Biometrics 74, 734–743
June 2018

DOI: 10.1111/biom.12761

A C-Index for Recurrent Event Data: Application to Hospitalizations among Dialysis Patients

Sehee Kim,[1,*] **Douglas E. Schaubel**,[1] **and Keith P. McCullough**[2]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[2]Arbor Research Collaborative for Health, Ann Arbor, Michigan 48104, U.S.A.
[*]*email:* seheek@umich.edu

Summary. We propose a C-index (index of concordance) applicable to recurrent event data. The present work addresses the dearth of measures for quantifying a regression model's ability to discriminate with respect to recurrent event risk. The data which motivated the methods arise from the Dialysis Outcomes and Practice Patterns Study (DOPPS), a long-running prospective international study of end-stage renal disease patients on hemodialysis. We derive the theoretical properties of the measure under the proportional rates model (Lin et al., 2000), and propose computationally convenient inference procedures based on perturbed influence functions. The methods are shown through simulations to perform well in moderate samples. Analysis of hospitalizations among a cohort of DOPPS patients reveals substantial improvement in discrimination upon adding country indicators to a model already containing basic clinical and demographic covariates, and further improvement upon adding a relatively large set of comorbidity indicators.

Key words: C-index; Model discrimination; Proportional rates model; Recurrent events; Wild bootstrap

1. Introduction

In the analysis of clinical or epidemiologic data, the event of interest is often recurrent (i.e., can occur multiple times for the same subject). Examples of recurrent events include hospital admissions, infections, relapses, and blood transfusions. Methods of analysis of recurrent event data can be broadly classified as marginal or conditional, the distinction being that marginal methods implicitly average over the prior recurrent event history. Conditional models are distinguished by conditioning on the event history, either implicitly (e.g., through a frailty variate correlating the events within-subject) or explicitly through time-dependent covariates (e.g., event counters). Examples of marginal methods include Lawless and Nadeau (1995), Lin et al. (2000), and Schaubel et al. (2006), while examples of conditional methods include Andersen and Gill (1982). A comprehensive review of recurrent event models and methods is given by Cook and Lawless (2007).

The data which motivated our current work arise from the renown Dialysis Outcomes and Practice Patterns Study (DOPPS). The DOPPS is a prospective, multi-center, international study of patients receiving hemodialysis (the most common form of dialysis). Note that dialysis is the most common form of renal replacement therapy (RRT), which is necessary for patients with end-stage renal disease (ESRD), a condition characterized by kidney function that has diminished to such an extent that survival is considered impossible without RRT.

The DOPPS has been ongoing for more than 20 years, with data collected through five Phases. Details regarding the design of the DOPPS study have been described by Young et al. (2000). In Section 5, we analyze data from DOPPS Phase 5, which is the most recently completed phase. The recurrent event of interest is hospitalization, which is an important event due to its connection with morbidity and mortality, patient quality of life, health care cost and resource utilization. Since the DOPPS contains patients from many countries, we have a rather unique ability to directly evaluate differences among countries with respect to hospitalization rates. Correspondingly, we place some focus on comparing covariate-adjusted hospitalization rates by country. Of chief interest is evaluating the degree to which the fitted model accurately discriminates hospitalization risk among patients.

With respect to time-to-event outcomes, the majority of analyses have focused on patient survival. Furthermore, the limited number of DOPPS studies evaluating outcomes that can occur repeatedly within patient (e.g., hospitalization) have generally been restricted to time to first event. For example, in the study of hospitalizations, the event time would be time to first admission. Perhaps the biggest disadvantage of using time-to-first-event is inefficiency, in that considerable precision is sacrificed by ignoring each patient's second and subsequent events. That said, a benefit (or, at least a perceived benefit) of time-to-first-event is the ability to utilize techniques which are well-established for univariate survival data, but less (or not) developed for recurrent events. Among the more prominent techniques are those for model discrimination. The C-index (also known as the index of concordance) has become the most frequently used measure of the discriminatory ability of a survival model. However, no such measure has been developed for recurrent event data.

Considerable advancement has been made in the last 15 years with respect to the breadth of analyses available for recurrent event data. The majority of such works has focused on developing recurrent event methods for more complicated data structures. For example, Miloslavsky et al. (2004) developed recurrent event methods for dependently censored data. Several methods have been developed for jointly analyzing recurrent/terminal event data; for example, Ghosh and Lin (2002), Huang and Wang (2004), Liu et al. (2004), and Ye et al. (2007).

Despite the continuing advances in recurrent event methodology, there are relatively few methods available for evaluating a fitted model. The degree of fit is generally described in terms of predictive accuracy and/or discrimination ability, where the former considers how closely the fitted values approximate the observed responses. Discrimination considers the extent to which a model accurately distinguishes higher and lower risk subjects, and could be argued to be the more relevant of the two criteria in settings where prediction, per se, is not the analytic objective. A frequently used discrimination measure is the C-index. The C-index is related to the area under the receiver-operating curve (ROC), and was considered in the context of censored data by authors such as Harrell et al. (1982, 1984, 1996) and Uno et al. (2007). Several methods have been developed for ROC curves for survival data; for example, Heagerty et al. (2000), Moskowitz and Pepe (2004), Heagerty and Zheng (2005), and Uno et al. (2011).

In this report, we propose a C-index applicable to recurrent event data. Although initially motivated by a need to evaluate the discriminatory ability of the proportional rates model to the DOPPS data, the work addresses the lack of a widely accepted measure of model discrimination when the response is a recurrent event. The C-index can be interpreted as the proportion of subject-pairs for which the survival time ordering is concordant with the ordering of the fitted model's linear predictor. In the presence of censoring, the denominator of the C-index is the number of subject-pairs for which the order of the failure times is observed. Applying this concept to recurrent event responses, two subjects are comparable during follow-up time intervals during which both subjects are uncensored. For example, in the presence of right censoring and absence of any left truncation, two subjects are comparable until the minimum of the two censoring times; the subject pair then contributes to the denominator if the two event counters are not tied (at 0 or otherwise).

Using counting processes and U-processes theories, we derive the large-sample distribution of the proposed C-index estimator, and then propose a simulation-based method for computing standard errors and hence confidence intervals. Due to its popularity among practitioners, we derive the theoretical properties of the proposed C-index assuming the proportional rates model of Lin et al. (2000); extension of our results to other models, such as the additive rates model (Schaubel et al., 2006), would be straightforward.

The remainder of this report is organized as follows. In Section 2, we notationalize the data structure and set out the proposed measure. In Section 3, we derive the theoretical properties, proofs for which appear in the Appendix. Simulations are carried out in Section 4. The proposed methods are applied to the afore-described DOPPS data in Section 5. In Section 6, we provide some discussion.

2. Proposed Methods

2.1. *Set-Up and Notation*

Let $N^*(t)$ denote the number of events that occur over the interval $[0, t]$ and C denote the follow-up or censoring time. Assume that $N^*(\cdot)$ and C are independent conditional on a p-dimensional covariate vector $\mathbf{Z}$. The observed event process is denoted as $N(t) = N^*(t \wedge C)$ over the total observation window $[0, \tau]$, where $a \wedge b = \min(a, b)$. In the set-up of interest, we have a random sample of n individuals, with observed data $\{N_i(t), C_i, \mathbf{Z}_i;\ 0 \le t \le \tau\}$ $(i = 1, \ldots, n)$.

The proportional rates/means model (Lin et al., 2000) is commonly used to analyze recurrent event data. This model formulates the mean function for $N^*(\cdot)$ is associated with covariates $\mathbf{Z}$ as follows,

$$\mu_Z(t) \equiv E\{N^*(t)|\mathbf{Z}\} = \mu_0(t)\exp(\boldsymbol{\beta}_0{}'\mathbf{Z}), \tag{1}$$

where $\mu_0(\cdot)$ is an unknown baseline mean function of the marginal recurrent event process, and $\boldsymbol{\beta}_0$ is an unknown vector of regression parameters. The estimating equation for $\boldsymbol{\beta}_0$ is given by

$$\begin{aligned} U(\boldsymbol{\beta}) &= \sum_{i=1}^{n} U_i(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \int_0^{\tau} \left\{ \mathbf{Z}_i - \frac{\sum_k Y_k(t)\mathbf{Z}_k \exp(\boldsymbol{\beta}'\mathbf{Z}_k)}{\sum_k Y_k(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_k)} \right\} dN_i(t) = 0, \end{aligned} \tag{2}$$

where $Y_i(t) = I(C_i \ge t)$ and $I(\cdot)$ is the indicator function. The solution of (2) is $\hat{\boldsymbol{\beta}}$, and the Aalen–Breslow-type estimator for $\mu_0(t)$ is $\hat{\mu}(t)$.

2.2. *Proposed C-Index for Recurrent Events*

Consider future observations on two independent patients indexed $k = 1$ and $k = 2$ with observed data $\{\mathring{N}_k^*(\mathring{C}_k),\ \mathring{C}_k,\ \mathring{\mathbf{Z}}_k\}$ for $k = 1, 2$. A natural way to evaluate the risk discrimination ability of a recurrent event rate/mean model is to measure concordance between the observed and predicted event counts over the time interval of common observation; that is, $\hat{\mu}_{Z_1}(\mathring{C}_1 \wedge \mathring{C}_2)$ versus $\hat{\mu}_{Z_2}(\mathring{C}_1 \wedge \mathring{C}_2)$, given $\mathring{N}_1^*(\mathring{C}_1 \wedge \mathring{C}_2)$ and $\mathring{N}_2^*(\mathring{C}_1 \wedge \mathring{C}_2)$. We consider event rate models which are monotone functions of the linear predictor, $\boldsymbol{\beta}'\mathring{\mathbf{Z}}_k$. Without loss of generality, assuming that $\mu_{Z_k}(t)$ is monotone increasing in $\boldsymbol{\beta}'\mathring{\mathbf{Z}}_k$, we propose summarizing the model's risk discrimination through the following C-index,

$$\mathbb{C}(\boldsymbol{\beta}) = \Pr\{\boldsymbol{\beta}'\mathring{\mathbf{Z}}_1 > \boldsymbol{\beta}'\mathring{\mathbf{Z}}_2 | \mathring{N}_1^*(\mathring{C}_1 \wedge \mathring{C}_2) > \mathring{N}_2^*(\mathring{C}_1 \wedge \mathring{C}_2)\}. \tag{3}$$

Note that $\boldsymbol{\beta}$ may be derived from a score obtained from the existing literature, in which case $\boldsymbol{\beta}$ in (3) could be replaced by a constant vector $\boldsymbol{\beta}_0$ to reflect the fact that the parameter implied by the score is known with certainty. Conversely, the basis of the score may be an event rate model fitted to the data at hand, such that $\boldsymbol{\beta}$ from (3) would be replaced by an estimate $\hat{\boldsymbol{\beta}}$ to reflect its randomness. In the development that follows, we focus on the latter case, keeping in the background the various simplifications that arise when $\boldsymbol{\beta}_0$ is known.

We propose that (3) be estimated by the proportion of pairs in which the risk prediction scores and the observed event counts are concordant, as given by

$$\hat{\mathbb{C}}(\boldsymbol{\beta}) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I\{N_i(C_i \wedge C_j) > N_j(C_i \wedge C_j)\} I(\boldsymbol{\beta}^{'}\mathbf{Z}_i > \boldsymbol{\beta}^{'}\mathbf{Z}_j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} I\{N_i(C_i \wedge C_j) > N_j(C_i \wedge C_j)\}}. \tag{4}$$

It is interesting to compare $\hat{\mathbb{C}}(\boldsymbol{\beta})$ for recurrent event data to a concordance index for survival data where a subject can have at most one event (e.g., death or time-to-first recurrent event). Let T_i be the time-to-first event for the ith subject, with $X_i = T_i \wedge C_i$ as the observed survival time, and set $\Delta_i = I(T_i < C_i)$. For right-censored survival data, Harrell et al. (1996) proposed a concordance index, which can be rewritten using counting process notation as follows,

$$\begin{aligned}\hat{\mathbb{C}}_S(\boldsymbol{\beta}) &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i\, I(X_i < X_j,\ \boldsymbol{\beta}^{'}\mathbf{Z}_i > \boldsymbol{\beta}^{'}\mathbf{Z}_j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i\, I(X_i < X_j)} \\ &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I\{N_i(X_i \wedge X_j) > N_j(X_i \wedge X_j)\} I(\boldsymbol{\beta}^{'}\mathbf{Z}_i > \boldsymbol{\beta}^{'}\mathbf{Z}_j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} I\{N_i(X_i \wedge X_j) > N_j(X_i \wedge X_j)\}}.\end{aligned} \tag{5}$$

In survival data, a simple derivation can show that $\hat{\mathbb{C}}(\boldsymbol{\beta}) = \hat{\mathbb{C}}_S(\boldsymbol{\beta})$. However, when subjects can experience multiple (recurrent) events, restricting responses to "time-to-first event" can result in reduced discriminatory power (i.e., $\hat{\mathbb{C}}_S < \hat{\mathbb{C}}$). See Section 4.3 for an example.

We note that $\hat{\mathbb{C}}(\hat{\boldsymbol{\beta}})$ converges in probability to a censoring-dependent quantity

$$\mathbb{C}_0 = \Pr\{\boldsymbol{\beta}_0' \mathring{\mathbf{Z}}_1 > \boldsymbol{\beta}_0' \mathring{\mathbf{Z}}_2 | \mathring{N}_1^*(\mathring{C}_1) > \mathring{N}_2^*(\mathring{C}_1),\ \mathring{C}_1 \leq \mathring{C}_2\},$$

provided that $\hat{\boldsymbol{\beta}}$ converges to a constant vector $\boldsymbol{\beta}_0$ as n goes to infinity. It is true regardless whether the model (1) holds. For right-censored survival data, Gerds et al. (2013) showed that Harrell's estimator $\hat{\mathbb{C}}_S$, which does not explicitly model the censoring mechanism, performed as well as several existing Inverse Probability of Censoring Weighted C-index estimators. Correspondingly, our simulation study also showed a good performance of the proposed C-index even under violation of the assumption of conditionally independent censoring given the predictors. Therefore, in the interests of practicality and computational simplicity, we consider an unweighted version of the C-index.

The proposed C-index, $\hat{\mathbb{C}}(\hat{\boldsymbol{\beta}})$, discriminate event risk between subjects based on the estimated linear predictor, $\hat{\boldsymbol{\beta}}^{'}\mathbf{Z}$, from a regression model. However, as implied earlier, an external score determined independently of the data at hand could also be used. Examples include quantities such as the Gail model for breast cancer risk (Gail and Mai, 2010); the Model for End-Stage Liver Disease (MELD) score (Wiesner et al., 2003); and the Kidney Donor Risk Index (KDRI) for deceased-donor kidneys (Rao et al., 2009). From this perspective, the true value, $\mathbb{C}_0$, simply represents the limiting value of the $\hat{\mathbb{C}}$ with respect to a particular score, irrespective of whether the risk score is based on the true model (or any model). If an externally derived score is the basis of risk discrimination, a consistent variance estimator can be obtained based on the first term in equation (6).

2.3. *Variance Estimation*

In the Appendix, we show that $\mathcal{W} = \sqrt{n}\,\{\hat{\mathbb{C}}(\hat{\boldsymbol{\beta}}) - \mathbb{C}_0\}$ is asymptotically normal with mean zero and variance σ^2. To estimate σ^2, we use a resampling-based method. In particular, we modify the perturbation resampling method by Uno et al. (2011) to the recurrent event setting. Specifically, we first formulate W^*, a perturbed version of $\mathcal{W}$, then show that it has the same limiting distribution as $\mathcal{W}$. Then, σ^2 can be estimated as the sample variance on B realizations of W^*.

First, we construct a perturbed W^* that depends on two sources of random variation, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbb{C}}(\boldsymbol{\beta})$ for a fixed $\boldsymbol{\beta}$. Perturbation of both sources can be done by the same random quantity, resampled from any known distribution with mean 1 and variance 1. For instance, we use $\epsilon \sim$ Exponential (1). By (repeatedly) generating a random variable $\{\epsilon_i;\ i = 1, \ldots, n\}$, a perturbed $\hat{\boldsymbol{\beta}}$ can be obtained by

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \binom{n}{2}^{-1} \sum_{i<j} \hat{A}^{-1}(\hat{\boldsymbol{\beta}})\{U_i(\hat{\boldsymbol{\beta}}) + U_j(\hat{\boldsymbol{\beta}})\}\, \epsilon_i \epsilon_j / 2,$$

where $\hat{A}(\boldsymbol{\beta}) = -n^{-1} \partial U(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$. The $\hat{\boldsymbol{\beta}}$ perturbation is done through the estimating equation for $\boldsymbol{\beta}_0$ in equation (2). To generate a perturbed counterpart of $\hat{\mathbb{C}}(\boldsymbol{\beta})$, we define

$$V_{ij}(\boldsymbol{\beta}) = \frac{I\{N_i(C_i \wedge C_j) > N_j(C_i \wedge C_j)\}\{I(\boldsymbol{\beta}^{'}\mathbf{Z}_i > \boldsymbol{\beta}^{'}\mathbf{Z}_j) - \hat{\mathbb{C}}(\boldsymbol{\beta})\}}{n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} I\{N_i(C_i \wedge C_j) > N_j(C_i \wedge C_j)\}}.$$

Then, a perturbed random variable W^* can be generated by

$$W^* = \sqrt{n} \binom{n}{2}^{-1} \sum_{i<j} \{V_{ij}(\hat{\boldsymbol{\beta}}) + V_{ji}(\hat{\boldsymbol{\beta}})\}\, \epsilon_i \epsilon_j / 2 + \sqrt{n}\,\{\hat{\mathbb{C}}(\boldsymbol{\beta}^*) - \hat{\mathbb{C}}(\hat{\boldsymbol{\beta}})\}. \tag{6}$$

Finally, a two-sided 95% confidence interval can be obtained by $\hat{\mathbb{C}}(\hat{\boldsymbol{\beta}}) \pm 1.96\,\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}$ is the sample standard deviation of W^*, following the asymptotic normality of W^* shown in the Appendix.

We note that the second term of (6) needs to assume that the risk prediction model (at least the regression components) is correctly specified. This assumption can be avoided by employing the nonparametric bootstrap. This gain in robustness is at the expense of increased computation time. For instance, in our analysis of the DOPPS data (Section 4), the model-based standard error estimates using (6) were very similar to the nonparametric bootstrap estimates (see Tables 5–6), but took one third of the computing time required for the nonparametric bootstrap.

3. Simulation

We evaluated the finite sample properties of the proposed estimator in (4) through a series of simulation studies under scenarios which differed by the intensity of recurrent event and the censoring mechanism. The recurrent event times were generated from a proportional intensity model:

$$\Lambda_{Z_i}(t) = \Lambda_0(t)\gamma_i \exp(\beta_1 Z_{1i} + \beta_2 Z_{2i}), \tag{7}$$

where $Z_{1i} \sim$ Uniform$[-1, 1]$, $Z_{2i} \sim$ Bernoulli(0.5), and true $(\beta_1, \beta_2)' = (1, 0.5)$. To vary the intensity of recurrent event, we included a subject-specific random effect γ_i, following a Gamma distribution with mean 1 and variance $V = 0.01$, 0.25, 0.5, or 1. As the V value increased, an individual tended to experience more events, but a percentage of individuals with zero event also increased. Hence, a total number of observed events was similar across different V values. To investigate sensitivity to the assumption on the censoring mechanism, the censoring time C_i was generated from the following four scenarios:

(i) Complete follow-up: $C_i = \tau$ for all i; set $\tau = 5$,
(ii) Completely independent censoring,
(iii) Covariate-dependent censoring,
(iv) Outcome-dependent censoring.

In Scenario (ii), C_i was generated as the minimum of Uniform $[1, \tau + 2]$ and τ. In Scenario (iii), C_i was generated from Uniform$[1, \tau + 1]$ if $Z_{2i} = 0$ and Exponential(1) + 1 if $Z_{2i} = 1$, and then truncated at τ. In Scenario (iv), we set $C_i = \min(\exp(\gamma_i), 100\tau)$, which yielded Spearman correlations between $N_i(C_i)/C_i$ and C_i, ranging 0.05–0.56. In Table 1, we fixed the baseline intensity to $\Lambda_0(t) = 0.5\,t$ for all scenarios, in order to demonstrate the dependence between the censoring distribution and C-index, with the baseline intensity equal. In contrast, in Table 2 $\Lambda_0(t) = \nu t$ differed across scenarios in order to demonstrate that similar C_0 can result from different censoring distributions. Specifically, ν was tuned to yield similar numbers of comparable pairs among the different censoring scenarios; we set $\nu = 0.5$ for Scenario (i), $\nu = 1$ for Scenario (ii), and $\nu = 1.4$ for Scenario (iii). For each setting, the true value of $\mathbb{C}_0$ was approximated based on a random sample of $\{\mathbf{Z}_i, N_i^*(\cdot), C_i\}$ from one hundred thousand individuals under true $\boldsymbol{\beta}_0$. To compute $\hat{\mathbb{C}}$, we first fitted a proportional rates model, then used the resulting estimates $\hat{\boldsymbol{\beta}}$ to compute the proposed C-index estimate from (4). Standard errors were computed under the proposed perturbation resampling methods.

Simulation results based on 1000 replications and $n = 200$ are presented in Tables 1–3. In each table, Bias equals the difference between the average of the C-index estimates and the true value; SD is the empirical standard deviation of the parameter estimates; SEE is the average of the standard error estimates; and CP is the coverage probability of the 95% confidence intervals.

From Tables 1–2, we notice that $\mathbb{C}_0$ strictly decreased as: the baseline intensity (ν) decreased; the frailty variance (V) increased; and the marginal proportion of subjects with zero observed events ($P\{N_i(\tau) = 0\}$) increased. However, $\mathbb{C}_0$ appeared only to be weakly dependent on the percentage of censored (unobserved) $N^*(\tau)$ and the specific form of censoring mechanism. In all settings we considered, including the outcome-dependent censoring scenario, the proposed estimator showed good performance. That is, the $\hat{\mathbb{C}}$ estimates were unbiased, while the standard error estimates closely approximated the true variability in $\hat{\mathbb{C}}$. In turn, CPs were close to the nominal level.

We also investigated the finite sample properties of the proposed $\hat{\mathbb{C}}$ when the proportional rates model assumption was violated. Under this scenario, the additive rates model $E[dN_i^*(t)|Z_i, \gamma_i] = \{m_0 + \gamma_i + \beta_1 Z_{1i} + \beta_2 Z_{2i}\}dt$ with $m_0 = 0.2$ (Schaubel et al., 2006) was used to generate recurrent event times, with the proportional rates model used to develop a risk score and to calculate the C-index. Table 3 shows that, even if the risk score development model was different from the data generation model, $\hat{\mathbb{C}}$ accurately estimated $\mathbb{C}_0$ with a small sample size, where $\mathbb{C}_0$ is defined with respect to the assumed risk score development model. The standard deviations of $\hat{\mathbb{C}}$ were well estimated using the nonparametric bootstrap method, whereas the model-based standard error estimates using equation (6) were slightly over-estimated when the frailty variance increased to $V = 1$.

4. Application

4.1. *Dialysis Outcomes and Practice Patterns Study (DOPPS)*

The DOPPS is a prospective multi-center international study of prevalent hemodialysis patients. Patients within each DOPPS facility were randomly sampled, with the intention of preserving the key characteristics of the base population of the selected facilities. Data were obtained from Arbor Research Collaborative for Health, which founded and serves a data coordinating center for the DOPPS. We analyzed data from DOPPS Phase 5, the most recently completed phase of the study. Patients were recruited for Phase 5 between 2012 and 2015. The total sample size included approximately 17,000 prevalent hemodialysis patients, from 465 facilities in 19 different countries. Active follow-up began at entry to DOPPS and continued until the earliest of death, receipt of a kidney transplant, switch to peritoneal dialysis, transfer to another facility, or the end of the observation period (12/31/2015). Further detail regarding the DOPPS is available in Robinson et al. (2012).

In the interests of constructing a cohort of (approximately) incident end-stage renal disease patients, we included only patients with ≤ 6 months on dialysis at DOPPS entry ($n = 3692$). Our study cohort included patients from the following countries: Belgium, Canada, China, Germany, the six Gulf Cooperation Council (GCC) countries (Bahrain, Qatar, Kuwait, Oman, Saudi Arabia, and United Arab Emirates), Italy, Japan, Spain, Sweden, the United Kingdom (UK), and the United States (U.S.).

4.2. *Objectives of Analysis*

The recurrent event of interest is hospitalization (i.e., hospital admission). Among the more than 140 peer-reviewed articles featuring the analysis of DOPPS data, relatively few have involved evaluating the ability of the assumed regression

Table 1

Simulation results for the proposed $\hat{\mathbb{C}}$ based on $n = 200$. The baseline intensity function was set as $\Lambda_0(t) = 0.5\,t$ regardless of censoring scenarios.

V[a]	Censored[b]	$P\{N_i(\tau) = 0\}$[c]	True	Bias	SD	SEE	CP
Complete follow-up							
0.01	0%	8.9%	0.804	0.002	0.016	0.017	94.7
0.25	0%	12.9%	0.745	0.001	0.020	0.020	95.5
0.5	0%	17.5%	0.707	0.002	0.022	0.022	94.9
1.0	0%	25.4%	0.667	−0.002	0.024	0.025	96.3
Completely independent censoring							
0.01	24.3%	17.8%	0.770	0.001	0.020	0.020	94.2
0.25	23.1%	21.9%	0.729	0.001	0.022	0.022	94.0
0.5	21.9%	26.2%	0.702	0.000	0.024	0.024	94.1
1.0	19.9%	33.6%	0.666	0.001	0.026	0.027	95.1
Covariate-dependent censoring							
0.01	42.6%	25.8%	0.749	0.001	0.022	0.022	94.8
0.25	40.9%	30.0%	0.717	0.003	0.024	0.025	93.9
0.5	38.7%	34.0%	0.696	0.002	0.025	0.026	94.7
1.0	35.2%	40.7%	0.667	0.001	0.028	0.028	94.6
Outcome-dependent censoring							
0.01	41.3%	21.8%	0.765	0.003	0.020	0.021	95.4
0.25	35.9%	29.4%	0.729	0.001	0.023	0.024	95.6
0.5	33.6%	35.2%	0.707	0.002	0.025	0.025	94.6
1.0	29.8%	43.0%	0.681	0.001	0.027	0.028	95.1

[a] The variance of frailty γ_i used for generating repeated events from model (7).
[b] $1 - E[N_i(\tau)]/E[N_i^*(\tau)]$.
[c] The marginal proportion of subjects with zero observed events.

Table 2

Simulation results for the proposed $\hat{\mathbb{C}}$ based on $n = 200$. The baseline intensity function $\Lambda_0(t) = \nu t$ was varying with different censoring scenarios.

ν	V[a]	Censored[b]	$P\{N_i(\tau) = 0\}$[c]	True	Bias	SD	SEE	CP
Complete follow-up								
0.5	0.01	0%	8.9%	0.804	0.002	0.016	0.017	94.7
	0.25	0%	12.9%	0.745	0.001	0.020	0.020	95.5
	0.5	0%	17.5%	0.707	0.002	0.022	0.022	94.9
	1.0	0%	25.4%	0.667	−0.002	0.024	0.025	96.3
Completely independent censoring								
1.0	0.01	26.2%	6.7%	0.809	0.000	0.015	0.015	94.3
	0.25	25.6%	9.9%	0.746	0.001	0.019	0.019	95.0
	0.5	24.6%	13.9%	0.712	−0.002	0.021	0.022	95.3
	1.0	22.6%	21.2%	0.664	0.000	0.023	0.024	95.2
Covariate-dependent censoring								
1.4	0.01	46.0%	6.2%	0.805	−0.001	0.015	0.016	95.9
	0.25	45.3%	9.6%	0.744	0.001	0.019	0.020	94.2
	0.5	44.2%	13.9%	0.708	0.001	0.021	0.022	95.3
	1.0	41.2%	21.2%	0.664	0.001	0.023	0.025	95.5

[a] The variance of frailty γ_i used for generating repeated events from model (7).
[b] $1 - E[N_i(\tau)]/E[N_i^*(\tau)]$.
[c] The marginal proportion of subjects with zero observed events.

Table 3

Simulation results for the proposed $\hat{\mathbb{C}}$ based on $n = 200$ under the misspecified risk prediction model. The baseline intensity function was set as $\Lambda_0(t) = 0.2\,t$ regardless of censoring scenarios.

						Model-based		Robust	
V[a]	Censored[b]	$P\{N_i(\tau) = 0\}$[c]	True	Bias	SD	SEE	CP	SEE	CP
Complete follow-up									
0.01	0%	0.0%	0.691	0.002	0.021	0.022	95.0	0.021	93.7
0.25	0%	0.2%	0.659	0.003	0.023	0.024	94.9	0.023	94.7
0.5	0%	0.5%	0.645	0.003	0.024	0.025	95.1	0.023	92.5
1.0	0%	1.0%	0.630	0.004	0.024	0.028	96.4	0.024	93.4
Completely independent censoring									
0.01	26.6%	1.5%	0.656	0.003	0.022	0.023	94.5	0.022	93.1
0.25	26.4%	2.1%	0.637	0.002	0.024	0.024	94.5	0.023	94.8
0.5	26.7%	2.6%	0.627	0.003	0.024	0.026	94.6	0.024	94.9
1.0	26.6%	3.4%	0.618	0.004	0.024	0.029	97.1	0.025	94.1
Covariate-dependent censoring									
0.01	46.1%	3.2%	0.633	0.003	0.024	0.025	95.8	0.024	94.3
0.25	46.2%	3.9%	0.622	0.002	0.025	0.026	95.4	0.024	95.2
0.5	46.1%	4.8%	0.615	0.004	0.024	0.027	97.4	0.025	94.8
1.0	45.9%	6.1%	0.610	0.004	0.024	0.030	96.8	0.026	94.1

[a] The variance of frailty γ_i used for generating repeated events from model (7).

[b] $1 - E[N_i(\tau)]/E[N_i^*(\tau)]$.

[c] The marginal proportion of subjects with zero observed events.

Table 4

Analysis of DOPPS data: Estimated covariate effects on hospitalization rates (based on Model 3)

Covariate	$\hat{\boldsymbol{\beta}}$	SE	p-value	$\exp\{\hat{\boldsymbol{\beta}}\}$
Age (per 15 yrs)	−0.024	0.030	0.425	0.98
Female	0.100	0.065	0.127	1.11
Height (per 10 cm)	0.009	0.036	0.804	1.01
Dialysis $\leq$ 3 months[a]	0.027	0.056	0.626	1.03
Graft (ref: AVF)	0.510	0.111	<0.0001	1.67
Catheter (ref: AVF)	0.454	0.058	<0.0001	1.57
Congestive heart failure	0.165	0.061	0.007	1.18
Coronary artery disease	0.122	0.058	0.035	1.13
Cerebrovascular disease	0.142	0.060	0.019	1.15
Peripheral vascular disease	0.163	0.059	0.006	1.18
Chronic obstructive pulmonary disease	0.221	0.071	0.002	1.25
Diabetes	−0.001	0.052	0.984	1.00
Cancer	0.189	0.073	0.010	1.21
Neurological disease	0.292	0.076	0.0001	1.34
Belgium (ref: U.S.)	0.381	0.097	0.0001	1.46
Canada (ref: U.S.)	−0.227	0.108	0.036	0.77
China (ref: U.S.)	−0.386	0.191	0.043	0.68
Germany (ref: U.S.)	0.336	0.082	<0.0001	1.40
Gulf (ref: U.S.)	−0.181	0.115	0.115	0.83
Italy (ref: U.S.)	−0.329	0.115	0.004	0.72
Japan (ref: U.S.)	0.001	0.097	0.996	1.00
Spain (ref: U.S.)	−0.368	0.113	0.001	0.69
Sweden (ref: U.S.)	0.750	0.113	<0.0001	2.12
United Kingdom (ref: U.S.)	0.134	0.111	0.226	1.14

AVF, arteriovenous fistula

[a] Time since initiating dialysis as of DOPPS entry; ref: $[3, 6]$ months.

Table 5
Analysis of DOPPS data: Comparison of $\hat{\mathbb{C}}$ for various models.

Model	$\hat{\mathbb{C}}$	Model-based SE	Robust SE
1[a]	0.596	0.008	0.007
2[b]	0.630	0.008	0.008
3[c]	0.654	0.008	0.008

Model comparison	$\hat{\mathbb{C}}_\Delta$	Robust SE
2 versus 1	0.034	0.006
3 versus 2	0.025	0.005

[a] Model 1 included age, sex, height, duration of dialysis, vascular access (AV fistula, AV graft, or tunneled catheter).

[b] Model 2 included all predictors in Model 1 + country (Belgium, Canada, China, Germany, Gulf, Italy, Japan, Spain, Sweden, UK, or U.S.).

[c] Model 3 included all predictors in Model 2 + eight comorbid conditions.

model to discriminate patients with respect to event risk. Motivated by this issue, we sought to quantify the improvement in discrimination resulting from adjusting for country, and further adjusting for an extensive list of comorbidity indicators. That is, we evaluate the improvement in the model based on the successive inclusion of covariates representing key distinguishing characteristics of DOPPS: its international component, and collection of information on an extensive list of comorbid conditions.

4.3. *Analysis of DOPPS Data*

Of the $n = 3692$ incident patients identified, 49% and 46% were dialyzing with fistulas and catheters, respectively, at the study entry. A mean follow-up time was 14 months, and maximum follow-up was 4 years. In terms of hospitalizations, 55% of patients were never admitted, 20% were admitted once, 11% were admitted twice, and 14% were hospitalized >2 times.

Three different sets of potential confounding factors were considered for the risk of hospitalization. Model 1 was the most basic model, and included age (15-year increments), sex (ref: male), height (10-cm increments), duration of dialysis at DOPPS entry (ref: 3–6 month), and two separate indicators for a graft and a catheter user (ref: fistula). In addition to the afore-listed covariates, Model 2 further adjusted for country, while Model 3 adjusted for country and the following eight comorbid conditions: congestive heart failure, coronary artery disease, cerebrovascular disease, peripheral vascular disease, chronic obstructive pulmonary disease, diabetes, cancer, and neurological disorder. In Table 4, the estimated regression coefficients for predictors in Model 3 are presented. The hospitalization rate was significantly increased for patients with a graft or a catheter serving as vascular access, relative to arteriovenous fistula (AVF). Each of the comorbid conditions, except diabetes, was associated with a significantly increased hospitalization rate. Canada, China, Italy, and Spain had significantly lower hospitalization rates related to the United States. Three countries had significantly higher hospitalization rates than the U.S.: Belgium, Germany, and Sweden, with the latter estimated to have the highest rate.

In Table 5, we compare C-index estimates for models with and without comorbidity and country. The standard error estimate of $\hat{\mathbb{C}}$ using the perturbation resampling method (see Model-based SE) was compared with the nonparametric bootstrap estimate (see Robust SE). Risk discrimination based on different models was compared through $\hat{\mathbb{C}}_\Delta$, the difference in $\hat{\mathbb{C}}$. To estimate the variance of $\hat{\mathbb{C}}_\Delta$, we recommend using the nonparametric bootstrap method since its validity does not require correctness of either of the models being compared. Suppose $\hat{\mathbb{C}}_\Delta^{(b)}$ is the difference in C-index estimates obtained from the bth bootstrapped set of random subjects. Then, the the robust SE of $\hat{\mathbb{C}}_\Delta$ was obtained as the sample standard deviation of $B = 100$ realizations of $\hat{\mathbb{C}}_\Delta^{(b)}$. A two-sided 95% confidence interval $\hat{\mathbb{C}}_\Delta \pm 1.96 * \mathrm{SE}(\hat{\mathbb{C}}_\Delta)$ can be used for the test for no difference in $\hat{\mathbb{C}}$s.

Table 5 shows that Model 1, having not adjusted for country and comorbidities, resulted in $\hat{\mathbb{C}} = 0.596$. This implies that 59.6% of pairs were concordant, in the sense that patient predicted to have higher hospitalization risk had more hospital admissions during the sub-interval of overlapping follow-up.

Table 6
Analysis of DOPPS data by country using Model 1 + comorbidities

Country	Sample size[a]	Hospitalization rate[b]	$\hat{\mathbb{C}}$	Model-based SE	Robust SE
Canada	315	6.7	0.644	0.030	0.030
Germany	359	10.8	0.654	0.023	0.023
Japan	631	5.0	0.645	0.022	0.021
U.S.	1047	7.7	0.619	0.018	0.015
All	3692	7.3	0.654	0.008	0.008

SE, standard error estimate.

[a] Patients.

[b] Per 100 patient-months.

By additionally including country in the model, the discriminatory power of Model 2 improved to 63% concordance, and the improvement was statistically significant at the level of 0.05 ($\hat{\mathbb{C}}_\Delta = 3.4\%$, SE = 0.6%). Finally, $\hat{\mathbb{C}}$ of Model 3 improved to 65.4% concordance by further adjusting for comorbidities; and Model 3 was significantly better than Model 2 ($\hat{\mathbb{C}}_\Delta = 2.5\%$, SE = 0.5%).

For comparison purposes, we also carried out a time-to-first event analysis, with the end-point re-defined as the time to first hospital admission. We observed the $\hat{\mathbb{C}}_S$ decreased from $\hat{\mathbb{C}}$ when the time-to-first-hospitalization was used only ($\hat{\mathbb{C}}_S = 0.585$, SE = 0.008 for Model 1; $\hat{\mathbb{C}}_S = 0.618$, SE = 0.008 for Model 2; $\hat{\mathbb{C}}_S = 0.641$, SE = 0.008 for Model 3). This can be explained by the fact that the use of recurrent event data allowed a longer common observation period for risk comparisons (i.e., $C_i \wedge C_j$ was longer than $X_i \wedge X_j$), which, in turn, increased the corresponding model's discriminatory power.

For those countries with more than 300 patients (Canada, Germany, Japan and U.S.), we have also carried out separate analyses by county and evaluated the country-specific models using the proposed C-index. The country-specific models included the same set of predictors (i.e., Model 1 + comorbidities), but the resulting C-index estimates were varying country to country (Table 6). In particular, the model for Germany obtained the highest $\hat{\mathbb{C}} = 0.654$ (best in predicting a higher risk), whereas the model for U.S. yielded a noticeably lower $\hat{\mathbb{C}} = 0.619$, comparing to $\hat{\mathbb{C}} = 0.644$ for Canada and $\hat{\mathbb{C}} = 0.645$ for Japan. The standard error estimate for $\hat{\mathbb{C}}$ consistently reduced as the sample size increased.

5. Discussion

Using counting processes, we have developed a C-index applicable to recurrent event data. Theoretical properties are derived under an assumed proportional rates model (Lin et al., 2000). The proposed C-index can be interpreted as the fraction of concordant subject-pairs, where concordance refers to the within-pair ordering of the linear predictor and the observed number of events for the follow-up subinterval during which both subject are uncensored. The measure reflects a rate model's ability to discriminate subjects with respect to recurrent event risk. The proposed C-index performed well in simulation studies. The use of perturbation methods (in lieu of traditional closed-form variance computation) permits relatively quick estimation of confidence intervals, and hence makes the proposed inference procedures quite attractive computationally; this is an important property in the big data era.

In our analysis of hospitalization rates using data from Phase 5 of the Dialysis Outcomes and Practice Patterns Study (DOPPS), $\hat{\mathbb{C}}$ increased by approximately 0.03 upon the addition of country to a model which contained demographic and basic clinical covariates, and then increased by an additional ≈0.02 upon further adjustment for comorbidity indicators. These increases are somewhat contrary to the reputation for insensitivity the C-index has earned in the context of standard survival data. Further study would reveal whether this is due to the nature of recurrent event data in general, or whether our real-data example is somewhat of an anomaly. It is true that the C-index tends to be higher for logistic regression models than for survival models, owing mostly to the latter being subject to censoring; for example, see Sharma et al. (2016). For right censored survival data, for a subject-pair to be usable, at least one member of the pair has to be an observed death. In contrast, for right censored recurrent event data, all untied subject-pairs are usable, albeit during the subinterval of overlapping follow-up (i.e., until the minimum of the two censoring times within-pair). In the DOPPS analysis, we actually observed that C-index with recurrent event data was higher than that with survival data. From this perspective, recurrent event data may lie in between survival data and binary responses with respect to the typical sensitivity of the C-index to the addition of model covariates.

6. Supplementary Materials

The source R codes for implementing the proposed methods are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors thank the Associate Editor and Reviewer for the constructive suggestions that led to substantial improvement of the manuscript. This work was supported in part by National Institutes of Health grant R01-DK070869. The DOPPS is administered by Arbor Research Collaborative for Health and supported by scientific research grants from Amgen (since 1996), Kyowa Hakko Kirin (since 1999, in Japan), Sanofi Renal (since 2009), Abbott (since 2009), Baxter (since 2011), and Vifor Fresenius Renal Pharma (since 2011), without restrictions on publications. The authors wish to thank Arbor Research Collaborative for Health for providing access to the DOPPS data.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10**, 1100–1120.

Cook, R. J. and Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. New Your, NY: Springer Science & Business Media.

Gail, M. H. and Mai, P. L. (2010). Comparing breast cancer risk assessment models. *Journal of the National Cancer Institute* **102**, 665–668.

Gerds, T. A., Kattan, M. W., Schumacher, M., and Yu, C. (2013). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine* **32**, 2173–2184.

Ghosh, D. and Lin, D. (2002). Marginal regression models for recurrent and terminal events. *Statistica Sinica* **12**, 663–688.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association* **247**, 2543–2546.

Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., and Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine* **3**, 143–152.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361–387.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics* **61**, 92–105.

Huang, C.-Y. and Wang, M.-C. (2004). Joint modeling and estimation for recurrent event processes and failure time data. *Journal of the American Statistical Association* **99**, 1153–1165.

Lawless, J. F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**, 158–168.

Lin, D., Wei, L., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 711–730.

Liu, L., Wolfe, R. A., and Huang, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics* **60**, 747–756.

Miloslavsky, M., Keleş, S., van der Laan, M. J., and Butler, S. (2004). Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 239–257.

Moskowitz, C. S. and Pepe, M. S. (2004). Quantifying and comparing the accuracy of binary biomarkers when predicting a failure time outcome. *Statistics in medicine* **23**, 1555–1570.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *The Annals of Statistics* **15**, 780–799.

Nolan, D. and Pollard, D. (1988). Functional limit theorems for U-processes. *The Annals of Probability* **16**, 1291–1298.

Rao, P. S., Schaubel, D. E., Guidinger, M. K., Andreoni, K. A., Wolfe, R. A., Merion, R. M., Port, F. K., and Sung, R. S. (2009). A comprehensive risk quantification score for deceased donor kidneys: the kidney donor risk index. *Transplantation* **88**, 231–236.

Robinson, B., Bieber, B., Pisoni, R., and Port, F. (2012). Dialysis outcomes and practice patterns study (DOPPS): Its strengths, limitations, and role in informing practices and policies. *Clinical Journal of the American Society of Nephrology* **7**, 1897–1905.

Schaubel, D. E., Zeng, D., and Cai, J. (2006). A semiparametric additive rates model for recurrent event data. *Lifetime Data Analysis* **12**, 389–406.

Sharma, P., Shu, X., Schaubel, D. E., Sung, R. S., and Magee, J. C. (2016). Propensity score-based survival benefit of simultaneous liver-kidney transplant over liver transplant alone for recipients with pretransplant renal dysfunction. *Liver Transplantation* **22**, 71–79.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30**, 1105–1117.

Uno, H., Cai, T., Tian, L., and Wei, L. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102**, 527–537.

Wiesner, R., Edwards, E., Freeman, R., Harper, A., Kim, R., and Kamath, P., et al. (2003). Model for end-stage liver disease (MELD) and allocation of donor livers. *Gastroenterology* **124**, 91–96.

Ye, Y., Kalbfleisch, J. D., and Schaubel, D. E. (2007). Semiparametric analysis of correlated recurrent and terminal events. *Biometrics* **63**, 78–87.

Young, E. W., Goodkin, D. A., Mapes, D. L., Port, F. K., Keen, M. L., Chen, K., Maroni, B. L., Wolfe, R. A., and Held, P. J. (2000). The Dialysis Outcomes and Practice Patterns Study (DOPPS): An international hemodialysis study. *Kidney International* **57**, 74–81.

Received September 2016. *Revised June* 2017.
Accepted June 2017.

Appendix

Asymptotic properties of $\hat{\mathbb{C}}(\hat{\boldsymbol\beta})$. To establish asymptotic properties of $\hat{\mathbb{C}}(\hat{\boldsymbol\beta})$, we impose the same regularity conditions as those in Section 2 of Lin et al. (2000). We begin with the consistency of $\hat{\mathbb{C}}(\hat{\boldsymbol\beta})$. We define the denominator of $\hat{\mathbb{C}}(\boldsymbol\beta)$ by

$$\hat\pi = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} I\{N_i(C_i \wedge C_j) > N_j(C_i \wedge C_j)\} = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{ij}.$$

Given that $\hat\pi$ (as well as $\hat{\mathbb{C}}(\boldsymbol\beta)$) is a functional of a U-process indexed by a class of indicator functions, by a uniform law of large numbers for U-processes (Nolan and Pollard, 1987) and the independent assumption between N^* and C, we can show $\hat\pi$ converges to $\Pr\{N_1^*(C_1) > N_2^*(C_1),\, C_1 \le C_2\}$ in probability. Following the strong consistency of $\hat{\boldsymbol\beta}$ (Lin et al., 2000) and a uniform law of large numbers for U-processes of $\hat{\mathbb{C}}(\boldsymbol\beta)$, we can then show that

$$\hat{\mathbb{C}}(\hat{\boldsymbol\beta}) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{ij} I(\hat{\boldsymbol\beta}^{'}\mathbf{Z}_i > \hat{\boldsymbol\beta}^{'}\mathbf{Z}_j)/\left\{n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{ij}\right\}$$

converges to $\mathbb{C}_0 = \Pr\{\boldsymbol\beta_0'\mathbf{Z}_1 > \boldsymbol\beta_0'\mathbf{Z}_2 \mid N_1^*(C_1) > N_2^*(C_1),\, C_1 \le C_2\}$ in probability.

To show the limiting distribution of $\mathcal{W}$, we decompose $\mathcal{W}$ into

$$\mathcal{W} = \sqrt{n}\,\{\hat{\mathbb{C}}(\boldsymbol\beta_0) - \mathbb{C}_0\} + \sqrt{n}\,\{\hat{\mathbb{C}}(\hat{\boldsymbol\beta}) - \hat{\mathbb{C}}(\boldsymbol\beta_0)\}. \qquad (8)$$

By a functional central limit theorem for U-processes (Nolan and Pollard, 1988), the first term in (8) is asymptotically equivalent to, for a fixed $\boldsymbol\beta_0$,

$$\begin{aligned}\sqrt{n}\{\hat{\mathbb{C}}(\boldsymbol\beta_0) - \mathbb{C}_0\} &= n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{ij}\{I(\boldsymbol\beta_0^{'}\mathbf{Z}_i > \boldsymbol\beta_0^{'}\mathbf{Z}_j)\\ &\quad -\mathbb{C}_0\}/\left\{n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{ij}\right\}\\ &\approx \sqrt{n}\binom{n}{2}^{-1}\sum_{i<j}\{\mathcal{V}_{ij}(\boldsymbol\beta_0) + \mathcal{V}_{ji}(\boldsymbol\beta_0)\}/2,\end{aligned}$$

where $\mathcal{V}_{ij}(\boldsymbol\beta_0) = I_{ij}\{I(\boldsymbol\beta_0^{'}\mathbf{Z}_i > \boldsymbol\beta_0^{'}\mathbf{Z}_j) - \mathbb{C}_0\}[\Pr\{N_1^*(C_1) > N_2^*(C_1),\, C_1 \le C_2\}]^{-1}$. Next, we show that the second term in (8) is asymptotically equivalent to

$$\sqrt{n}\,\{\hat{\mathbb{C}}(\hat{\boldsymbol{\beta}}) - \hat{\mathbb{C}}(\boldsymbol{\beta}_0)\} = \partial_\beta \mathbb{C}_0\,\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(1),$$

where $\partial_\beta \mathbb{C}_0$ is an approximation of the first derivative of $\mathbb{C}_0$ with respect to $\boldsymbol{\beta}_0$. The proof starts from the fact that for an indicator function, we can find at least one continuously differentiable approximation function. For simplicity, let's consider an indicator function $I(\beta) = 1$ if $\beta > 0$, and $I(\beta) = 0$ otherwise. Then, there exists a continuously differentiable function $L(\beta, \nu)$ such that, $\forall \beta \in \mathbb{R}$, it holds that $\lim_{\nu \to \infty} L(\beta, \nu) = I(\beta)$, where

$$L(\beta, \nu) = \frac{1}{1 + \exp\{-\nu(\beta + 1/\sqrt{\nu})\}}.$$

The first derivative of $L(\beta, \nu)$ with respect toβ is $\partial L(\beta, \nu)/\partial\beta = \nu \exp\{-\nu(\beta + 1/\sqrt{\nu})\}\ [1 + \exp\{-\nu(\beta + 1/\sqrt{\nu})\}]^{-2}$. As ν goes to positive infinity, the limit of $\partial L(\beta, \nu)/\partial\beta$ approaches 0 for any fixed β. Limits of higher order derivatives are bounded as well. By analogical arguments, approximations of the first derivatives of $\mathbb{C}_0$ and $\hat{\mathbb{C}}(\boldsymbol{\beta}_0)$ exist, and they are denoted as $\partial_\beta \mathbb{C}_0$ and $\partial_\beta \hat{\mathbb{C}}(\boldsymbol{\beta}_0)$, respectively. Now, applying the Taylor series expansion at $\boldsymbol{\beta}_0$ from the consistency of $\hat{\boldsymbol{\beta}}$ yields

$$\sqrt{n}\,\{\hat{\mathbb{C}}(\hat{\boldsymbol{\beta}}) - \hat{\mathbb{C}}(\boldsymbol{\beta}_0)\} \approx \partial_\beta \hat{\mathbb{C}}(\boldsymbol{\beta}_0)\,\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(1).$$

We then show $\partial_\beta \hat{\mathbb{C}}(\boldsymbol{\beta}_0)$ converges to $\partial_\beta \mathbb{C}_0$ in probability from the consistency of $\hat{\mathbb{C}}(\boldsymbol{\beta})$ and the continuity of $\mathbb{C}_0$. Note that the asymptotic expansion of $\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ shown by Lin et al. (2000) can be re-written with respect to a U-statistic as follows:

$$\begin{aligned}\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &\approx n^{-1/2} \sum_{i=1}^{n} A^{-1}(\boldsymbol{\beta}_0)\, U_i(\boldsymbol{\beta}_0)\\ &= \sqrt{n} \binom{n}{2}^{-1} \sum_{i<j} A^{-1}(\boldsymbol{\beta}_0)\{U_i(\boldsymbol{\beta}_0) + U_j(\boldsymbol{\beta}_0)\}/2.\end{aligned}$$

Finally, it then follows, by a functional central limit theorem for U-processes, that

$$\mathcal{W} = \sqrt{n} \binom{n}{2}^{-1} \sum_{i<j} \tilde{W}_{ij} + o_p(1) \tag{9}$$

converges in distribution to a Gaussian process with zero mean and variance $\sigma^2 \equiv E(\tilde{W}_{12}\tilde{W}_{13})$, where $\tilde{W}_{ij} = \{\mathcal{V}_{ij}(\boldsymbol{\beta}_0) + \mathcal{V}_{ji}(\boldsymbol{\beta}_0)\}/2 + \partial_\beta \mathbb{C}_0\, A^{-1}(\boldsymbol{\beta}_0)\{U_i(\boldsymbol{\beta}_0) + U_j(\boldsymbol{\beta}_0)\}/2$.

To approximate the distribution of $\mathcal{W}$, we simulate a number of realizations from W^*, given by

$$W^* = \sqrt{n} \binom{n}{2}^{-1} \sum_{i<j} \{V_{ij}(\hat{\boldsymbol{\beta}}) + V_{ji}(\hat{\boldsymbol{\beta}})\}\, \epsilon_i \epsilon_j/2 + \sqrt{n}\,\{\hat{\mathbb{C}}(\boldsymbol{\beta}^*) - \hat{\mathbb{C}}(\hat{\boldsymbol{\beta}})\},$$

by repeatedly sampling $\{\epsilon_i;\ i = 1, \ldots, n\}$, conditioning on the observed data $\{N_i(t), C_i, \mathbf{Z}_i\}$. Note that, from the consistency in $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbb{C}}$, the limiting quantity of $V_{ij}(\hat{\boldsymbol{\beta}})$ is $\mathcal{V}_{ij}(\boldsymbol{\beta}_0)$. After replacing all unknown quantities in $\mathcal{W}$ with their respective consistent estimates and limits, the only random components in W^* are the i.i.d. $\{\epsilon_i\}$ that has mean one and variance one. Therefore, the conditional distribution of W^* given $\{N_i(t), C_i, \mathbf{Z}_i\}$ has the same limiting distribution as $\mathcal{W}$.