CityU
香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative
For The World

CityU Scholars

U-Statistics for left truncated and right censored data

Sudheesh, K. K.; Anjana, S.; Xie, M.

Published in:
Statistics

Published: 01/01/2023

Document Version:
Post-print, also known as Accepted Author Manuscript, Peer-reviewed or Author Final version

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1080/02331888.2023.2217314

Publication details:
Sudheesh, K. K., Anjana, S., & Xie, M. (2023). U-Statistics for left truncated and right censored data. *Statistics*, *57*(4), 900-917. https://doi.org/10.1080/02331888.2023.2217314

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.

Download date: 21/01/2025

This is an Accepted Manuscript of an article published by Taylor & Francis in Statistics on 24 May 2023, available online: http://www.tandfonline.com/10.1080/02331888.2023.2217314.

U-STATISTICS FOR LEFT TRUNCATED AND RIGHT CENSORED DATA

Sudheesh, K. K.[a†], Anjana, S.[b] and Xie, M.[c]

[a]Indian Statistical Institute, Chennai, India,

[b]University of Hyderabad, Hyderabad, India,

[c]City University of Hong Kong, Hong Kong.

Abstract. The analysis of left truncated and right censored data is very common in survival and reliability analysis. In lifetime studies patients are often subject to left truncation in addition to right censoring. For example, in bone marrow transplant studies based on International Bone Marrow Transplant Registry (IBMTR), the patients who die while waiting for the transplants will not be reported to the IBMTR. In this paper, we develop novel U-statistics under left truncation and right censoring. We prove the $\sqrt{n}$-consistency of the proposed U-statistics. We derive the asymptotic distribution of the U-statistics using the counting process technique. As an application of the U-statistics, we develop a simple non-parametric test for testing the independence between time to failure and cause of failure in competing risks when the observations are subject to left truncation and right censoring. The finite sample performance of the proposed test is evaluated through a Monte Carlo simulation study. Finally, we illustrate our test procedure using the lifetime data of transformers.

Key words: Competing risks; Left truncation; Right censoring; U-statistics.

1. Introduction

A common problem in practice is the presence of censoring event C that makes the variable of interest (lifetime) X unobservable. That is, some of

†Corresponding author E-mail: skkattu@isichennai.res.in.

the lifetimes may not be observed because of censoring. For example, in the context of medical trials where patients often survive beyond the end of the trial period or are lost to follow-up for some reason. In addition to censoring, one may have encountered left truncation, where we do not observe lifetime when $X < L$ where L is the truncation random variable. Left truncation occurs when the failure time of the subject under study is included in the sample if the failure time is greater than the truncation time. This happens when the study subjects enter into the study at a time point, which is not necessarily the time origin for the event of interest. Then we observe the subjects from this particular time until they experience the event of interest or until they are censored (Kalbfleisch and Prentice (2002), Klein and Moeschberger (2006)). Recently, Michimae and Emura (2022) developed likelihood inference for copula-based competing risks models under left truncation.

The present study was motivated by a real example of left truncated right censored data that appeared in Hong et al. (2009). The data consist of the lifetimes of high voltage power transformers from an energy company in the USA. There were approximately 15000 transformers and the company started recording information about the transformers in 1980. The data contain information on transformers which are installed before or after 1980 but failed after 1980. In this paper, we considered the data till 2008. Thus, the lifetime of the transformers which are still in service in 2008 is considered as right censored. Moreover, no information was available for the units which are installed and failed before 1980. Hong et al. (2009) were interested in predicting the remaining lifetime of the transformers and the rate of failure of these transformers over time. They used a parametric model for explaining the distribution of the lifetime of transformers. For the analysis of transformers data, Balakrishna and Mitra (2011, 2012, 2014) considered

different parametric models for the lifetime distribution of transformers under left truncation and right censoring. Kundu et al. (2017) used the same data with further information on the causes of failures of transformers. They considered the parametric analysis of the data in the presence of competing risks. The parametric models are more accurate only when we are able to correctly specify the underlying distribution. These facts motivated us to develop non-parametric inferences for such data.

Another example is the bone marrow transplant (BMT) study using International Bone Marrow Transplant Registry (IBMTR). The patients who die while waiting for the transplants will not be reported to the IBMTR and the patients who are lost to follow-up are subject to random right censoring. Hence it is important to study and develop methods to deal with left truncated and right censored samples. In the analysis of left truncated right censored data, researchers assume the quasi-independence between the truncation time and time to event. The above BMT example is the one where the longer waiting time is a common predictor for survival outcome and thus the left truncation is not random. Many researchers considered the copula-based models to handle the problem of dependent left truncation (Chaieb et al. (2006), Emura and Murotani (2015), Emura and Pan (2020)). Recently, Vakulenko-Lagun et al. (2022) proposed an inverse probability weighted approach to estimate the failure time distribution when shared covariates induce the dependence between the truncation time and event time. In competing risks case, the problem of dependent truncation is addressed by Stegherr et al. (2020), where they considered the inverse probability weighted approach to model the dependent left truncation. We refer interested readers to Jiang et al. (2005), Klein and Moeschberger (2006), Geskus (2011), Zhang et al. (2011), Su and Wang (2012), Vakulenko-Lagun and Mandel (2016), Cortese et al. (2017), Chen et al. (2017), Chen and Shen

(2018), Efromovich and Chu (2018), Hou et al. (2018), Jiang et al. (2020), Chen and Yi (2021) and Emura and Michimae (2022) and the references therein for some recent works based on left truncated and right censored data.

The theory of U-statistics has a major role in finding non-parametric estimators of parameters of interest. Interested readers may refer to Lee (1990) and Kowalski and Tu (2007) for more discussion about the application of U-statistics in different fields. Based on U-statistics, Jing et al. (2009) devolved jackknife empirical likelihood inference which had considerable attention recently. Due to the plethora of use cases in non-parametric inference, it is desirable to develop U-statistics under different censoring schemes and truncations. In this scenario, an important concern is the reworking and extension of the procedures which exist for completely observed data. Using the inverse probability of censoring weighted (IPCW) approach, Datta et al. (2010) developed a right-censored version of U-statistics. Satten et al. (2018) and Chen and Datta (2019) discussed comparison of two distributions using two sample U-statistics in the presence of right censoring and confounding covariates. Motivated by these works, in this paper, an attempt is made to develop U-statistics for left truncated and right censored data.

The rest of the paper is organized as follows. In Section 2, we develop novel U-statistics for left truncated and right censored data. We prove the consistency and asymptotic normality of the proposed U-statistics. We also obtain a consistent estimator of the asymptotic variance. In Section 3, making use of the U-statistics, we develop a new test for testing the independence between the cause of failure and failure time in competing risks under the left truncated and right censored data setup. The finite sample performance of the test is evaluated through a Monte Carlo simulation study. The proposed

method is illustrated using the lifetime data of transformers. Concluding remarks along with some open problems are given in Section 4.

2. **Proposed U-statistics**

Suppose X, C and L denote the failure time, censoring time and truncation time random variables respectively. Let $F(.)$ be the distribution function of X and $\bar{F}(x) = 1 - F(x)$. We assume that (L, C) is independent of X and follows a joint distribution G with $P(L < C) = 1$. The left truncated right censored data are the truncated part of an independent and identically distributed (iid) sample (Andersen et al., 1988). These truncated observations are independent and identically distributed conditional on the sample size (Weißbach and Dorre, 2022). Now suppose that data are available on n independent and identically distributed random vectors $(T_i\epsilon_i, \delta_i)$ drawn from $(T\epsilon, \delta)$, where $T = \min(X, C)$, $\delta = I(X < C)$ and $\epsilon = I(L < T)$. Here δ is the censoring indicator, while ϵ is used to specify the truncation. Under left truncation, the lifetime T is observed only when $T > L$ and thus the number of individual is not known at time zero (see Section 3 of Friedrich et al. (2017)).

We start by defining U-statistics for complete (uncensored) data. Let $X_1, ..., X_n$ be a random sample of size n from F. Let $h(X_1, ..., X_m)$ be a symmetric kernel of degree m with the property $E\big(h(X_1, ..., X_m)\big) = \theta$, where θ is real. The U-statistics with symmetric kernel h is defined as

$$U = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < ... < i_m \le n} h(X_{i_1}, ..., X_{i_m}),$$

where we use $i_1, ..., i_m$ to indicates m integers chosen from $(1, \ldots, n)$. U-statistics are widely used for finding estimators of several statistical functionals as well as for developing non-parametric tests as their asymptotic

properties are well established. Lehmann (1951) proved the strong consistency of U-statistics. Moreover, the asymptotic distribution of U is Gaussian with mean θ and variance $m^2\sigma^2$ (see Theorem 1, Chapter 3 of Lee (1990)) where $\sigma^2 = Var\big(E\big(h(X_1, ..., X_m)|X_1\big)\big)$.

Next we define U-statistics for left truncated and right censored (LTRC) data. We use IPCW approach to define U-statistics. To find the weight used in IPCW approach we consider

$$\begin{aligned} E(\delta\epsilon T) &= E(\delta\epsilon X) \\ &= E\big(E(\delta\epsilon X|X)\big) \\ &= E(XP(C > X > L|X)). \end{aligned} \tag{1}$$

The first identity follows from the fact that $T = X$ when $\delta = 1$. In view of equation (1), we consider a weight function $\frac{\delta\epsilon}{P(L<T<C)}$ for defining the U-statistics when the sample contain LTRC observations.

We define U-statistics for LTRC data as

$$U_m = \frac{1}{\binom{n}{m}} \sum_{1\le i_1<\ldots<i_m\le n} \frac{h(T_{i1}, ..., T_{im}) \prod_{l\in \underline{i}} \delta_l\epsilon_l}{\prod_{l\in \underline{i}} P(L_l < T_l < C_l)}. \tag{2}$$

provided $P(L_i < T_i < C_i) > 0$, for each i, with probability one. Here, the notation $l \in \underline{i}$ is used to indicate that l is one of the integers $\{i_1, i_2 \ldots, i_m\}$ chosen from $(1, \ldots, n)$. For $m = 1$ and $m = 2$ we have

$$U_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{h(T_i)\delta_i\epsilon_i}{P(L_i < T_i < C_i)}$$

and

$$U_2 = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{h(T_i, T_j)\delta_i\delta_j\epsilon_i\epsilon_j}{P(L_i < T_i < C_i)P(L_j < T_j < C_j)}.$$

Using equation (1), it can be easily verified that the U-statistics defined in equation (2) is an unbiased estimator of θ. Since U_m is a U-statistic with

kernel $\frac{h(T_{i1},...,T_{im})\prod_{l\in\underline{i}}\delta_l\epsilon_l}{\prod_{l\in\underline{i}}P(L_l<T_l<C_l)}$ it is a consistent estimator of θ and has asymptotic normal distribution.

As $P(L_i < T_i < C_i)$ appeared in the equation (2) is not known, we need to estimate it. We estimate it by $\widehat{K}_c(T_i)$ and is given in equation (6) below. Hence we define IPCW U-statistics under left truncation and right censoring as

$$\widehat{U}_m = \frac{1}{\binom{n}{m}} \sum_{1\le i_1<...<i_m\le n} \frac{h(T_{i1}, T_{i2}, ..., T_{im})\prod_{l\in\underline{i}}\delta_l\epsilon_l}{\prod_{l\in\underline{i}}\widehat{K}_c(T_l)}.$$

Next, we study the asymptotic properties of $\widehat{U}_m$. First, we establish the consistency of $\widehat{U}_m$. The proof of the following theorem is given in Appendix.

Theorem 2.1. *Assume $E|h(T_1, T_2, ..., T_m)| < \infty$. As $n \to \infty$, $\widehat{U}_m$ converges in probability to θ.*

Next, we obtain the asymptotic distribution of $\widehat{U}_m$. For $i = 1, 2, ..., n$, we define $N_i(t) = I(L_i < T_i \le t, \delta_i = 1)$ and $N_i^c(t) = I(L_i < T_i \le t, \delta_i = 0)$ as the counting process corresponding to the failure time and the censoring time, respectively. Also, denote $N(t) = \sum_{i=1}^n N_i(t)$, $N^c(t) = \sum_{i=1}^n N_i^c(t)$. We define risk indicator as $Y_i(t) = I(T_i \ge t \ge L_i)$ and $Y(t) = \sum_{i=1}^n Y_i(t)$. Note that the risk set $Y(t)$ at t contains the subjects who entered the study before t and are still under study at t. Clearly $N_i^c(t)$ is local sub-martingale with appropriate filtration $\mathbb{F}_t$. The martingale associated with the censoring counting process with filtration $\mathbb{F}_t$ is given by

$$M_i^c(t) = N_i^c(t) - \int_0^t Y_i(u)\lambda_c(u)du, \qquad i = 1, 2 \ldots, n, \tag{3}$$

where $\lambda_c(.)$ is the hazard function corresponding to the censoring variable C under left truncation. The cumulative hazard function of C is given by $\Lambda_c(t) = \int_0^t \lambda_c(u)du$. Denote $M^c(t) = \sum_{i=1}^n M_i^c(t)$.

Now, we define the sub-distribution function of T_1 corresponding to $\delta_1 = 1$ and $\epsilon_1 = 1$ as

$$S(x) = P(T_1 \leq x, \delta_1 \epsilon_1 = 1). \tag{4}$$

Let

$$w(t) = \int_0^\infty \frac{h_1(x)}{P(L_1 \leq x \leq C_1)} I(x > t) dS(x), \tag{5}$$

where $h_1(x) = E(h((T_1, \delta_1), \ldots, (T_m, \delta_m))|(T_1, \delta_1) = (x, \delta_1))$. Also, denote $y(t) = P(T_1 \geq t \geq L_1)$. An estimator of the survival function of censoring variable C under left truncation, denoted by $\widehat{K}_c$, is given by

$$\widehat{K}_c(\tau) = \prod_{t \leq \tau} \Big(1 - \frac{dN^c(t)}{Y(t)}\Big). \tag{6}$$

As an analog to the Nelson-Aalen estimator, the estimator for cumulative hazard function for C under left truncation is defined as

$$\widehat{\Lambda}_c(\tau) = \int_0^\tau \frac{dN^c(t)}{Y(t)}. \tag{7}$$

In both the definitions given in equations (6) and (7) we assume $Y(t)$ is non-zero with probability one. The relationship between $\widehat{K}_c(\tau)$ and $\widehat{\Lambda}_c(\tau)$ is given by

$$\widehat{K}_c(\tau) = \exp[-\widehat{\Lambda}_c(\tau)]. \tag{8}$$

Next, we state the assumptions needed to prove the asymptotic distributions.

C1: $E(h((T_1, \delta_1), \ldots, (T_m, \delta_m))) < \infty$,
C2: $\int \frac{h_1^2(x)}{\widehat{K}_c^2(x)} dS(x) < \infty$,
C3: $\int \frac{w^2(x)\lambda_c(x)}{y(x)} dx < \infty$.

The asymptotic distribution of $\widehat{U}_m$ is given in the next theorem and the proof of the same is given in the Appendix.

Theorem 2.2. *Under the conditions C1-C3, as* $n \to \infty$, $\sqrt{n}(\widehat{U}_m - \theta)$ *converges in distribution to Gaussian random variable with mean zero and variance* $m^2\sigma_c^2$, *where* σ_c^2 *is given by*

$$\sigma_c^2 = \sigma_1^2 + \sigma_2^2 \tag{9}$$

with

$$\sigma_1^2 = Var\Big(\frac{h_1(X)\delta_1\epsilon_1}{K_c(X)}\Big)$$

and

$$\sigma_2^2 = \int_0^\infty \frac{w^2(x)d\Lambda_c(x)}{y(x)}. \tag{10}$$

Next, we find a consistent estimator of the asymptotic variance σ_c^2. Using the re-weighting principle an estimator of $h_1(x)$ is given by

$$\widehat{h}_1(x) = \frac{1}{n^m} \sum_{1\le i_2<\ldots<i_m\le n} \frac{h(x, T_{i_2}, ..., T_{i_m})\delta_{i_2}\ldots\delta_{i_m}\epsilon_{i_2}\ldots\epsilon_{i_n}}{\widehat{K}_c(T_{i_2})\ldots\widehat{K}_c(T_{i_m})}. \tag{11}$$

A consistent estimator σ_1^2 is given by

$$\widehat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^n (V_i - \bar{V})^2,$$

where

$$V_i = \frac{\widehat{h}_1(T_i)\delta_i\epsilon_i}{\widehat{K}_c(T_i)} \qquad \text{and} \qquad \bar{V} = \frac{1}{n}\sum_{i=1}^n V_i.$$

Using equations (5) and (11), we find an estimator of $w(x)$ as

$$\begin{aligned} \widehat{w}(x) &= \int_0^\infty \frac{\widehat{h}_1(z)I(z>x)}{\widehat{K}_c(z)} d\widehat{S}(z) \\ &= \frac{1}{n}\sum_{i=1}^n \frac{\widehat{h}_1(T_i)I(T_i>x)\delta_i\epsilon_i}{\widehat{K}_c(T_i)}. \end{aligned} \tag{12}$$

We can estimate $y(x)$ by $\frac{Y(x)}{n}$. Hence using (7), from equation (10) we obtain an estimator of σ_2^2 as

$$\widehat{\sigma}_2^2 = \sum_{i=1}^{n} \frac{n\widehat{w}^2(T_i)(1-\delta_i)}{Y^2(T_i)}.$$

Therefore, an estimator of the asymptotic variance σ_c^2 is given by

$$\widehat{\sigma}_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(V_i - \bar{V})^2 + \sum_{i=1}^{n} \frac{\widehat{w}^2(T_i)(1-\delta_i)}{Y^2(T_i)/n}.$$

3. **Test for competing risks**

In classical survival studies, the study subjects are at risk of one terminal event. However, in many survival studies, the failure (death) of an individual may be due to one of k (say) causes. In such situations, each unit under study is exposed to k causes of failure, but its failure can be due to exactly one of these causes of failure. The data that arise from such contexts is known as competing risks data. In the literature, competing risks data are modeled using a latent failure time approach or through a bivariate random pair (X, J), where X is the failure time of the unit and $J \in \{1, 2, ..., k\}$ is the corresponding cause of failure. The latent failure time approach has identifiability issues (Tsiatis, 1975). The approach based on the observable random pair (T, J) helps us to overcome the identifiability issue that may arise with the latent failure time approach. In our study, we restrict to the case $k = 2$. For modeling and analysis of competing risks data, one may refer to Kalbfleisch and Prentice (2002), Lawless (2003) and Crowder (2012).

The joint distribution of (X, J) is specified by the sub-distribution functions (cumulative incidence functions)

$$F_r(t) = P(X \leq t, J = r), \quad r = 1, 2. \tag{13}$$

Then $F(x) = F_1(x) + F_2(x)$. We define $P_r = P(J = r)$, $r = 1, 2$ with $P_1 + P_2 = 1$. Also, define sub-survival function $S_r(t) = P(T > t, J = r)$, $r = 1, 2$ and $S(t) = P(T > t) = S_1(t) + S_2(t)$.

In the analysis of competing risks data, researchers considered the problem of testing the independence between X and J. If X and J are independent, then $F_r(t) = P(J = r)F(t)$ and hence one can study X and J separately. Accordingly, the testing problem reduces to test whether $P(J = 1) = P(J = 2) = \frac{1}{2}$ and the bivariate problem is converted to the problem involving only J. Furthermore, the time to failure and cause of failure are independent if and only if the cause-specific hazard rate functions are proportional (Crowder, 2012). We refer interested readers to Dewan et al. (2004) and Anjana et al. (2019) and the reference therein for more details on testing the independence between X and J. Anjana et al. (2019) discussed how to incorporate the right censored observation in their methodology. In analysing the transformer data discussed above, we are interested in developing a test for testing the independence between X and J under left truncated and right censored situation.

3.1. **Test statistics.** Next, we discuss the testing problem discussed above in the presence of left truncation and right censoring. Suppose T, X, C and L are the random variables as defined in Section 2 and J denotes the random variable corresponding to the cause of failure. Under right censoring and left truncation, we observe the competing risks data as $(T\epsilon, \delta, J\delta)$, where ϵ and δ are defined as in Section 2. Let $(T_i\epsilon_i, \delta_i, J_i\delta_i)$, $i = 1, 2, ..., n$ be independent copies of $(T\epsilon, \delta, J\delta)$. On the basis of the observed data, we are interested to test the null hypothesis

$$H_0 : T \text{ and } J \text{ are independent}$$

against the alternative hypothesis

$$H_1 : T \text{ and } J \text{ are not independent.}$$

To detect the departure from the null hypothesis H_0 towards the alternative hypothesis H_1, we consider a measure Δ given by

$$\Delta = \int_0^\infty \int_0^\infty (S_2(t_1)S_1(t_2) - S_2(t_2)S_1(t_1))dF_1(t)dF_2(t). \tag{14}$$

Next, we express Δ in a simple form. Consider

$$\begin{aligned}
&P(T_1 > T_2 > T_3 > T_4, J_1 = 2, J_3 = 1) \\
&= \int_0^\infty \int_0^\infty P(T_1 > t_1 > T_3 > t_2 J_1 = 2, J_3 = 1)dF(t_1)dF(t_2) \\
&= \int_0^\infty \int_0^\infty (P(T_1 > t_1)P(t_1 > T_3 > t_2))\, dF(t_1)dF(t_2) \\
&= \int_0^\infty \int_0^\infty (S_2(t_1)S_1(t_2) - S_2(t_1)S_1(t_1))\, dF(t_1)dF(t_2). \qquad (15)
\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}
&P(T_1 > T_2 > T_3 > T_4, J_1 = 1, J_3 = 2) \\
&= \int_0^\infty \int_0^\infty (S_2(t_1)S_2(t_2) - S_2(t_1)S_1(t_1))\, dF(t_1)dF(t_2). \qquad (16)
\end{aligned}$$

Substituting (15) and (16) in (14), we obtain

$$\begin{aligned}
\Delta = \; & P(T_1 > T_2 > T_3 > T_4, J_1 = 2, J_3 = 1) \\
& - P(T_1 > T_2 > T_3 > T_4, J_1 = 1, J_3 = 2).
\end{aligned}$$

It can be easily verified that Δ is zero under H_0 and positive under H_1. We use the U-statistics defined in Section 2 to find the test statistic $\widehat{\Delta}$. Define

the kernel ψ_c^*

$$\psi_c^*((T_1,J_1),(T_2,J_2),(T_3,J_3)(T_4,J_4)) = \begin{cases} 1 & \text{if} \begin{cases} T_1 > T_2 > T_3 > T_4, J_1 = 1, J_3 = 2 \end{cases} \\ -1 & \text{if} \begin{cases} T_1 > T_2 > T_3 > T_4, J_1 = 2, J_3 = 1. \end{cases} \end{cases}$$

Hence the test statistics is given by

$$\widehat{\Delta} = \frac{1}{\binom{n}{4}} \sum_{1\le i<j<l<k\le n} \frac{\psi_c((T_i,J_i),(T_j,J_j),(T_l,J_l)(T_k,J_k))\delta_i\delta_j\delta_l\delta_k\epsilon_i\epsilon_j\epsilon_l\epsilon_k}{\widehat{K}_c(T_i)\widehat{K}_c(T_j)\widehat{K}_c(T_l)\widehat{K}_c(T_k)},$$

where ψ_c is the symmetric version corresponding to ψ_c^*. Test procedure is to reject the null hypothesis H_0 against the alternative hypothesis H_1 for large values of $\widehat{\Delta}$. We obtain a critical region of the test based on the asymptotic distribution of $\widehat{\Delta}$. In the next theorem, we find the limiting distribution of $\widehat{\Delta}$. The proof of the following theorem follows from Theorem 2.2.

Theorem 3.1. *Let* $\psi_1^c(x) = E\big(\psi_c((T_1,J_1),(T_2,J_2),(T_3,J_3)(T_4,J_4))|T_1 = x, J_1 = j\big)$. *Assume* $E(\psi_c(T_1,T_2,T_3,T_4,J_1,J_2,J_3,J_4)) < \infty$, $\int \frac{(\psi_1^c(t))^2}{y^2(t)} dS(t) < \infty$ *and* $\int \frac{w^2(t)}{y(t)}\lambda_c(t)dt < \infty$. *As* $n \to \infty$, $\sqrt{n}(\widehat{\Delta} - \Delta)$ *is distributed as Gaussian with mean 0 and variance* $16\sigma_{1c}^2$, *where* σ_{1c}^2 *is given by*

$$\sigma_{1c}^2 = Var\left(\frac{\psi_1^c(T_1)\epsilon_1\delta_1}{\widehat{K}_c(T_1)}\right) + \int \frac{w^2(t)}{y(t)}\lambda_c(t)dt.$$

A consistent estimator of σ_{1c}^2 is given by

$$\widehat{\sigma}_{1c}^2 = \frac{1}{n-1}\sum_{i=1}^n (V_i - \bar{V})^2 + \sum_{i=1}^n \frac{n\widehat{w}^2(T_i)(1-\delta_i)}{Y^2(T_i)},$$

where

$$V_i = \frac{\widehat{\psi}_1^c(T_i)\epsilon_i\delta_i}{\widehat{K}_c(T_i)} \qquad \text{and} \qquad \bar{V} = \frac{1}{n}\sum_{i=1}^n V_i$$

and the estimators $\widehat{\psi}_1^c(x)$ and $\widehat{w}(x)$ can be obtained using the expressions (11) and (12), respectively by considering the kernel $\psi_c(.)$.

Using the asymptotic distribution obtained above, we obtain a critical region of the test. For large values of n, we reject the null hypothesis H_0 in favour of H_1 if

$$\frac{\sqrt{n}\widehat{\Delta}_c}{\widehat{\sigma}_{1c}} > Z_\alpha,$$

where Z_α is the upper α-percentile points of the standard normal distribution.

3.2. **Simulation study.** Next, we report the results of the Monte Carlo simulation study carried out to evaluate the performance of the proposed test procedure. The simulations are carried out using R software and repeated 10000 times. Lifetime random variable X is generated from a stan-

Table 1. Empirical type 1 error of the test

(a, P_1)	n	20% Censored		40% Censored	
		5 % level	1 % level	5 % level	1 % level
$(1, 0.45)$	50	0.0541	0.0145	0.0552	0.0154
	75	0.0524	0.0130	0.0538	0.0132
	100	0.0512	0.0119	0.0521	0.0121
	150	0.0506	0.0111	0.0513	0.0115
	200	0.0503	0.0105	0.0511	0.0110
$(1, 0.48)$	50	0.0642	0.0162	0.0661	0.0164
	75	0.0622	0.0155	0.0632	0.0160
	100	0.0592	0.0140	0.0603	0.0151
	150	0.0566	0.0132	0.0573	0.0136
	200	0.0523	0.0115	0.0530	0.0119

dard exponential distribution. Censoring variable C is generated from an exponential distribution with parameter γ, where γ is chosen in such a way that the sample contains the desired percentage of censored observations, that is, $P(X > C) = p$, $0 < p < 1$. In the present study, we considered two situations with 20% and 40% of the observations that are censored. The truncated variable L is generated from an exponential distribution with λ, where λ satisfy $P(L > X) = 0.2$, which guaranteed 20% observations are truncated.

For generating the random samples from competing risks with two causes of failure, we consider the parametric family of sub-distribution functions given by (Dewan and Kulathinal, 2009).

$$F_1(t) = P_1 F^a(t) \quad \text{and} \quad F_2(t) = F(t) - F_1(t),$$

where $1 \leq a \leq 2$ and $0 < P_1 < 0.5$. If $a = 1$, then T and J are independent.

We computed the empirical type 1 error at 5% and 1% levels of significance and the result is presented in Table 1. From Table 1, we observe that the empirical type 1 error are close to the chosen level of significance. We calculated the empirical power of the test which is given in Table 2. From Table 2, we observe that the test has good power in general and the power increases when the sample size increases and the value of a deviates from the null hypothesis value ($a = 1$). We also observe from Table 2 that the power of the test decreases as the censoring percentage increases.

Table 2. Empirical power the test: Exponential distribution

(a, P_1)	n	20% Censored		40% Censored	
		5 % level	1 % level	5 % level	1 % level
	50	0.5412	0.4652	0.3404	0.2863
	75	0.6564	0.5809	0.3898	0.3279
	100	0.8022	0.6944	0.5683	0.5021
	150	0.9682	0.8048	0.7656	0.6892
$(1.5, 0.3)$	200	1.0000	0.9620	0.8834	0.8238
	50	0.6802	0.5585	0.4121	0.3782
	75	0.7106	0.6011	0.4983	0.4387
	100	0.8604	0.7240	0.6482	0.5741
	150	0.9805	0.8436	0.8639	0.7930
$(1.9, 0.3)$	200	1.0000	0.9829	0.9422	0.9058

Next, we simulate lifetime from Weibull random variable with shape parameter $\theta = 2$ where the distribution function is given by $F(x) = 1 - e^{-x^\lambda}$, $\lambda > 1$, $x \geq 0$. The censoring variable C and truncated variable L are simulated as above. The empirical power obtained in this case is reported in

Table 3. In this case also, from Table 3, we observe that power increases as the sample size increases.

Table 3. Empirical power the test: Weibull distribution

(a, P_1)	n	20% Censored		40% Censored	
		5 % level	1 % level	5 % level	1 % level
$(1.5, 0.3)$	50	0.5219	0.4363	0.3331	0.2728
	75	0.6436	0.5602	0.3938	0.3040
	100	0.8001	0.6761	0.5459	0.4983
	150	0.9529	0.7883	0.7581	0.6634
	200	1.0000	0.9636	0.8776	0.8033
$(1.9, 0.3)$	50	0.6466	0.5375	0.4121	0.3584
	75	0.7049	0.5980	0.4831	0.4158
	100	0.8448	0.7042	0.6160	0.5681
	150	0.9763	0.8389	0.8445	0.7708
	200	1.0000	0.9801	0.9384	0.9003

3.3. **Data analysis.** In this section, we illustrate the proposed test procedure using an extract of the real data set mentioned in Hong et al. (2009). The data provide information on the lifetime of transformers from an energy company. There were approximately 15000 transforms and the company started recording information about the transformers in 1980. For the analysis, Hong et al. (2009) considered the data till 2008. The data contain information on transformers which are installed before or after 1980 but failed after 1980. The lifetime of the transformers which are still in service in 2008 is considered as right censored. Moreover, no information was available for the units which were installed and failed before 1980. Also, the data contain information about the possible causes of failure of each unit. Hence the lifetime of the transformers can be treated as left truncated and right censored competing risks data. Many authors considered the analysis of this data. Balakrishnan and Mitra (2012) and Kundu et al. (2017) considered the extract of this data (data of sample size 100) for the analysis. For analysing transformer data, Kundu et al. (2017) developed a parametric model for

latent failure times under left truncated and right censored competing risks setup. We consider the same data set for the analysis. In our study, we are interested to test the independence of the lifetime of transformers and associated causes of failure.

Table 4. Transformer Data

S.N.	Year Inst.	Year Exit	ν	K	S.N.	Year Inst.	Year Exit	ν	K	S.N.	Year Inst.	Year Exit	ν	K
1	1961	1996	0	2	35	1989	2008	1	0	69	1983	2006	1	2
2	1964	1985	0	1	36	1981	2008	1	0	70	1983	1993	1	1
3	1962	2007	0	2	37	1985	2008	1	0	71	1989	2008	1	0
4	1962	1986	0	2	38	1986	2004	1	2	72	1989	2008	1	0
5	1961	1992	0	2	39	1980	1987	1	2	73	1986	2008	1	0
6	1962	1987	0	1	40	1986	2005	1	1	74	1982	1999	1	2
7	1964	1993	0	2	41	1980	2008	1	0	75	1985	2008	1	0
8	1960	1984	0	2	42	1982	2008	1	0	76	1986	2008	1	0
9	1963	1997	0	2	43	1986	2008	1	0	77	1982	2008	1	0
10	1962	1995	0	2	44	1984	2008	1	0	78	1988	2004	1	1
11	1963	2008	0	0	45	1986	1995	1	2	79	1980	2008	1	0
12	1963	2000	0	1	46	1986	2008	1	0	80	1982	2002	1	2
13	1960	1981	0	2	47	1987	2008	1	0	81	1981	2006	1	2
14	1963	1984	0	2	48	1986	2008	1	0	82	1988	1996	1	1
15	1963	1993	0	2	49	1986	2008	1	0	83	1985	2002	1	2
16	1964	1992	0	2	50	1984	2008	1	0	84	1984	2008	1	0
17	1961	1981	0	2	51	1984	2001	1	2	85	1980	2008	1	0
18	1960	1995	0	1	52	1983	2008	1	0	86	1982	2008	1	0
19	1961	2008	0	0	53	1988	2008	1	0	87	1981	1995	1	2
20	1960	2002	0	1	54	1988	2008	1	0	88	1986	1997	1	2
21	1960	1988	0	1	55	1985	2008	1	0	89	1986	2008	1	0
22	1961	1993	0	2	56	1986	2008	1	0	90	1986	2008	1	0
23	1961	1990	0	2	57	1988	2008	1	0	91	1982	2008	1	0
24	1960	1986	0	1	58	1982	2008	1	0	92	1989	2008	1	0
25	1962	2008	0	0	59	1985	2008	1	0	93	1984	2008	1	0
26	1964	1982	0	2	60	1988	2008	1	0	94	1980	2008	1	0
27	1963	1984	0	1	61	1982	2004	1	2	95	1988	2008	1	0
28	1960	1987	0	2	62	1980	2008	1	0	96	1986	2008	1	0
29	1962	1996	0	2	63	1980	2002	1	2	97	1982	1996	1	2
30	1963	1994	0	1	64	1984	2008	1	0	98	1982	2008	1	0
31	1987	2008	1	0	65	1981	1999	1	1	99	1982	2008	1	0
32	1980	2008	1	0	66	1986	2007	1	2	100	1989	2008	1	0
33	1988	2008	1	0	67	1987	2008	1	0					
34	1985	2008	1	0	68	1983	2008	1	0					

The transformer data of size 100 is presented in Table 4. In Table 4, ν represents the truncation indicator. Here $\nu = 1$ specifies that the transformer was installed after 1980 and $\nu = 0$ specifies that the transformer was

installed before 1980. $K = 1$ indicates that the failure is due to cause 1 and $K = 2$ indicates that the failure is due to cause 2. $K = 0$ denoted the right censored observations. For the transformer data, we estimated the cumulative incidence functions given in equation (13) using the Aalen-Johansen estimator. Figure 1 displays the plot of the estimators of the cumulative incidence functions along with 95% confidence intervals corresponding to cause 1 and cause 2. From Figure 1, it can be observed that there are time intervals without any failure for cause 1 compared to cause 2. It may be due to the fact that the number of failures in the data due to cause 1 is less than the number of failures due to cause 2. It is also evident that the chance of failure due to cause 2 is more compared to cause 1 as the lifetime increases. Now, we calculated the value of $\hat{\Delta}$ which is obtained as 88.16, indicating that the lifetime and causes of failure are dependent.

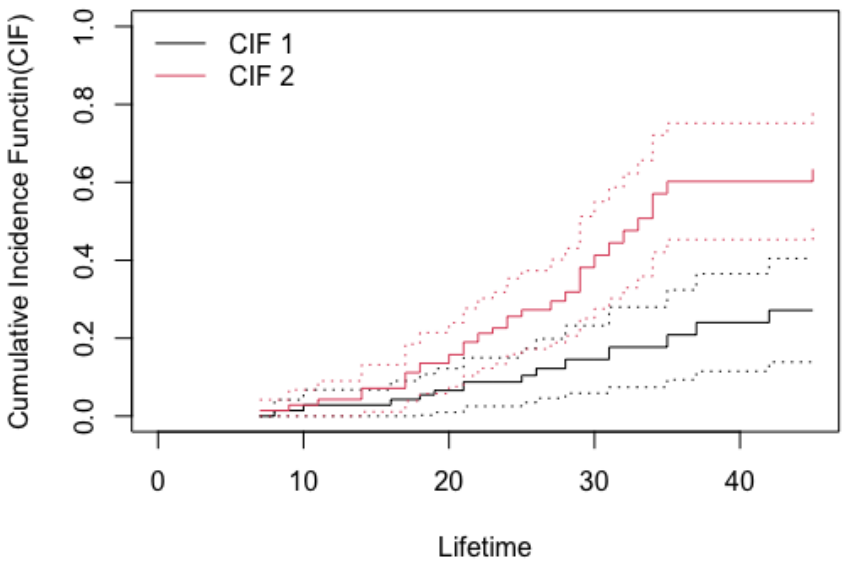

Figure 1. Cumulative incidence functions

4. Conclusion

In many applications involving follow-up studies, the lifetime of a patient is often subject to left truncation in addition to random right censoring. We develop U-statistics for left truncated and right censored data. We use the inverse probability weighted technique for developing the U-statistics. The U-statistics can even be used when we have ties in the data since the Kaplan-Meier estimator defined in equation (6) can incorporate the ties present in the data. The asymptotic properties of the U-statistics are studied. We proved the $\sqrt{n}$- consistency of the proposed U-statistics. We derived the asymptotic distribution of the U-statistics as normal. We then obtained a consistent estimator of the asymptotic variance. As an application, we develop a new test for testing the independence between the cause of failure and failure time in competing risks. The finite sample performance of the test is evaluated through a Monte Carlo simulation study. The test procedure is illustrated using the failure time data of transformers reported by Hong et al. (2009). We established that the failure time of the transformers is dependent on the cause of failure.

In this study, we looked at the situation of random left truncation along with right censoring to develop the U-statistics. The problem of developing U-statistics in dependent left truncation can be considered. As mentioned in the introduction, a large number of parameters were estimated using U-statistics. Hence it is important to develop U-statistics under the different censoring schemes. Some important censoring mechanisms that appeared in medical research are interval censoring and double censoring (see Chapter 8 of Sun (2006)). Extending our works to these censoring schemes can be considered for future works.

Acknowledgment

We thank the anonymous reviewers for their constructive comments on the earlier version of the manuscript which enabled us to improve substantially.

References

[1] Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes.* Springer Science & Business Media, New York.

[2] Anjana, S., Dewan, I. and Sudheesh, K. K. (2019). Test for independence between time to failure and cause of failure in competing risks with k causes of failure. *Journal of Nonparametric Statistics,* 31, 322–339.

[3] Balakrishnan, N. and Mitra, D. (2011). Likelihood inference for lognormal data with left truncation and right censoring with an illustration. *Journal of Statistical Planning and Inference*, 141, 3536–3553.

[4] Balakrishnan, N. and Mitra, D. (2012). Left truncated and right censored Weibull data and likelihood inference with an illustration. *Computational Statistics & Data Analysis*, 56, 4011–4025.

[5] Balakrishnan, N. and Mitra, D. (2014). Some further issues concerning likelihood inference for left truncated and right censored lognormal data. *Communications in Statistics-Simulation and Computation*, 43, 400–416.

[6] Chaieb, L. L., Rivest, L. P. and Abdous, B. (2006). Estimating survival under a dependent truncation.*Biometrika*, 93, 655–669.

[7] Chen, Y. and Datta, S. (2019). Adjustments of multi-sample U-statistics to right censored data and confounding covariates. *Computational Statistics & Data Analysis,* 135, 1–14.

[8] Chen, C. M. and Shen, P. S. (2018). Conditional maximum likelihood estimation in semiparametric transformation model with LTRC data. *Lifetime Data Analysis*, 24, 250–272.

[9] Chen, C. M., Shen, P. S., Wei, J. C. C. and Lin, L. (2017). A semiparametric mixture cure survival model for left truncated and right censored data. *Biometrical Journal,* 59, 270–290.

[10] Chen, L. P. and Yi, G. Y. (2021). Semiparametric methods for left-truncated and right-censored survival data with covariate measurement error. *Annals of the Institute of Statistical Mathematics,* 73, 481–517.

[11] Crowder, M. J. (2012). *Multivariate Survival Analysis and Competing Risks.* CRC Press, Boca Raton.

[12] Cortese, G., Holmboe, S. A. and Scheike, T. H. (2017). Regression models for the restricted residual mean life for right censored and left truncated data. *Statistics in Medicine,* 36, 1803–1822.

[13] Datta, S., Bandyopadhyay, D. and Satten, G. A. (2010). Inverse probability of censoring weighted U-statistics for right censored data with an application to testing hypotheses. *Scandinavian Journal of Statistics,* 37, 680–700.

[14] Dewan, I., Deshpande, J. and Kulathinal, S. (2004). On testing dependence between time to failure and cause of failure via conditional probabilities. *Scandinavian Journal of Statistics,* 31, 79–91.

[15] Dewan, I. and Kulathinal, S. (2007). On testing dependence between time to failure and cause of failure when causes of failure are missing. *PloS One*, 2, 1255–1264.

[16] Efromovich, S. and Chu, J. (2018). Hazard rate estimation for left truncated and right censored data. *Annals of the Institute of Statistical Mathematics*, 70, 889–917.

[17] Emura, T. and Michimae, H. (2022). Left-truncated and right-censored field failure data: Review of parametric analysis for reliability. *Quality and Reliability Engineering International*, 38, 3919–3934.

[18] Emura, T. and Murotani, K. (2015). An algorithm for estimating survival under a copula-based dependent truncation model. *Test*, 24, 734–751.

[19] Emura, T. and Pan, C. H. (2020). Parametric likelihood inference and goodness-of-fit for dependently left-truncated data, a copula-based approach. *Statistical Papers*, 61, 479–501.

[20] Friedrich, S., Beyersmann, J., Winterfeld, U., Schumacher, M. and Allignol, A. (2017). Nonparametric estimation of pregnancy outcome probabilities. *The Annals of Applied Statistics*, 11, 840–867.

[21] Geskus, R. B. (2011). Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring. *Biometrics*, 67, 39–49.

[22] Hong, Y., Meeker, W. Q. and McCalley, J. D. (2009). Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *The Annals of Applied Statistics*, 3, 857–879.

[23] Hou, J., Chambers, C. D. and Xu, R. (2018). A nonparametric maximum likelihood approach for survival data with observed cured subjects, left truncation and right-censoring. *Lifetime Data Analysis*, 24, 612–651.

[24] Jiang, H., Fine, J. P. and Chappell, R. (2005). Semiparametric analysis of survival data with left truncation and dependent right censoring. *Biometrics,* 61, 567–575.

[25] Jiang, W., Ye, Z. and Zhao, X. (2020). Reliability estimation from left-truncated and right-censored data using splines. *Statistica Sinica,* 30, 845–875.

[26] Jing, B. Y., Yuan, J. and Zhou, W. (2009). Jackknife empirical likelihood. *Journal of the American Statistical Association*, 104, 1224–1232.

[27] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York.

[28] Klein, J. P. and Moeschberger, M. L. (2006). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media, New York.

[29] Kowalski, J. and Xin, M. T. (2008). *Modern Applied U-Statistics*. John Wiley & Sons, New Jersey.

[30] Kundu, D., Mitra, D. and Ganguly, A. (2017). Analysis of left truncated and right censored competing risks data. *Computational Statistics & Data Analysis*, 108, 12–26.

[31] Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, New York.

[32] Lee, A. J. (1990). *U-Statistics: Theory and Practice*. CRC Press, Boca Raton.

[33] Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, 22, 165–179.

[34] Michimae, H. and Emura, T. (2022). Likelihood inference for copula models based on left-truncated and competing risks data from field studies. *Mathematics,* 10, 2163.

[35] Satten, G. A., Kong, M. and Datta, S. (2018). Multisample adjusted U-statistics that account for confounding covariates. *Statistics in Medicine*, 37, 3357–3372.

[36] Stegherr, R., Allignol, A., Meister, R., Schaefer, C. and Beyersmann, J. (2020). Estimating cumulative incidence functions in competing risks data with dependent left-truncation. *Statistics in Medicine*, 39, 481–493.

[37] Su, Y. R. and Wang, J. L. (2012). Modeling left-truncated and right-censored survival data with longitudinal covariates. *The Annals of Statistics*, 40, 1465–1488.

[38] Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data*, Springer, New York.

[39] Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72, 20–22.

[40] Vakulenko-Lagun, B. and Mandel, M. (2016). Comparing estimation approaches for the illness–death model under left truncation and right censoring. *Statistics in Medicine*, 35, 1533–1548.

[41] Vakulenko-Lagun, B., Qian, J., Chiou, S. H., Wang, N. and Betensky, R. A. (2022). Nonparametric estimation of the survival distribution under covariate-induced dependent truncation. *Biometrics*, 78, 1390–1401

[42] Weißbach, R. and Dörre, A. (2022). Retrospective sampling of survival data based on a Poisson birth process: conditional maximum likelihood. *Statistics*, 56, 844–866.

[43] Zhang, X., Zhang, M. J. and Fine, J. (2011). A proportional hazards regression model for the subdistribution with right censored and left truncated competing risks data. *Statistics in Medicine*, 30, 1933–1951.

Appendix

Proof of Theorem 1: We prove the theorem for $m = 1, 2$ and proofs for the other cases are similar. Consider

$$\begin{aligned}
\widehat{U}_1 &= \frac{1}{n}\sum_{i=1}^{n}\frac{h(T_i)\delta_i\epsilon_i}{\widehat{K}_c(T_i)} \\
&= -\frac{1}{n}\sum_{i=1}^{n}\frac{h(T_i)\delta_i\epsilon_i(\widehat{K}_C(T_i) - P(L_i < T_i < C_i))}{\widehat{K}_c(T_i)P(L_i < T_i < C_i)} + \frac{1}{n}\sum_{i=1}^{n}\frac{h(T_i)\delta_i\epsilon_i}{P(L_i < T_i < C_i)} \\
&= A_1 + A_2 \qquad \text{(say)}. \qquad (17)
\end{aligned}$$

Since A_2 is a U-statistic with kernel $\frac{h(T_i)\delta_i\epsilon_i}{P(L_i<T_i<C_i)}$, as $n \to \infty$, A_2 converges in probability to θ. Note that $\widehat{K}_c(T_i)$ is a consistent estimator of $P(L_i < T_i < C_i)$. Also we have

$$\max_{1\le i\le n}\frac{\sqrt{n}\left|\widehat{K}_c(T_i) - P(L_i < T_i < C_i))\right|}{\widehat{K}_c(T_i)} = O_p(1),$$

Now, consider

$$\begin{aligned}
|A_1| &= \Big| -\frac{1}{n}\sum_{i=1}^{n} \frac{h(T_i)\delta_i\epsilon_i(\widehat{K}_C(T_i) - P(L_i < T_i < C_i))}{\widehat{K}_C(T_i)P(L_i < T_i < C_i)}\Big| \\
&\leq \max_{1\leq i\leq n} \frac{\sqrt{n}\Big|(\widehat{K}_C(T_i) - P(L_i < T_i < C_i))\Big|}{\widehat{K}_C(T_i)} \frac{1}{\sqrt{n}}.\frac{1}{n}\sum_{i=1}^{n}\Big|\frac{h(T_i)\delta_i\epsilon_i}{P(L_i < T_i < C_i)}\Big| \\
&\leq O_p(1)o_p(1)O_p(1) = o_p(1).
\end{aligned}$$

Hence from the representation given in (17), we have the result for $m = 1$.

Next, consider the case $m = 2$, where $\widehat{U}_c$ is given by

$$\begin{aligned}
\widehat{U}_2 &= \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j<i;j=1}^{n} \frac{h(T_i,T_j)\delta_i\delta_j\epsilon_i\epsilon_j}{\widehat{K}_c(T_i)\widehat{K}_c(T_j)} \\
&= \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j<i;j=1}^{n} \frac{h(T_i,T_j)\delta_i\delta_j(\widehat{K}_c(T_j) - P(L_j < T_j < C_j))}{\widehat{K}_c(T_i)\widehat{K}_c(T_j)} \\
&\qquad \times\frac{(\widehat{K}_c(T_i) - P(L_i < T_i < C_i))}{P(L_i < T_i < C_i)P(L_j < T_j < C_j)} \\
&\quad + \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j<i;j=1}^{n} \frac{h(T_i,T_j)\delta_i\delta_j}{\widehat{K}_c(T_i)P(L_j < T_j < C_j)} \\
&\quad + \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j<i;j=1}^{n} \frac{h(T_i,T_j)\delta_i\delta_j}{P(L_i < T_i < C_i)\widehat{K}_c(T_j)} \\
&\quad - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j<i;j=1}^{n} \frac{h(T_i,T_j)\delta_i\delta_j}{P(L_i < T_i < C_i)P(L_j < T_j < C_j)} \\
&= \widehat{U}_{12} + \widehat{U}_{22} + \widehat{U}_{32} - \widehat{U}_{42}. \qquad \text{(say)}
\end{aligned} \tag{18}$$

Note that $\widehat{K}_c(T_1)$ is a consistent estimator of $P(L_1 < T_1 < C_1)$. Hence, as $n \to \infty$

$$\begin{aligned}
|\widehat{U}_{12}| &\leq \sup_{T_i}|(\widehat{K}_c(T_i) - P(L_i < T_i < C_i))| \sup_{T_j}|(\widehat{K}_c(T_j) - P(L_j < T_j < C_j))| \\
&\qquad \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j<i;j=1}^{n} \Big|\frac{h(T_i,T_j)\delta_i\delta_j}{\widehat{K}_c(T_i)\widehat{K}_c(T_j)P(L_i < T_i < C_i)P(L_j < T_j < C_j)}\Big| \\
&= o_p(1)o_p(1)O_p(1) = o_p(1).
\end{aligned} \tag{19}$$

Similar lines as above we can show that

$$\begin{aligned}\widehat{U}_{22} &= \widehat{U}_{42} + \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j<i;j=1}^{n}\frac{h(T_i,T_j)\delta_i\delta_j(\widehat{K}_c(T_i) - P(L_i < T_i < C_i))}{P(L_i < T_i < C_i)\widehat{K}_c(T_i)P(L_j < T_j < C_j)} \\ &= \widehat{U}_{42} + o_p(1). \qquad (20)\end{aligned}$$

and

$$\begin{aligned}\widehat{U}_{32} &= \widehat{U}_{42} + \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j<i;j=1}^{n}\frac{h(T_i,T_j)\delta_i\delta_j(\widehat{K}_c(T_j) - P(L_j < T_j < C_j))}{P(L_i < T_i < C_i)\widehat{K}_c(T_j)P(L_j < T_j < C_j)} \\ &= \widehat{U}_{42} + o_p(1). \qquad (21)\end{aligned}$$

Substituting equations (19), (20) and (21) in equation (18) we obtain

$$\widehat{U}_2 = \widehat{U}_{42} + o_p(1).$$

We observe that $\widehat{U}_{42}$ is a U-statistic with kernel $\frac{h(Y_i,Y_j)\delta_i\delta_j\epsilon_i\epsilon_j}{P(L_i<T_i<C_i)P(L_j<T_j<C_j)}$. Hence, as $n \to \infty$, $\widehat{U}_{42}$ converges in probability to θ (Lehmann, 1951). Accordingly, for $m = 2$, as $n \to \infty$, $\widehat{U}_2$ converges in probability to θ.

Proof of Theorem 2: For convenience, denote $K_c(T_i) = P(L_i < T_i < C_i)$ for any $i = 1, \ldots, n$. First, we consider the decomposition

$$\sqrt{n}(\widehat{U}_m - \theta) = \sqrt{n}(U_m - \theta) + \sqrt{n}(\widehat{U}_m - U_m). \qquad (22)$$

Since $\widehat{K}_c(T_i)$ is a $\sqrt{n}$-consistent for $K_c(T_i)$, we have $\sqrt{n}(\widehat{K}_c(T_i) - K_c(T_i)) = O_p(1)$. Hence the second term in the decomposition (22) can be written as (for the detailed steps see the proof above to show $\widehat{U}_{12} = o_p(1)$)

$$\begin{aligned}\sqrt{n}(\widehat{U}_m - U_m) &= \frac{\sqrt{n}}{\binom{n}{m}}\sum_{1\le i_1<\ldots<i_m}\frac{h(T_{i_1},\ldots,T_{i_m})\prod_{l\in\underline{i}}\delta_l\epsilon_l\Big(\prod_{l\in\underline{i}}(\widehat{K}_c(T_l) - K_c(T_l))\Big)}{\prod_{l\in\underline{i}}K_c(T_l)\prod_{l\in\underline{i}}K_c(T_l)} \\ &\quad + o_p(1).\end{aligned}$$

Hence using Hoeffding (1948) decomposition, from equation (22) we have

$$\begin{aligned}\sqrt{n}(\widehat{U}_m-\theta) &= \frac{m}{\sqrt{n}}\sum_{i=1}^{n}\frac{h_1(T_i)\delta_i\epsilon_i}{K_c(T_i)}-\theta \\ &- \frac{m}{\sqrt{n}}\sum_{i=1}^{n}\frac{h_1(T_i)\delta_i\epsilon_i(\widehat{K}_c(T_i)-K_c(T_i))}{K_c^2(T_i)}+o_p(1), \quad (23)\end{aligned}$$

where

$$h_1(x)=E\big(h(T_1,...,T_m)|T_1=x\big).$$

Using the relationship between $\widehat{K}_c(.)$ and $\widehat{\Lambda}_c(.)$ given in (8) and by delta method we have

$$\sqrt{n}(\widehat{K}_c(T_i)-K_c(T_i)) = -\sqrt{n}K_c(T_i)(\widehat{\Lambda}_c(T_i)-\Lambda_c(T_i))+o_p(1).$$

Hence using the martingale representation of $\widehat{\Lambda}_c(T_i)$ (see Page 178 of Andersen et al. (1993)), equation (23) becomes

$$\begin{aligned}\sqrt{n}(\widehat{U}_m-\theta) &= \frac{m}{\sqrt{n}}\sum_{i=1}^{n}\frac{h_1(T_i)\delta_i\epsilon_i}{K_c(T_i)}-\theta \\ &\quad +\frac{m}{\sqrt{n}}\sum_{i=1}^{n}\frac{h_1(T_i)\delta_i\epsilon_i(\widehat{\Lambda}_c(T_i)-\Lambda_c(T_i))}{K_c(T_i)}+o_p(1) \\ &= \frac{m}{\sqrt{n}}\sum_{i=1}^{n}\frac{h_1(T_i)\delta_i\epsilon_i}{K_c(T_i)}-\theta \\ &\quad +\frac{m}{\sqrt{n}}\sum_{i=1}^{n}\frac{h_1(T_i)\delta_i\epsilon_i}{K_c(T_i)}\int_0^{T_i}\frac{dM^c(x)}{Y(x)}+o_p(1), \quad (24)\end{aligned}$$

provided $Y(x)>0$ with probability one.

Now, recall the definition of $S(x)$ given in equation (4), we can express

$$m\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{h_1(T_i)\delta_i\epsilon_i}{K_c(T_i)}\int_0^{T_i}\frac{dM^c(x)}{Y(x)}=m\sqrt{n}\int_0^{\infty}\frac{h_1(z)}{K_c(z)}\left(\int_0^{z}\frac{dM^c(x)}{Y(x)}\right)dS(z).$$

Using Fubini's theorem, changing the order of integration gives

$$m\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{h_1(T_i)\delta_i\epsilon_i}{K_c(T_i)}\int_0^{T_i}\frac{dM^c(x)}{Y(x)}=m\sqrt{n}\int_0^{\infty}\left(\int_x^{\infty}\frac{h_1(z)}{P(L<z<C)}dS(z)\right)\frac{dM^c(x)}{Y(x)}.$$

We denote the right hand side of the above equation as

$$m\sqrt{n}\int_0^\infty \frac{w(x)}{Y(x)}dM^c(x),$$

where $w(x) = \int_0^\infty \frac{h_1(z)I(z>x)}{K_c(z)}dS(z)$. Hence we can write (24) as

$$\begin{aligned}\sqrt{n}(\widehat{U}_m - \theta) &= \frac{m}{\sqrt{n}}\sum_{i=1}^n \frac{h_1(T_i)\delta_i\epsilon_i}{K_c(T_i)} - \theta + m\sqrt{n}\int_0^\infty \frac{w(x)}{Y(x)}dM^c(x) + o_p(1)\\ &= D_1 + D_2 + o_p(1). \qquad \text{(say)}\end{aligned}$$

Using central limit theorem, as $n \to \infty$, D_1 converges in distribution to Gaussian random variable with mean zero and variance σ_1^2 where σ_1^2 is given by

$$\sigma_1^2 = m^2 Var\Big(\frac{h_1(X)\delta_1\epsilon_1}{K_c(X)}\Big).$$

As $n \to \infty$, $Y(x)/n$ converges in probability to $y(x)$. Using martingale central limit theorem, as $n \to \infty$, D_2 converges in distribution to Gaussian random variable with mean zero and variance σ_2^2, where σ_2^2 is the limit of the predictable variation process given by

$$\begin{aligned}\sigma_2^2 &= \lim_{n\to\infty} m^2 n\int_0^\infty \frac{w^2(x)Y(x)d\Lambda_c(x)}{Y^2(x)}\\ &= \lim_{n\to\infty} m^2\int_0^\infty \frac{w^2(x)d\Lambda_c(x)}{Y(x)/n}\\ &= m^2\int_0^\infty \frac{w^2(x)d\Lambda_c(x)}{y(x)}.\end{aligned}$$

Hence we have the variance expression given in the Theorem 2.2. The proof is completed when we show that the asymptotic covariance between D_1 and D_2 is zero.

Consider

$$\begin{aligned}
&\Big|\big(\frac{m}{\sqrt{n}}\sum_{i=1}^{n}\frac{h_1(T_i)\delta_i\epsilon_i}{K_c(T_i)}-\theta\big)\big(m\sqrt{n}\int_0^\infty\frac{w(x)}{Y(x)}dM^c(x)\big)\Big| \\
&= \quad \Big|\frac{m}{n}\sum_{i=1}^{n}\frac{h_1(T_i)\delta_i\epsilon_i}{K_c(T_i)}-\theta\Big|\Big|mn\int_0^\infty\frac{w(x)}{Y(x)}dM^c(x)\Big| \\
&\leq \quad \frac{m}{n}\sum_{i=1}^{n}\Big|\frac{h_1(T_i)\delta_i\epsilon_i}{K_c(T_i)}-\theta\Big|\Big|m\int_0^\infty\frac{w(x)}{Y(x)/n}dM^c(x)\Big| \\
&\leq \quad O_p(1).o_p(1)=o_p(1).
\end{aligned}$$

This completes the proof of the theorem.