Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

This is the peer reviewed version of the following article: [A note on the applicability of the standard nonparametric maximum likelihood estimator for combined incident and prevalent cohort data. Stat 9, 1 (2020)], which has been published in final form at https://doi.org/10.1002/sta4.280

ARTICLE TYPE

A note on the applicability of the standard non-parametric maximum likelihood estimator for combined incident and prevalent cohort data

James H. McVittie* | David B. Wolfson | David A. Stephens

Department of Mathematics and Statistics, McGill University, Montreal, H3A 0B9, Canada

Correspondence
*James H. McVittie, Department of Mathematics and Statistics, McGill University, Montreal, H3A 0B9, Canada
Email: james.mcvittie@mail.mcgill.ca

Abstract

Nonparametric estimation of the survival function for either incident or prevalent cohort failure time data, exclusively, has been well studied in the literature; the Kaplan-Meier (KM) estimator is routinely used for right-censored incident cohort failure time data while a modified form of the KM estimator, sometimes referred to as the Tsai-Jewell-Wang (TJW) estimator, is the default estimator used for prevalent cohort data with follow-up. Often, failure time data comprise observations from a combination of incident and prevalent cohorts. In this note, we justify the use of the TJW estimator for a combined sample of incident and prevalent cohort data with follow-up. We suggest how the TJW estimator forms the basis for density estimation and hypothesis testing problems, when incident and prevalent cohorts are combined.

KEYWORDS:
Censored Data, Large Sample Theory, Nonparametric Methods, Survival Analysis

1 | INTRODUCTION

It is common in survival analysis for the collected data to comprise the union of two data sources; typically, one source is a classical incident cohort study with follow-up while the other source is a prevalent cohort study with follow-up. In the latter, subjects with prevalent disease are followed forward in time. The development of statistical procedures for such combined cohort data allows researchers to avoid the shortcomings of analyses based on the two types of data separately. Combined cohort data may arise through the use of complex study designs in which multiple groups are simultaneously sampled, as well as through the sharing of subject-level data from different studies (Tierney et al. 2015; Wolfson, Best, Addona, Wolfson, & Gadalla 2019). For example, Humbert et al. (2010) examined the survival of subjects with Pulmonary Arterial Hypertension (PAH) taken from the French PAH registry. From this registry, the subjects with idiopathic, familial or anoreixgen-associated PAH were classified either as incident cases (subjects who acquired PAH during the observation period) or prevalent cases (subjects who had acquired PAH prior to the observation period). Other examples of combined cohort data occur when independent studies are combined, as might be done in a meta-analysis (Abner et al. 2015). Analyses using combined cohort data may be found in a variety of fields including medicine/health, sports, public policy and finance (Daepp, Hamilton, West, & Bettencourt 2015; Groothuis & Hill 2011; Kingwell et al. 2012; Lee, Ning, Kryscio, & Shen 2019; Welch 1998).

Irrespective of whether the observed combined cohort data were collected from a single study or collected from various sources, a single survival analysis using all available data can have substantial benefits. For, when used alone: (i) Incident cohort studies may be of limited duration owing to logistical considerations and/or funding constraints. As a result, many failure times will be administratively censored at the end of the study rendering the classic Kaplan-Meier estimator undefined beyond this point (Kaplan & Meier 1958). Moreover, increasing the cohort size in a study of restricted duration does not solve this problem; (ii) On the other hand, increasing the size of the prevalent cohort in a prevalent cohort study with follow-up is likely to ameliorate this problem even with restricted follow-up (Wolfson et al. 2019). Yet, a modified form of the Kaplan-Meier estimator, called the Tsai-Jewell-Wang (TJW) estimator, used for pure prevalent cohort studies with follow-up, can produce absurd estimates of the survivor function. This happens when the risk set is underpopulated at observed failure times close to $t = 0$ (Pan & Chappell 1998; Wolfson et al.

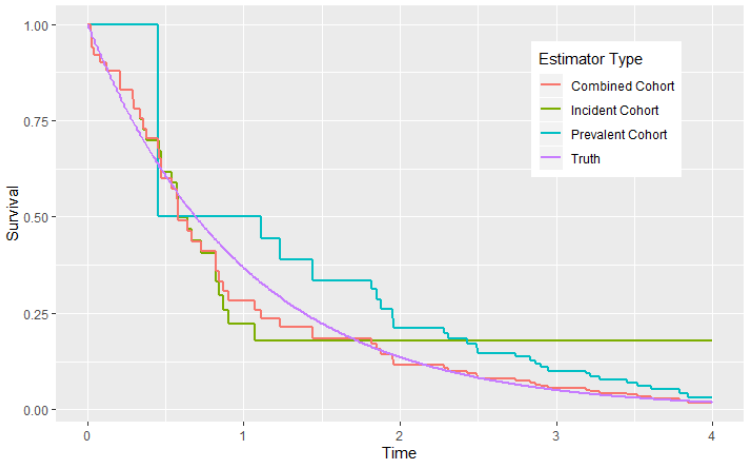

FIGURE 1 A comparison of the nonparametric product-limit estimates of the survival function for purely prevalent cohort data (blue), incident cohort data (green), and combined cohort data (orange) to the true underlying survival curve (purple). The underlying failure time distribution is Weibull with an Exponential censoring time distribution and a Gamma truncating distribution where each individual cohort comprise 50 observations. The TJW estimate derived from the prevalent cohort with follow-up has a substantial drop near the origin. The Kaplan-Meier estimate derived from an incident cohort, is undefined after the end of the study. The TJW estimate based on the combination of prevalent and incident cohorts ameliorates both of these issues.

2019). It is suggested through simulations, by Wolfson et al. (2019), that combining incident and prevalent cohort data simultaneously addresses both of these issues, through a data-based "fix". For a visual representation of these properties, see Figure 1. Further, use of all the available data will enhance the precision of the estimators; for example, the onset dates of a certain disease in a pure prevalent cohort study will follow-up, retrospectively obtained, can induce additional uncertainty (McVittie, Wolfson, & Stephens 2019; Zhong & Cook 2014).

Although intended for use with pure prevalent cohort data, the TJW estimator may be used with combined incident-prevalent cohort data by regarding the incident cohort as a prevalent cohort with truncation times set equal to zero. The wide availability of the TJW estimator in statistical software packages makes its use for combined data sets particularly attractive. However, the asymptotic properties of the TJW estimator originally relied on the assumed continuity of the truncation time distribution, which is not the case when some of them are allowed to be zero (Tsai, Jewell, & Wang 1987; Wang, Jewell, & Tsai 1986). Since then, various authors have analyzed the asymptotic properties of the TJW estimator under different assumptions on the truncating distribution and restrictions on the supports of the random variables involved in the estimation procedure, all within the context of a purely prevalent cohort (Gijbels & Wang 1993; Zhou 1996; Zhou & Yip 1999).

In this note, we show how the approach taken by Gijbels and Wang (1993), which allows for an arbitrary truncating distribution (not necessarily continuous), may be used to establish the asumptotic properties of the TJW (product-limit) estimator in the combined cohort setting. We conclude with a discussion on two applications, in density estimation and hypothesis testing problems.

2 | THE TSAI-JEWELL-WANG PRODUCT-LIMIT ESTIMATOR: DEFINITION AND ASYMPTOTIC PROPERTIES

Let $T_1, T_2, \ldots, T_n$ be i.i.d. failure times with distribution function $F(\cdot)$. We assume each T_i is potentially randomly right-censored by the random variable C_i which has distribution function $H(\cdot)$, for all $i = 1, 2, \ldots, n$. Our incident cohort consists of the independent pairs of observations $\{(X_i, \delta_i) = (\min(T_i, C_i), 1(T_i < C_i)) : i = 1, 2, \ldots, n\}$. Let $T_1, T_2, \ldots$ be a sequence of i.i.d. failure times with distribution function $F(\cdot)$ and let $A_1, A_2, \ldots$ be a sequence of i.i.d. truncation times independent of $T_1, T_2, \ldots$, with distribution function $G(\cdot)$. Our prevalent cohort consists of the observations for which $T_i > A_i$ where each failure time T_i is subject to right-censoring by the random variable C_i^* where, by assumption, $C_i^* > A_i$ and $C_i^* \sim H^*(\cdot)$. Thus, after follow-up, the prevalent cohort consists of the observation triples $\{(X_i, A_i, \delta_i) = (\min(T_i, C_i^*), A_i, 1(T_i < C_i^*)) : i = 1, 2, \ldots, m\}$. As discussed by Wolfson et al. (2019), incident cohort data may be embedded in $\mathbb{R}^3$ by setting the truncation times of the incident cohort failure/censoring times to 0. A combined cohort consisting of both incident and prevalent cohort data would then consist of the quadruples $\{(X_i, A_i, \delta_i, \gamma_i) : i = 1, 2, \ldots, n + m\}$ where γ_i is the cohort inclusion indicator function.

Let $\{(x_i, a_i, \delta_i, \gamma_i) : i = 1, 2, ..., n + m\}$ denote the observed quadruples of a combined incident and prevalent cohort, and let $t_{(1)}, t_{(2)}, ..., t_{(k)}$, denote the distinct ordered lifetimes of the uncensored x's. The TJW estimator of the survival function $S = 1 - F$ is given by:

$$\hat{S}_{n+m}(t) = \begin{cases} 1 & \text{if } t < \min\{x_i : \delta_i = 1\} \\ \prod_{j:t_{(j)}<t} \left(1 - \frac{\sum_i^{n+m} 1(t_{(j)}=t_i)}{\sum_i^{n+m} 1(a_i \le t_{(j)} \le x_i)}\right) & \text{otherwise} \end{cases} \quad (1)$$

Not only is $\hat{S}$ a product-limit estimator, it is also the non-parametric maximum likelihood estimator (NPMLE) of S (see (Wang 1991) for further details).

To establish the asymptotic properties of the TJW estimator, we note that the observed data may arise via two mechanisms. Either two studies (incident and prevalent cohorts, respectively) are combined or, data are collected from a single study in which the numbers of incident and prevalent cases are random. In the former case, we can embed the observed incident cases into the same space as the prevalent cases by considering their truncation times to be equal to 0. Thus, the combined cohort truncation times follow a mixture distribution $G^*(\cdot)$, where either $G^*(\cdot)$ has non-zero probability mass at 0 or follows the distribution $G(\cdot)$ for all positive truncation times. A similar scenario arises in the case in which the grand sample size is fixed but the numbers of incident and prevalent cases are random. Again, the combined cohort truncating distribution will follow the mixture distribution $G^*(\cdot)$.

The distribution of the truncation times in the combined cohort is necessarily of mixed type as there is a discontinuity at time $t = 0$. Therefore, the asymptotic properties of the TJW estimator must be established under this assumption. With this in mind, we consider the approach taken by Gijbels and Wang (1993). While the authors' proof requires that the supports of the failure time and truncation time distributions are bounded intervals, in practice, this restriction is hardly serious for combined incident and prevalent cohort data. For any distribution function K, denote the left and right endpoints of its non-zero support by

$$a_K = \inf\{t : K(t) > 0\} \text{ and } b_K = \inf\{t : K(t) = 1\}$$

Let the distribution function W be defined through the equality $1 - W(\cdot) = (1 - F(\cdot))(1 - H^\dagger(\cdot))$ where $H^\dagger(\cdot)$ is the distribution function of the combined cohort censoring random variables. Note that F is identifiable if $a_G \le a_W$ and $b_G \le b_W$ (Gijbels & Wang 1993). Gijbels and Wang establish an i.i.d. representation for the NPMLE $\hat{F}_{n+m}(\cdot)$ when $a_G < a_W$. Now, in most applied settings where combined incident and prevalent cohort data are available, this support assumption will hold since the incident cohort cases will contribute a non-zero probability mass at $t = 0$. Thus, we may assume $a_G = 0$. More precisely, in practice, it can be assumed that there exists $\epsilon > 0$ such that $F(x) = H^\dagger(x) = 0$ for all $0 \le x \le \epsilon$. Further, we may adapt the strong representation approach of Gijbels and Wang (1993) for a pure prevalent cohort with follow-up to a mixed cohort setting. Crucially, Gijbels and Wang place no continuity requirement on the truncation distribution, which they allow to be arbitrary. In mixed cohorts, the truncation distribution, $G(\cdot)$, is inevitably discontinuous. Therefore, there is a strong representation for the NPMLE of the failure time distribution function, $\hat{F}_{n+m}(\cdot)$, as given in Corollary 1 (d) of (Gijbels & Wang 1993). This representation opens the way to the following theorem:

Theorem 1. Let $a_G < a_W$ and $b < b_W$.

1. For $0 < x < b_W$, $\hat{F}_{n+m}(x) \to F(x)$ a.s.
2. $\sup_{0 \le x \le b} |\hat{F}_{n+m}(x) - F(x)| = O(((n + m)^{-1} \ln \ln n + m)^{\frac{1}{2}})$ a.s.
3. Let $D[0, b]$ be the space of all right-continuous functions with left limits on the interval $[0, b]$. The stochastic process $\sqrt{n+m}\left[\hat{F}_{n+m}(x) - F(x)\right]$ converges weakly on $D[0, b]$ to a mean zero Gaussian process with covariance function

$$\Gamma(x_1, x_2) = \int_a^{x_1 \wedge x_2} [\mathbb{P}(T \le x \le X | T \le X)]^{-2} \frac{\mathbb{P}(T \le x \le C^\dagger)}{\mathbb{P}(T \le X)} dF(x)$$

where $C^\dagger$ is the combined cohort censoring random variable.

3 | APPLICATIONS AND DISCUSSION

The NPMLE may also be applied to density estimation and hypothesis testing problems. For pure prevalent cohort data, Gijbels and Wang (1993) define the kernel density estimator:

$$\hat{f}_m(z) = \frac{1}{b_m^{-1}} \int K((z - t)/b_m) d\hat{F}_m(t).$$

Invoking the asymptotic properties of Theorem 1 above, it follows that $\hat{f}_{m+n}(z)$, defined analogously, in the mixed cohort setting, has matching asymptotic properties for estimating the probability density function f of the common survivor function.

Ning, Qin, and Shen (2010) discuss how the product-limit estimator may be used to test for distributional differences between two random variables when the observed samples are both length-biased (i.e. the onset process of the failure times is assumed to follow a stationary Poisson process) and right-censored. They argue that in the more general left-truncation setting, the TJW estimator may be used for such a test. In the combined cohort setting, as the testing procedure of Ning et al. is likelihood-based, their approach may be generalized to test whether data collected from an incident cohort and a prevalent cohort come from the same underlying distribution. This procedure would allow researchers to test whether survival is the same in two different cohort studies, one an incident cohort study and the other, a prevalent cohort study with follow-up. It could also be used to test for a change in survival before and after recruitment of the two cohorts.

A product-limit type non-parametric estimator of the survivor function has been proposed by Lai and Ying (1991) in the pure prevalent cohort setting. Importantly, they established the asymptotic properties of their estimator under much weaker conditions than those required for the proof of Theorem 1; they require independence of the failure times and of the truncaiton times. However, they do not require continuity of the truncation time distribution or of the failure time distribution. Moreover, they do not require the truncation times to be identically distributed. Nevertheless, the Lai and Ying estimator comes with a price for its implementation as it requires the specification of two tuning parameters which seem hard to determine in a systematic way. The simulation results presented by Wolfson et al. (2019) suggest that the TJW estimator performs at least as well as the Lai and Ying estimator in most cases and is clearly better in several. Importantly, the TJW estimator is included in most statistical software packages.

ACKNOWLEDGMENTS

The first author was supported by the Natural Sciences and Engineering Research Council of Canada PGSD-3 Award.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

Abner, E. L., Schmitt, F. A., Nelson, P. T., Lou, W., Wan, L., Gauriglia, R., … Kryscio, R. J. (2015). The statistical modeling of aging and risk of transition project: Data collection and harmonization across 11 longitudinal cohort studies of aging, cognition, and dementia. *Observational Studies*, *1*, 56-73.

Daepp, M. I. G., Hamilton, M., West, G., & Bettencourt, L. (2015). The mortality of companies. *Journal of the Royal Society Interface*, *12*.

Gijbels, I., & Wang, J. L. (1993). Strong representations of the survival function estimator for truncated and censored data with applications. *Journal of Multivariate Analysis*, *47*, 210-229.

Groothuis, P. A., & Hill, J. (2011). Pay discrimination, exit discrimination or both? Another look at an old issue using NBA data. *Journal of Sports Economics*, *14*, 171-185.

Humbert, M., Sitbon, O., Yaïci, A., Montani, D., O'Callaghan, D. S., Jaïs, X., … on behalf of the French Pulmonary Arterial Hypertension Network (2010). Survival in incident and prevalent cohorts of patients with pulmonary arterial hypertension. *European Respiratory Journal*, *36*, 549-555.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*, 457-481.

Kingwell, E., van der Kop, M., Zhao, Y., Shirani, A., Zhu, F., Oger, J., & Tremlett, H. (2012). Relative mortality and survival in multiple sclerosis: findings from british columbia, canada. *Journal of Neurology, Neurosurgery, and Psychiatry*, *83*, 61-66.

Lai, T. L., & Ying, Z. (1991). Estimating a distribution function with truncated and censored data. *The Annals of Statistics*, *19*, 417-442.

Lee, C., Ning, J., Kryscio, R., & Shen, Y. (2019). Analysis of combined incident and prevalent cohort data under a proportional mean residual life model. *Statistics in Medicine*, *38*, 2103-2114.

McVittie, J. H., Wolfson, D. B., & Stephens, D. A. (2019). Parametric modelling of prevalent cohort data with uncertainty in the measurement of the initial onset date. *Lifetime Data Analysis*. DOI: 10.1007/s10985-019-09481-1.

Ning, J., Qin, J., & Shen, Y. (2010). Nonparametric tests for right-censored data with biased sampling. *Journal of the Royal Statistical Society Series B Statistical Methodology*, *72*, 609-630.

Pan, W., & Chappell, R. (1998). A nonparametric estimator of survival functions for arbitrarily truncated and censored data. *Lifetime Data Analysis*, *4*, 187-202.

Tierney, J. F., Pignon, J. P., Gueffyier, F., Clarke, M., Askie, L., Vale, C. L., … On behalf of the Cochrane IPD Meta-analysis Methods Group (2015). How individual participant data meta-analyses have influenced trial design, conduct, and analysis. *Journal of Clinical Epidemiology*, *68*, 1325-35.

Tsai, W. Y., Jewell, N. P., & Wang, M. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, *74*, 883-886.

Wang, M. C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, *86*, 130-143.

Wang, M. C., Jewell, N. P., & Tsai, W. Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *The Annals of Statistics*, *14*, 1597-1605.

Welch, S. (1998). Nonparametric estimates of the duration of welfare spells. *Economics Letters*, *60*, 217-221.

Wolfson, D. B., Best, A. F., Addona, V., Wolfson, J., & Gadalla, S. M. (2019). Benefits of combining prevalent and incident cohorts: An application to myotonic dystrophy. *Statistical Methods in Medical Research*, *28*, 3333-3345.

Zhong, Y., & Cook, R. J. (2014). Measurement error for age of onset in prevalent cohort studies. *Applied Mathematics*, *5*, 1672-83.

Zhou, Y. (1996). A note on the TJW product-limit estimator for truncated and censored data. *Statistics and Probability Letters*, *26*, 381-387.

Zhou, Y., & Yip, P. S. F. (1999). A strong representation of the product-limit estimator for left truncated and right censored data. *Journal of Multivariate Analysis*, *69*, 261-80.