model have been widely used in clinical study design and decision making. For example, the relative risk score (i.e. $\hat{\beta}^T z$) obtained from Cox PH model can be useful to select a subgroup of MCI patients with high risk of conversion to AD into clinical trials.

Unlike the Harrell's C-statistic, K-index is robust to the unknown censoring distribution. K-index also has simple closed-form formula thus easy to compute. Moreover, it is straightforward to compare two risk score systems by comparing the corresponding K-indices. In particular, one can calculated the difference of the two K-indices with bootstrapped confidence intervals [30]. These confidence intervals are easy to obtain and straightforward for interpretation. In contrast, the difference of two C-statistics can be shown to depend on the unknown censoring distribution in general [21,31], and thus it does not have much practical utility. As discussed by Gönen and Heller [23], K-index can be viewed as an extension of the AUC for

binary outcome to censored time-to-event outcome under Cox PH models. When proportional odds (PO) models and other transformational models are more suited than Cox PH models for the data-set, extended K-index in Zhang and Shao [30] might be used in similar fashion to evaluate predictive accuracy of related models for disease progression.

Although we have shown some advantages of K-index compared to Harrell's C-statistic and IPCW C-statistic, the K-index also has some limitations. First, it is suitable only under Cox PH models. If the PH assumption is not satisfied, K-index may give misleading estimates. Second, the K-index itself does not reflect absolute risk. Another limitation of our study is that it is possible that other cohort does not have the neuropsychological tests we selected using the ADNI data. Another model may have to be established using other available test scores in similar manner and K-index can be used to measure the overall discriminative power. Nevertheless, our results indicate that the more easily available cognitive tests instead of more expensive (and possibly invasive) biomarkers can have great potential to facilitate the selection of the 'right' MCI population for future clinical trials.

To sum up, we suggest using K-index to evaluate overall predictive accuracy when PH assumption is satisfied. As a comprehensive example, we developed a risk score to predict the progression from MCI to AD using a Cox PH model, and applied K-index to evaluate the fitted Cox PH model after confirmed the PH assumption was satisfied. One major strength of this risk score is the simplicity. It only includes six simple items, i.e. APOE $\varepsilon 4$ allele, medical history-neurological disorder other than AD, cognitive dementia rating scale sum of boxes (CDR-SB), Alzheimer's Disease Assessment Scale cognitive subscale (ADAS-Cog) 13 items, mini-mental state examination score (MMSE), and Functional Activities Questionnaire (FAQ). In addition, this risk score is cost-effective. APOE $\varepsilon 4$ allele, medical history and neuropsychological test performance are relative easy to gather, compared to the more expensive MRI scan, PET scan, as well as CSF biomarkers which requires invasive lumbar puncture. Interestingly, the accuracy of this risk score was very good with a K-index 0.77. Our results highlight the important role of neuropsychological tests in terms of prediction of Alzheimer's disease risk for MCI patients.

## Acknowledgments

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

## Notes on contributors

*Xiaoxia Han* is a PhD student in Biostatistics Division, Department of Population Health, New York University School of Medicine, New York, NY.

*Yilong Zhang* is a senior scientist in BARDS at Merck & Co., Inc., Rahway, NJ.

*Yongzhao Shao* is a professor, Division of Biostatistics, Department of Population Health, New York University School of Medicine, New York, NY. Dr Yongzhao is the core director of NYU Alzheimer's Disease Center Data and Statistics Core.

## References

[1]  Alzheimer's A. Alzheimer's disease facts and figures. Alzheimers Dement. 2016; 12(4):459–509. PubMed PMID: 27570871.
[2]  Petersen RC, Smith GE, Waring SC, et al. Mild cognitive impairment: clinical characterization and outcome. Arch Neurol. 1999;56(3):303–308. PubMed PMID: 10190820.
[3]  Hanninen T, Hallikainen M, Tuomainen S, et al. Prevalence of mild cognitive impairment: a population-based study in elderly subjects. Acta Neurol Scand. 2002;106(3):148–154. PubMed PMID: 12174174.
[4]  Blasko I, Jellinger K, Kemmler G, et al. Conversion from cognitive health to mild cognitive impairment and Alzheimer's disease: prediction by plasma amyloid beta 42, medial temporal lobe atrophy and homocysteine. Neurobiol Aging. 2008;29(1):1–11. doi:10.1016/j.neurobiolaging.2006.09.002. PubMed PMID: 17055615.
[5]  Morris JC, Storandt M, Miller JP, et al. Mild cognitive impairment represents early-stage Alzheimer disease. Arch Neurol. 2001;58(3):397–405. PubMed PMID: 11255443.
[6]  Tierney MC, Szalai JP, Snow WG, et al. Prediction of probable Alzheimer's disease in memory-impaired patients: a prospective longitudinal study. Neurology. 1996;46(3):661–665. PubMed PMID: 8618663.
[7]  Lange C, Suppa P, Frings L, et al. Optimization of statistical single subject analysis of brain FDG PET for the prognosis of mild cognitive impairment-to-Alzheimer's disease conversion. J Alzheimers Dis. 2016;49(4):945–959. doi:10.3233/JAD-150814. PubMed PMID: 26577523.

[8] Zhang S, Han D, Tan X, et al. Diagnostic accuracy of 18 F-FDG and 11 C-PIB-PET for prediction of short-term conversion to Alzheimer's disease in subjects with mild cognitive impairment. Int J Clin Pract. 2012;66(2):185–198. doi:10.1111/j.1742-1241.2011.02845.x. PubMed PMID: 22257044.

[9] Tong T, Gao Q, Guerrero R, et al. A novel grading biomarker for the prediction of conversion from mild cognitive impairment to Alzheimer's disease. IEEE Trans Biomed Eng. 2017;64 (1):155–165. doi:10.1109/TBME.2016.2549363. PubMed PMID: 27046891.

[10] Wei R, Li C, Fogelson N, et al. Prediction of conversion from mild cognitive impairment to Alzheimer's disease using MRI and structural network features. Front Aging Neurosci. 2016;8:76. doi:10.3389/fnagi.2016.00076. PubMed PMID: 27148045; PubMed Central PMCID: PMCPMC4836149.

[11] Liu H, Zhou X, Jiang H, et al. A semi-mechanism approach based on MRI and proteomics for prediction of conversion from mild cognitive impairment to Alzheimer's disease. Sci Rep. 2016;6:26712. doi:10.1038/srep26712. PubMed PMID: 27273250; PubMed Central PMCID: PMCPMC4896009.

[12] Nesteruk M, Nesteruk T, Styczynska M, et al. Predicting the conversion of mild cognitive impairment to Alzheimer's disease based on the volumetric measurements of the selected brain structures in magnetic resonance imaging. Neurol Neurochir Pol. 2015;49(6):349–353. doi:10.1016/j.pjnns.2015.09.003. PubMed PMID: 26652867.

[13] Koppara A, Wolfsgruber S, Kleineidam L, et al. The latent dementia phenotype delta is associated with cerebrospinal fluid biomarkers of Alzheimer's disease and predicts conversion to dementia in subjects with mild cognitive impairment. J Alzheimers Dis. 2016;49(2):547–560. doi:10.3233/JAD-150257. PubMed PMID: 26484902.

[14] Papaliagkas VT, Anogianakis G, Tsolaki MN, et al. Prediction of conversion from mild cognitive impairment to Alzheimer's disease by CSF cytochrome c levels and N200 latency. Curr Alzheimer Res. 2009;6(3):279–284. PubMed PMID: 19519309.

[15] Ewers M, Walsh C, Trojanowski JQ, et al. Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. Neurobiol Aging. 2012;33(7):1203–1214. doi:10.1016/j.neurobiolaging.2010.10.019. PubMed PMID: 21159408; PubMed Central PMCID: PMCPMC3328615.

[16] Westman E, Muehlboeck JS, Simmons A. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. Neuroimage. 2012;62(1):229–238. doi:10.1016/j.neuroimage.2012.04.056. PubMed PMID: 22580170.

[17] Anstey KJ, Cherbuin N, Herath PM, et al. A self-report risk index to predict occurrence of dementia in three independent cohorts of older adults: the ANU-ADRI. PLoS One. 2014;9(1): e86141. doi:10.1371/journal.pone.0086141. PubMed PMID: 24465922; PubMed Central PMCID: PMCPMC3900468.

[18] Barnes DE, Beiser AS, Lee A, et al. Development and validation of a brief dementia screening indicator for primary care. Alzheimers Dement. 2014;10(6):656–665.. doi:10.1016/j. jalz.2013.11.006. PubMed PMID: 24491321; PubMed Central PMCID: PMCPMC4119094.

[19] Exalto LG, Quesenberry CP, Barnes D, et al. Midlife risk score for the prediction of dementia four decades later. Alzheimers Dement. 2014;10(5):562–570. doi:10.1016/j.jalz.2013.05.1772. PubMed PMID: 24035147.

[20] Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statist Med. 2004;23 (13):2109–2123.

[21] Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30(10):1105–1117.

[22] Liu X, Jin Z, Graziano JH. Comparing paired biomarkers in predicting quantitative health outcome subject to random censoring. Stat Methods Med Res. 2016;25(1):447–457.

[23] Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. Biometrika. 2005;92(4):965–970.

[24] Harrell FE, Jr., Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. Statist Med. 1984;3(2):143–152. PubMed PMID: 6463451.

[25] Saunders AM, Strittmatter WJ, Schmechel D, et al. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology. 1993;43(8):1467–1472. PubMed PMID: 8350998.

[26] Liu CC, Liu CC, Kanekiyo T, et al. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat Rev Neurol. 2013;9(2):106–118. doi: 10.1038/nrneurol.2012.263. PubMed PMID: 23296339; PubMed Central PMCID: PMCPMC3726719.

[27] Kivipelto M, Ngandu T, Laatikainen T, et al. Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study. Lancet Neurol. 2006;5(9):735–741. doi:10.1016/S1474-4422(06)70537-3. PubMed PMID: 16914401.

[28] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B. 1996;58(1):267–288.

[29] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–1429.

[30] Zhang Y, Shao Y. Concordance measure and discriminatory accuracy in transformation cure models. Biostatistics. 2017; doi:https://doi.org/10.1093/biostatistics/kxx016

[31] Han X, Zhang Y, Shao Y. On comparing two correlated C indices with censored survival data. Stat Med. 2017;provisionally accepted.