

A novel approach for analyzing data on recurrent events with duration to estimate the combined cumulative rate of both variables over time



Sudipta Bhattacharya

Biometrics and Information Sciences, 101 ORD, One MedImmune Way, Gaithersburg, 20878, United States

ARTICLE INFO

Keywords:

Recurrent events
Duration per event
Intensity
Nelson-aalen estimator

ABSTRACT

Recurrent adverse events, once occur often continue for some duration of time in clinical trials; and the number of events along with their durations is clinically considered as a measure of severity of a disease under study. While there are methods available for analyzing recurrent events or durations or for analyzing both side by side, no effort has been made so far to combine them and present as a single measure. However, this single-valued combined measure may help clinicians assess the wholesome effect of recurrence of incident comprising events and durations. Non-parametric approach is adapted here to develop an estimator for estimating the combined rate of both, the recurrence of events as well as the event-continuation, that is the duration per event. The proposed estimator produces a single numerical value, the interpretation and meaningfulness of which are discussed through the analysis of a real-life clinical dataset. The algebraic expression of variance is derived, asymptotic normality of the estimator is noted, and demonstration is provided on how the estimator can be used in the setup of testing of statistical hypothesis. Further possible development of the estimator is also noted, to adjust for the dependence of event occurrences on the history of the process generating recurrent events through covariates and for the case of dependent censoring.

1. Introduction

In clinical trials on diseases like Chronic Obstructive Pulmonary Disease (COPD), asthma, or migraine, etc., the event-durations are of interest along with the event-counts, as together they define severity of the disease.

Poisson regression or Negative-Binomial regression described by Lawless [1] for analyzing data on recurrent events when covariates are considered not time dependent; or for time dependent covariates, estimating the mean or rate function of recurrent events, e.g., method introduced by Lin et al. [2,3] and by Miloslavsky et al. [4] (all the three based on the definition of intensity function introduced by Andersen-Gill [5]) are the standard approaches. Otherwise, if event occurrence is considered dependent on previous events, then stratified Andersen-Gill model (Cook and Lawless [6], pp 175–176) can be used. In addition, non-parametric Nelson Aalen estimator ([7]) for the rate or mean function of recurrent events, and the extensions by Cook et al. [8]) for event dependent censoring and termination are commonly used methods for analyzing data on recurrent events.

For the analysis of waiting times (with assumption of independence among waiting times and deviating from that assumption), detailed discussion is provided in chapter 4 of Cook and Lawless [6], pp 121–160. Otherwise, the modeling of proportional hazard ratio using

stratified Cox-type models ([9]) based on total time as well as on gap times introduced by Prentice, William and Peterson [10] and marginal Cox-models based on total time introduced by Wei, Lin and Weisfeld [11] are used as well.

On methods for analyzing data on duration, Metcalfe et al. [12] made a thorough coverage in their article. Otherwise, X. Joan Hu et al. [13] also proposed some methods for analyzing event-duration. The bivariate approach to deal with recurrent events with duration is through an alternating two-state process ('exacerbation state' and 'exacerbation-free state' being the two alternative states) as described by Cook and Lawless ([6] section 6.5, pp 216–218 and section 6.7.2, pp 232–236).

However, none of the methods mentioned above present an estimate for combined cumulative rate or mean of recurrent events and duration of events over time.

Here in this paper, a non-parametric estimator is proposed that takes the totality of the data into account through dealing with both, the recurrence of events and the duration of them simultaneously; and as a result, produces a single numerical value, which estimates the wholesome effect of the incident. Consequently, the proposed estimator can be looked upon as a joint or combined rate of both, the event recurrences as well as the duration per event.

Following is how the concept of the proposed estimator is developed

E-mail address: sudipta.bhattacharya@astrazeneca.com.

<https://doi.org/10.1016/j.conctc.2018.03.002>

Received 20 October 2017; Received in revised form 23 February 2018; Accepted 8 March 2018

Available online 12 March 2018

2451-8654/ © 2018 The Author. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

over the next few sections. In section 2, the mathematical motivation, development and properties of the estimator are described. In section 3, the interpretation, usefulness and meaningfulness of the single value, produced by the estimator based on a real-life clinical dataset are discussed; and in section 4, the possible use and advantages of the proposed estimator are described and further potential developments are mentioned.

2. Mathematical development

The mathematical framework for this paper is built on a process that generates recurrence of events to individual subjects, who constitute a population; and the assumption regarding occurrence of events here is that, an event can occur and end both at a time instant t (like as it happens in case of any Poisson process), or can occur at one time instant and then continues for some time before it ends.

2.1. Definitions

Let us denote by X_t the number of events that are observed to have occurred to an individual subject by (i.e., on or before) time t .

The intensity function is defined (Cook and Lawless [6], chapter 1.3, p 10) as

$$\chi(t|\mathcal{H}_t) = \lim_{\Delta t \downarrow 0} \frac{\Pr(\Delta X_t = 1 | \mathcal{H}_t)}{\Delta t} \quad (1)$$

where $\Delta X_t = X_{t+\Delta t} - X_t$ and $\mathcal{H}_t = \{X(s), 0 \leq s < t\}$ is the history of a process.

Note that the intensity function can also be looked upon as $\chi(t|\mathcal{H}_t)dt = P(dX_t = 1 | \mathcal{H}_t) = E\{dX_t | \mathcal{H}_t\}$.

Based on the definition of intensity function for the occurrence (or, onset, to be precise) of a new event presented above, let us define the intensity function for ending of events for an event that has already occurred (and started) to an individual subject at time-point $\tilde{t}_0 < t$ (i.e., the onset of the event was at time-point $\tilde{t}_0 < t$) and is not continuing until (i.e., has ended by) time t as:

$$\chi^Z(t|\mathcal{H}_t) = \lim_{\Delta t \downarrow 0} \frac{\Pr(-\Delta Z_t = 1 | \mathcal{H}_t)}{\Delta t} \quad (2)$$

where $\mathcal{H}_t = \{X(s), 0 \leq s < t\}$ is the history of a process and $\Delta Z_t = Z_{t+\Delta t} - Z_t$, with Z_t denoting an indicator function such that,

$Z_t = 1$, when an event that has already occurred (and started) at time-point $\tilde{t}_0 < t$ (i.e., the onset of the event was at time-point $\tilde{t}_0 < t$), continues until time t , or,

$Z_t = 0$, when an event that has already occurred (and started) at time-point $\tilde{t}_0 < t$ (i.e., the onset of the event was at time-point $\tilde{t}_0 < t$), has also ended by time t .

Note that the intensity function can also be looked upon as $\chi^Z(t|\mathcal{H}_t)dt = \Pr(-\Delta Z_t = 1 | \mathcal{H}_t) = E\{-dZ_t | \mathcal{H}_t\}$.

2.2. Mathematical motivation

Let us now define the following variables:

N_t = total count of the onset of events occurred to the population of n subjects by time t , which is non-decreasing over time, and N_t^Z = count of events that have already occurred to the population of n subjects before time t and are continuing until time t .

Clearly, $N_t \geq N_t^Z$ at any given time t .

If we define a new variable N_t^S as:

$$N_t^S = N_t + N_t^Z, \text{ then } \Delta N_t^S = \Delta N_t + \Delta N_t^Z.$$

Then, $\Delta N_t = \Delta N(t) = \sum_{k=1}^n [X^{(k)}(t + \Delta t) - X^{(k)}(t)] Y^{(k)}(t) C^{(k)}(t)$,

where $X^{(k)}(t)$ is an indicator function such that:

$X^{(k)}(t) = 1$ when a new event has occurred (in terms of onset of that new event) by time t to the k^{th} individual subject, or

$X^{(k)}(t) = 0$, when a new event has not occurred (in terms of onset of that new event) until time t since the preceding event occurred and ended to the k^{th} individual subject;

$Y^{(k)}(t)$ is an indicator function with following such that:

$Y^{(k)}(t) = 1$, when the k^{th} subject belongs to the risk set at time t for having a new event, since the preceding event occurred and ended, $Y^{(k)}(t) = 0$, when the k^{th} subject does not belong to the risk set at time t for having a new event, since the preceding event has occurred and is continuing;

and $C^{(k)}(t)$ is an indicator function with following such that:

$C^{(k)}(t) = I(t \leq C^{(k)})$ is an indicator function of whether the k^{th} subject is under observation at time t .

Clearly, $C^{(k)}(t)$ is the indicator for censoring of a subject and here we assume data to be missing at random after censoring.

We also consider $\Delta N_t^Z = \sum_{k=1}^n [1 - \{Z^{(k)}(t) - Z^{(k)}(t + \Delta t)\}] Z^{(k)}(t) C^{(k)}(t)$, where $Z^{(k)}(t)$ is an indicator function such that:

$Z^{(k)}(t) = 1$ if an event that has already occurred (and started) to the k^{th} subject and is continuing until time t ;

$Z^{(k)}(t) = 0$ if an event that has already occurred (and started) to the k^{th} subject and has also ended by time t .

It should be noted that since $Y^{(k)}(t) + Z^{(k)}(t) = 1$ at any given time t for any k ,

$$\begin{aligned} \sum_{k=1}^n Y^{(k)}(t) C^{(k)}(t) + \sum_{k=1}^n Z^{(k)}(t) C^{(k)}(t) &= \sum_{k=1}^n \{Y^{(k)}(t) + Z^{(k)}(t)\} C^{(k)}(t) \\ &= \sum_{k=1}^n C^{(k)}(t). \end{aligned} \quad (3)$$

2.3. Development of the estimator

It is already defined that $N_t^S = N_t + N_t^Z$, implies $\Delta N_t^S = \Delta N_t + \Delta N_t^Z$, where $\Delta N_t = N_{t+\Delta t} - N_t$.

Since by the total probability theorem, $P(a \cup b) = P(\varepsilon)P(a \cup b | \varepsilon) + P(\varepsilon^C)P(a \cup b | \varepsilon^C) = P(\varepsilon)P(a | \varepsilon) + P(\varepsilon^C)P(b | \varepsilon^C)$, in case a and b are disjoint and $P(b | \varepsilon) = 0$ and $P(a | \varepsilon^C) = 0$, Where

ε : risk set of subjects at time t for having a new event

ε^C : set of (subjects with) existing events (i.e., events that did not end by time t) that continuing until time t

a : occurrence of a new event to a subject within the interval of $[t, t + \Delta t)$

b : an existing event continuing during the interval of $[t, t + \Delta t)$

Which implies that $\frac{\Delta N_t^S}{\sum_{k=1}^n C^{(k)}(t)} = \frac{\Delta N_t + \Delta N_t^Z}{\sum_{k=1}^n C^{(k)}(t)} = \frac{\Delta N_t}{\sum_{k=1}^n C^{(k)}(t)} + \frac{\Delta N_t^Z}{\sum_{k=1}^n C^{(k)}(t)}$

$$\begin{aligned} &= \left[\frac{\sum_{k=1}^n Y^{(k)}(t) C^{(k)}(t)}{\sum_{k=1}^n C^{(k)}(t)} \right] \times \frac{\Delta N_t}{\sum_{k=1}^n Y^{(k)}(t) C^{(k)}(t)} \\ &+ \left[\frac{\sum_{k=1}^n Z^{(k)}(t) C^{(k)}(t)}{\sum_{k=1}^n C^{(k)}(t)} \right] \times \frac{\Delta N_t^Z}{\sum_{k=1}^n Z^{(k)}(t) C^{(k)}(t)}, \end{aligned}$$

where $\sum_{k=1}^n C^{(k)}(t) = \sum_{k=1}^n Y^{(k)}(t) C^{(k)}(t) + \sum_{k=1}^n Z^{(k)}(t) C^{(k)}(t)$, from equation (3).

Integrating over time, $\int_0^t \frac{dN_u^S}{\sum_{k=1}^n C^{(k)}(u)} = \int_0^t \frac{dN_u + dN_u^Z}{\sum_{k=1}^n C^{(k)}(u)} = \int_0^t \frac{dN_u}{\sum_{k=1}^n C^{(k)}(u)} + \int_0^t \frac{dN_u^Z}{\sum_{k=1}^n C^{(k)}(u)}$

$$= \int_0^t \left[\frac{\sum_{k=1}^n Y^{(k)}(u)C^{(k)}(u)}{\sum_{k=1}^n C^{(k)}(u)} \right] \times \frac{dN_u^S}{\sum_{k=1}^n \sum_{k=1}^n Y^{(k)}(u)C^{(k)}(u)} + \int_0^t \left[\frac{\sum_{k=1}^n Z^{(k)}(u)C^{(k)}(u)}{\sum_{k=1}^n C^{(k)}(u)} \right] \times \frac{dN_u^Z}{\sum_{k=1}^n Z^{(k)}(u)C^{(k)}(u)}.$$

Then $\int_0^t \frac{dN_u^S}{\sum_{k=1}^n C^{(k)}(u)} = \int_0^t p_u \chi(u|\mathcal{H}_u^n) du + \int_0^t p_u^Z \{1 - \chi^Z(u|\mathcal{H}_u^n)\} du$ may be looked upon as the estimator of the cumulative weighted event-time recurrence rate, where

$$p_u = \left[\frac{\sum_{k=1}^n \{Y^{(k)}(u)C^{(k)}(u)\}}{\sum_{k=1}^n C^{(k)}(t)} \right],$$

$$\int_0^t \chi(u|\mathcal{H}_u^n) du = \int_0^t \left[\frac{dN_u}{\sum_{k=1}^n \{Y^{(k)}(u)C^{(k)}(u)\}} \right],$$

$$p_u^Z = \left[\frac{\sum_{k=1}^n Z^{(k)}(u)C^{(k)}(u)}{\sum_{k=1}^n C^{(k)}(u)} \right],$$

$$\text{and } \int_0^t \{1 - \chi^Z(u|\mathcal{H}_u^n)\} du = \int_0^t \left[\frac{dN_u^Z}{\sum_{k=1}^n Z^{(k)}(u)C^{(k)}(u)} \right].$$

A detailed derivation of these individual components of the estimator is provided in the [Appendix](#) section of this paper.

2.4. Properties

2.4.1. Asymptotic property and variance

Due to the very way of the mathematical development here, the estimator of the cumulative weighted event-time recurrence rate defined above should retain the asymptotic normality property held by the Nelson-Aalen estimator; in addition, the derivation for the estimate of variance for Nelson-Aalen estimator, as derived in Cook and Lawless ([6] chapter 3.4.1, pp 68–69), can be adapted to derive the variance estimate of the proposed estimator in the following way.

$$\begin{aligned} \widehat{\text{Var}} \left(\int_0^t p_u \chi(u|\mathcal{H}_u^n) du + \int_0^t p_u^Z \{1 - \chi^Z(u|\mathcal{H}_u^n)\} du \right) &= \text{Var} \left(\int_0^t \frac{dN_u^S}{\sum_{k=1}^n C^{(k)}(u)} \right) \\ &= \text{Var} \left(\int_0^t \frac{dN_u}{\sum_{k=1}^n C^{(k)}(u)} + \int_0^t \frac{dN_u^Z}{\sum_{k=1}^n C^{(k)}(u)} \right) = \text{Var} \left(\int_0^t \frac{dN_u}{\sum_{k=1}^n C^{(k)}(u)} \right) + \text{Var} \left(\int_0^t \frac{dN_u^Z}{\sum_{k=1}^n C^{(k)}(u)} \right), \\ \text{since } \text{Cov}(\Delta N_t, \Delta N_t^Z) &= 0; \\ \Rightarrow \widehat{\text{Var}} \left(\int_0^t p_u \chi(u|\mathcal{H}_u^n) du + \int_0^t p_u^Z \{1 - \chi^Z(u|\mathcal{H}_u^n)\} du \right) &= \sum_{h:t(h) \leq t} \frac{\Delta N_{t(h)}^S}{\left\{ \sum_{k=1}^n C^{(k)}(t(h)) \right\}^2} \\ &+ \sum_{h:t(h) \leq t} \frac{\Delta N_{t(h)}^Z}{\left\{ \sum_{k=1}^n C^{(k)}(t(h)) \right\}^2} = \sum_{h:t(h) \leq t} \frac{\Delta N_{t(h)}^S}{\left\{ \sum_{k=1}^n C^{(k)}(t(h)) \right\}^2}, \text{ since } \Delta N_t^S = \Delta N_t + \Delta N_t^Z. \end{aligned}$$

Hence, $\widehat{\text{Var}} \left(\int_0^t p_u \chi(u|\mathcal{H}_u^n) du + \int_0^t p_u^Z \{1 - \chi^Z(u|\mathcal{H}_u^n)\} du \right) = \sum_{h:t(h) \leq t} \frac{\Delta N_{t(h)}^S}{\left\{ \sum_{k=1}^n C^{(k)}(t(h)) \right\}^2}$
 The practical utility of the proposed estimators' asymptotic normality property is that, in case of large samples it can be used for the testing of the statistical hypothesis regarding comparison of treatment effects on the cumulative weighted event-time recurrence rate.

2.4.2. Poisson properties

In case of a Poisson process or a renewal process without duration, that is, when the events occur and ends at the same time instant t , the estimator of the cumulative weighted event-time recurrence rate boils down to the regular Nelson-Aalen estimator for the cumulative rate (or mean in case of Poisson process, as discussed in details in Cook and Lawless ([6] p 68) of event occurrences (equivalently known as the cumulative intensity function or the integrated hazard function over time).

2.4.3. Equivalence with mean total duration

Under the assumption of no drop outs, $\sum_{k=1}^n C^{(k)}(t) = n$, $0 \leq t \leq \tilde{T}$, where n is the total number of subjects in the population and \tilde{T} is the total time all these n subjects were under study, the estimator for the

$$\text{cumulative weighted event-time recurrence rate, } \int_0^t \left[\frac{dN_u^S}{\sum_{k=1}^n C^{(k)}(u)} \right] = \sum_{h:t(h) \leq \tilde{T}} \frac{\Delta N_{t(h)}^S}{\sum_{k=1}^n C^{(k)}(t(h))} = \sum_{h:t(h) \leq \tilde{T}} \frac{\Delta N_{t(h)}^S + \Delta N_{t(h)}^Z}{\sum_{k=1}^n C^{(k)}(t(h))} = \frac{1}{n} \sum_{k=1}^n \sum_{i=1}^{\left\lfloor \frac{m^k}{2} \right\rfloor} d_{t_{(2i-1)}^k - t_{2i}^k};$$

where $\{d_{t_{(2i-1)}^k - t_{2i}^k}^k\}$, $i = 1(1)\left\lfloor \frac{m^k}{2} \right\rfloor$, $0 \leq t_1^k \leq \dots \leq t_{m^k-1}^k \leq t_{m^k}^k \leq \tilde{T}$ are the successive duration-times for k^{th} subject (i.e., the time-ordered sequence of time-intervals presenting duration of recurrent events occurred to k^{th} subject) over a given time-period $(0, \tilde{T})$ and $t_1^k \leq \dots \leq t_{m^k}^k$ are the time-points when recurrent events occurred and then ended (i.e., the time-ordered sequence of onset-time-point and the end-time-point of the recurrent events, which occurred) to the k^{th} subject.

Note that the right-hand side of the algebraic expression above is nothing but the mean of the total duration-time, i.e., the average of the sum of time-intervals, or in other words, the average of the sum of durations of recurrent events occurred to subjects during the study.

Consequently, if the event occurrence (or onset, to be precise) rate and duration distribution are stationary over time, then by construction, the cumulative weighted event-time recurrence rate is essentially the same as the mean of the total duration-time (per some unit time). That is, the estimator proposed here is equivalent to $E[D_T | T = t]$ for large t , where D_T is a random variable denoting the sum of time-intervals or durations of recurrent events occurred to a subject over a given time-period (i.e., $D_t = \sum_{i=1}^{\left\lfloor \frac{m}{2} \right\rfloor} d_{t_{(2i-1)} - t_{2i}}$, where $\{d_{t_{(2i-1)} - t_{2i}}\}$, $i = 1(1)\left\lfloor \frac{m}{2} \right\rfloor$, $0 \leq t_1 \leq \dots \leq t_{m-1} \leq t_m \leq t$ are the random variables denoting the successive time-intervals presenting duration of recurrent events occurred to a subject over a given time-period $(0, t)$, and $t_1 \leq \dots \leq t_m$ are the time-points when recurrent events occurred and then ended (i.e., the time-ordered sequence of onset-time-point and the end-time-point of the recurrent events, which occurred) to that subject. However, since the event occurrence (or onset, to be precise) rate and the distribution of event durations both might vary with time, the estimates from the proposed estimator might only be comparable to the numerical values of D_t as t increases.

3. Data analysis

Since the non-parametric estimator introduced here is developed to estimate a novel parameter for assessing the disease condition in a population through a single value combining recurrence of events and duration over time, which has not been dealt with by any other existing and/or standard methods so far, the purpose of the data analysis here is to understand the purpose of such an estimator in the context of data on recurrent events with duration, through the meaning, usefulness and interpretation of the numerical value the estimator produces, instead of comparing the new estimator to any existing estimator of the event-time recurrence rate, which in fact does not exist until now.

It should be noted that if an event to a subject ends at any time t , then another event to that subject can begin immediately afterwards, that is, at time $t + \Delta t$, which will make the previous event appear to be continuing for that subject through the time $t + \Delta t$. Hence when day is the time-unit in any data set used for estimating the cumulative weighted event-time recurrence rate, then if an event to a subject ends on a particular day, then continuation of event will be considered for that subject until the day before.

An exacerbation dataset from one of the historical COPD studies is used for the data analysis and estimation of the cumulative weighted event-time recurrence rate with COPD exacerbations being the recurrent events. The Nelson-Aalen estimator of the event rate and the descriptive means for total-duration and also for the duration of the disease (in days) have been calculated on that dataset, so that the results can be compared side by side, in order to understand the meaning and usefulness of the estimator introduced here.

The dataset, thus considered, has a total of 165 patients, of which

Table 1
Comparison of estimators for event rate and for duration of events with the proposed estimator.

Treatment	N	Cumulative Weighted Event-Time Recurrence Rate (SE)	Mean (SD) of total duration-time	Mean Duration (SD)	Nelson-Aalen Event Rate (SE)	Product of N.A. Event Rate and Mean Duration
A	117	9.0652 (0.331)	9.730 (5.579)	7.5608 (3.635)	1.4383 (0.126)	10.8747
B	48	10.1556 (0.759)	9.766 (6.273)	7.650 (4.100)	1.4315 (0.185)	10.9510
Treatment comparison (p-value)	–	0.4879	0.9060			

117 were under an active drug under study (drug A, say) and the rest of the 48 patients were under an active comparator (drug B, say) for one year, which was the study-period. All the patients under both the treatment arms had at least one occurrence of exacerbation to them. 28 patients under drug A and 11 patients under drug B had more than one exacerbation during the one-year of on-treatment study period. The maximum number of exacerbations to a patient under drug A was three and the maximum number of exacerbations to a patient under drug B was four. The minimum duration of COPD exacerbation was one day under drug A (maximum 34 days) but was two days under drug B (maximum 31 days). Two patients under drug A and one patient under drug B were observed with an event but without any respective duration-time recorded in the dataset. 38 patients under drug A and 20 under drug B had dropped out of the study early.

Table 1 shows the results of the different estimators: cumulative weighted event-time recurrence rate introduced here, the mean duration and the Nelson-Aalen event rate.

3.1. Meaningfulness

As seen in Table 1, the numerical comparison of the estimated cumulative weighted event-time recurrence rate and the product of the Nelson-Aalen estimate of the event rate and the mean duration indicates that the new estimator may be looked upon as a combined rate (over time) of event recurrence and duration. This is because the product of Nelson-Aalen estimate and the mean duration (average event-day of 10.8747 for drug A and 10.9510 for drug B) can be looked upon as a single numerical value giving an idea about the totality of the data through dealing with both: the intensity of the disease through the Nelson-Aalen estimate of the recurrent event rate and the average duration of events. Likewise, the estimated cumulative weighted event-time recurrence rate (9.0652 event-days for drug A and 10.1556 event-days for drug B, both in 1 patient-year) gives an idea about the combined cumulative weighted rate (within a year) of the number of event occurrences (intensity of the disease) along with duration of time those events continued, presented as a single numerical value.

3.2. Usefulness

Because of the asymptotic normality property of the estimator mentioned in section 2.4.1, in case of large samples it can be used as a statistic for the *t*-test in testing a statistical hypothesis regarding comparison of treatment effect on the cumulative weighted event-time recurrence rate, as presented in Table 1. Since neither the p-value for proposed estimator nor the p-value for N.A. estimator is significant, it's not meaningful to compare between them. However, the ability of performing hypothesis testing using the proposed estimator signifies the purposefulness of the estimator over both: The Nelson-Aalen estimator, which does not take duration into account, and the simple product of Nelson-Aalen event rate and mean duration, which cannot be used as a statistic for the *t*-test in testing a statistical hypothesis regarding comparison of treatment effect.

Fig. 1 presents the over-time graph comparing the progression of the cumulative weighted event-time recurrence rate and of the Nelson-Aalen estimate over one year from the COPD dataset described above. As observed in the graph (Fig. 1), the different in treatment effects between drugs A and B are more clearly visible in the curves for the

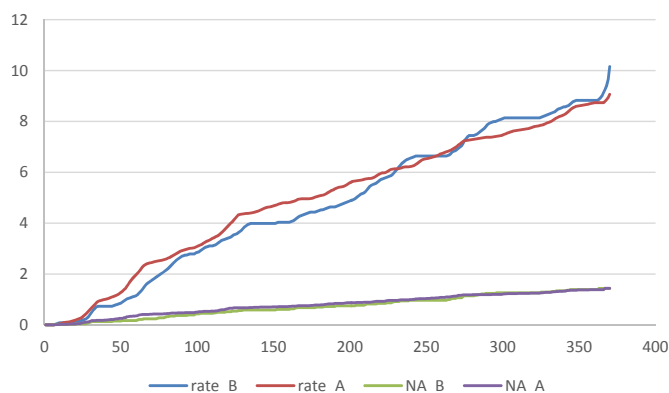


Fig. 1. Comparison of the proposed combined rate and N.A. rate over time.

cumulative weighted event-time recurrence rate than in the curves for Nelson-Aalen estimates for the cumulative intensity function of event recurrence.

The estimates of the cumulative weighted event-time recurrence rate in Table 1 (9.0652 event-days for treatment A and 10.1556 event-days for treatment B both in 1 patient-year) are numerically comparable to the means of total duration-time (9.73 for treatment A and 9.766 for treatment B) in the one-year time-period, as it is discussed in section 2.4.3.

3.3. Interpretability

When covariance of two variables is considered, it is viewed as a single numerical value, expressing the joint behavior of the two variables. Likewise, the proposed rate should be viewed as a linear combination of progression of event-recurrence and duration over time.

If the cumulative weighted rate of 9.0652 event-days for drug A of the event-time recurrences over 1 patient-year (10.1556 event-days in 1 patient-year for drug B) is viewed as a combined, single-valued rate of event-recurrence and duration, then the proposed estimator can be used to understand the wholesome effect of the recurrence of COPD exacerbations in these subjects under the two different treatments. In other words, this combined, single-valued cumulative weighted rate of event-time recurrences may be used as a measure to assess the severity of the disease involving recurrence of events (intensity of the disease) and the duration of events. This is because, the proposed estimator takes the totality of data on disease causing recurrent events with duration into account and produces a single value containing all information (recurrence of events, duration of events and total observation-time) about that data.

4. Discussion

The estimator of the cumulative weighted event-time recurrence rate introduced in this paper accounts for both, the recurrence of events as well as the duration per event over time. In case of the recurrent adverse events with duration that occur in clinical studies, the occurrence of events are dependent on whether or not the preceding event is still continuing; and hence, the issue of event continuation should be

taken into account while considering the risk set(s) over time for the occurrence of the following event(s). Moreover, in a process generating recurrent events where every event, once occurred, may continue for some duration of time, events and the associated time per event should be considered inseparable in defining such a process generating recurrent events, as done in section 2.2 of this paper. Naturally, one way of analyzing the data of recurrent events with durations is by dealing with them simultaneously, by introducing an extra parameter in the method of estimation, which has been done in the estimator introduced, as shown in section 2.3 of this paper. The proposed estimator, named as the cumulative weighted event-time recurrence rate in section 2.3 of this paper, is discussed in section 3 as a tool for understanding the wholesome effect of the recurrence of events with duration, e.g., the severity of a disease observed in a real-life data-set, as it gives an idea about the totality of the data regarding the disease causing recurrent events with duration through the estimation of the combined rate (over a given period of time) of event recurrences (intensity of the disease) along with duration of time those events continued, presented as a single numerical value.

The proposed estimator is very easy to calculate and also easy to interpret. In fact, the estimator can be interpreted in any of the following two ways. The estimator can be looked upon as an estimate of the combined rate of both: the recurrence of events and of the duration of events calculated over a given period of time. Or, under certain conditions (e.g., the event occurrence (onset) rate and duration distribution being stationary over time and/or when the time of study-period is very large, as described in section 2.4.3) it can be viewed as an estimate of the mean of the total duration-time (or, in other words, the average of the sum of the time-intervals of duration of recurrent events occurred to a subject during the study).

Appendix

Since $\Delta N_i = \sum_{k=1}^n [X^{(k)}(t + \Delta t) - X^{(k)}(t)]Y^{(k)}(t)C^{(k)}(t) = \sum_{k=1}^n \Delta X^{(k)}(t)Y^{(k)}(t)C^{(k)}(t)$ denotes the number of events occurred (i.e., the onset of the events that are observed) to the population of n subjects within the time-interval of $[t, t + \Delta t]$, then, denoting the history of the process generating recurrent events for the population of n subjects by time t by \mathcal{H}_t^n ,

$$E[\Delta N_i | \mathcal{H}_t^n] = E\left[\sum_{k=1}^n \Delta X^{(k)}(t)Y^{(k)}(t)C^{(k)}(t) \middle| \mathcal{H}_t^n\right] = \sum_{k=1}^n Y^{(k)}(t)C^{(k)}(t)E[\Delta X^{(k)}(t) | \mathcal{H}_t^{(k)}] = \sum_{k=1}^n Y^{(k)}(t)C^{(k)}(t)Pr(\Delta X^{(k)}(t) = 1 | \mathcal{H}_t^{(k)})$$

$$= \sum_{k=1}^n [Y^{(k)}(t)C^{(k)}(t)\chi^{(k)}(t|\mathcal{H}_t^{(k)})\Delta t].$$

The above equation is derived by following equation (1), since for any k :

$$Pr(\Delta X^{(k)}(t) = 1 | \mathcal{H}_t^{(k)}) = \chi^{(k)}(t|\mathcal{H}_t^{(k)})\Delta t,$$

where $\chi^{(k)}(t|\mathcal{H}_t^{(k)})$ is the intensity function for the occurrence (onset, to be precise) of an event to the k^{th} subject at time t given $\mathcal{H}_t^{(k)}$, which is denoting the history of the process to the k^{th} subject by time t .

However, under the assumption that for all k , $\chi^{(k)}(t|\mathcal{H}_t^{(k)}) = \chi(t|\mathcal{H}_t^n)$, which denotes the population intensity function for event recurrence to anyone in the population of n subjects at any time t ,

$$E[\Delta N_i | \mathcal{H}_t^n] = \chi(t|\mathcal{H}_t^n) \left[\sum_{k=1}^n \{Y^{(k)}(t)C^{(k)}(t)\} \right] \Delta t, \text{ which implies}$$

$$E\left[\left\{\Delta N_i - \chi(t) \left[\sum_{k=1}^n \{Y^{(k)}(t)C^{(k)}(t)\} \right] \Delta t \right\} \middle| \mathcal{H}_t^n\right] = 0.$$

Considering $M(t) = N_t - \int_0^t \chi(u|\mathcal{H}_u^n) \left[\sum_{k=1}^n \{Y^{(k)}(u)C^{(k)}(u)\} \right] du$, it can be written that $E[dM(t)] = 0$, indicating the martingale property for $M(t)$, as explained in Andersen, Borgen, Gill and Kleiding [14].

Now, because $\{\sum_{k=1}^n Y^{(k)}(t)C^{(k)}(t)\}$ and $\sum_{k=1}^n C^{(k)}(t)$ both are predictable at any time t , the integrals can be written in the following form:

$$\int_0^t \left[\frac{dN_u}{\sum_{k=1}^n C^{(k)}(u)} \right] = \int_0^t \left[\frac{\chi(u|\mathcal{H}_u^n) \left[\sum_{k=1}^n \{Y^{(k)}(u)C^{(k)}(u)\} \right] du}{\sum_{k=1}^n C^{(k)}(u)} \right] + \int_0^t \left[\frac{dM_u}{\sum_{k=1}^n C^{(k)}(u)} \right]$$

Or equivalently,

$$\int_0^t \left[\frac{dN_u}{\sum_{k=1}^n C^{(k)}(u)} \right] = \int_0^t \chi(u|\mathcal{H}_u^n) \left[\frac{\sum_{k=1}^n \{Y^{(k)}(u)C^{(k)}(u)\}}{\sum_{k=1}^n C^{(k)}(u)} \right] du + \int_0^t \left[\frac{dM_u}{\sum_{k=1}^n C^{(k)}(u)} \right]$$

While the alternating two-stage model described in Cook and Lawless ([6] section 6.5, pp 216–218 and section 6.7.2, pp 232–236) is a very appropriate approach to analyze data on recurrent events with duration, the method uses many assumptions (e.g. distribution for the frailty type parameter representing the conditional transitional intensities between states, assumption that the hazard function be dependent on the preceding events i.e., on the process generating recurrent events only through the time-dependent covariates, etc.) that are sometimes hard to justify. Whereas the estimator proposed in this paper uses minimum amount of assumption due to its developmental background being in the non-parametric methodology.

The estimator developed in this paper may be extended to fit into different conditions (namely, Poisson or even general point process in the sense that jumps are greater than of unit one, time-dependent covariates and dependent censoring) by adapting various methods for estimating the mean or the rate function of recurrent events, e.g., method introduced by Lin et al. [2], by Lin et al. [3], and by Miloslavsky et al. [4] i.e., methods wherever increments in the number of events at any time and the risk sets for having the following events over time are considered in the estimation procedure. This is a work left for future research.

Acknowledgements

The author is deeply thankful to Dr. Professor Jerry F. Lawless from University of Waterloo, Canada for all his critical inputs, without which this work of the author would never have been complete. The author sincerely thanks AstraZeneca Respiratory statisticians and the Information sharing group for their kind support by providing the author with a relevant dataset.

(following the same mathematical theory as that used in defining the Nelson Aalen estimator ([7]).

Which implies that $\int_0^t p_u \chi(u|\mathcal{H}_u^n) du = \int_0^t \left[\frac{dN_u}{\sum_{k=1}^n C^{(k)}(u)} \right]$, where $p_u = \left[\frac{\sum_{k=1}^n \{Y^{(k)}(u)C^{(k)}(u)\}}{\sum_{k=1}^n C^{(k)}(u)} \right]$

(also, naturally, $\int_0^t \chi(u|\mathcal{H}_u^n) du = \int_0^t \left[\frac{dN_u}{\sum_{k=1}^n \{Y^{(k)}(u)C^{(k)}(u)\}} \right]$).

For the estimation of the rate of the continuing events, what is note-worthy is that since the recurrence of events along with their duration of sustenance is of interest here, the consideration of the ending time of events may not be relevant in the context. This is because, if an event ends at any time t , by definition of instantaneous intensity function for event ending, that event is no longer visible at time $t + \Delta t$. Whereas, events that occur at time t and some or all of which continue beyond the time t (i.e., do not also end at time t) or other events that have occurred before time t and continue till time t (i.e., end at some point of time beyond t) are all what are visible at time $t + \Delta t$.

Consequently, as the objective here is to find the rate of such continuing cases of events, the following is derived.

Since $\Delta N_t^Z = \sum_{k=1}^n [1 - \{Z^{(k)}(t) - Z^{(k)}(t + \Delta t)\}]Z^{(k)}(t)C^{(k)}(t) = \sum_{k=1}^n \{(1 - (-\Delta Z^{(k)}(t)))Z^{(k)}(t)C^{(k)}(t)\}$ denotes the number of already occurred (i.e., existing) events continuing during the interval of $[t, t + \Delta t]$, then, denoting the history of the process generating recurrent events for the population of n subjects by time t by \mathcal{H}_t^n ,

$$E[\Delta N_t^Z | \mathcal{H}_t^n] = E \left[\sum_{k=1}^n \{(1 - (-\Delta Z^{(k)}(t)))Z^{(k)}(t)C^{(k)}(t)\} | \mathcal{H}_t^n \right] = \sum_{k=1}^n Z^{(k)}(t)C^{(k)}(t)E[\{1 - (-\Delta Z^{(k)}(t))\} | \mathcal{H}_t^{(k)}]$$

$$= \sum_{k=1}^n Z^{(k)}(t)C^{(k)}(t)Pr\{\{1 - (-\Delta Z^{(k)}(t))\} = 1 | \mathcal{H}_t^{(k)}\} = \left[\sum_{k=1}^n \{1 - \chi^{Z^{(k)}}(t|\mathcal{H}_t^{(k)})\Delta t\} Z^{(k)}(t)C^{(k)}(t) \right].$$

The above equation is derived by following equation (2), since for any k :

$$Pr\{1 - \{-\Delta Z^{(k)}(t)\} = 1 | \mathcal{H}_t^{(k)}\} = Pr\{\Delta Z^{(k)}(t) = 0 | \mathcal{H}_t^{(k)}\} = \{1 - \chi^{Z^{(k)}}(t|\mathcal{H}_t^{(k)})\Delta t\};$$

where $\chi^{Z^{(k)}}(t|\mathcal{H}_t^{(k)})$ is the intensity function for the ending of an already occurred event to the k^{th} subject at t given $\mathcal{H}_t^{(k)}$, which is denoting the history of the process generating recurrent events to the k^{th} subject by time t .

However, under the assumption that for all k , $\chi^{Z^{(k)}}(t|\mathcal{H}_t^{(k)}) = \chi^Z(t|\mathcal{H}_t^n)$, which denotes the population intensity function for ending of an already occurred event to anyone in that population of n subjects at any time t ,

$$E[\Delta N_t^Z | \mathcal{H}_t^n] = \{1 - \chi^Z(t|\mathcal{H}_t^n)\Delta t\} \left[\sum_{k=1}^n \{Z^{(k)}(t)C^{(k)}(t)\} \right], \text{ which implies}$$

$$E \left[\left\{ \Delta N_t^Z - \left\{ 1 - \chi^Z(t|\mathcal{H}_t^n)\Delta t \right\} \left\{ \sum_{k=1}^n \left\{ Z^{(k)}(t)C^{(k)}(t) \right\} \right\} \right\} \middle| \mathcal{H}_t^n \right] = 0.$$

Considering

$$M^Z(t) = N_t^Z - \int_0^t \left[\sum_{k=1}^n \{Z^{(k)}(u)C^{(k)}(u)\} \{1 - \chi^Z(u|\mathcal{H}_u^n)\} du \right] = N_t^Z - \left[\sum_{0 \leq u \leq t} \sum_{k=1}^n \{Z^{(k)}(u)C^{(k)}(u)\} - \int_0^t \chi^Z(u|\mathcal{H}_u^n) \left[\sum_{k=1}^n \{Z^{(k)}(u)C^{(k)}(u)\} \right] du \right], \text{ it can}$$

be written that $dM^Z(t) = 0$, indicating the martingale property for $M^{(1)}(t)$, as explained in Andersen, Borgen, Gill and Kleiding [14].

Now, once again, because $\sum_{k=1}^n Z^{(k)}(t)C^{(k)}(t)$ and $\sum_{k=1}^n C^{(k)}(t)$ both are predictable at time t , the integrals can be written in the following form:

$$\int_0^t \left[\frac{dN_u^Z}{\sum_{k=1}^n C^{(k)}(u)} \right] = \int_0^t \left[\frac{\sum_{k=1}^n Z^{(k)}(u)C^{(k)}(u)}{\sum_{k=1}^n C^{(k)}(u)} \right] \times \{1 - \chi^Z(u|\mathcal{H}_u^n)\} du + \int_0^t \left[\frac{dM_u^Z}{\sum_{k=1}^n C^{(k)}(u)} \right]$$

Or, equivalently,

$$\int_0^t \left[\frac{dN_u^Z}{\sum_{k=1}^n C^{(k)}(u)} \right] = \sum_{0 \leq u \leq t} \left[\frac{\sum_{k=1}^n Z^{(k)}(u)C^{(k)}(u)}{\sum_{k=1}^n C^{(k)}(u)} \right] - \int_0^t \chi^Z(u|\mathcal{H}_u^n) \left[\frac{\sum_{k=1}^n Z^{(k)}(u)C^{(k)}(u)}{\sum_{k=1}^n C^{(k)}(u)} \right] du + \int_0^t \left[\frac{dM_u^{(2)}}{\sum_{k=1}^n C_k(u)} \right],$$

(following the same mathematical theory as that used in defining the Nelson Aalen estimator ([7]).

Which implies that $\int_0^t p_u^Z \{1 - \chi^Z(u|\mathcal{H}_u^n)\} du = \int_0^t \left[\frac{dN_u^Z}{\sum_{k=1}^n C^{(k)}(u)} \right]$, where $p_u^Z = \left[\frac{\sum_{k=1}^n Z^{(k)}(u)C^{(k)}(u)}{\sum_{k=1}^n C^{(k)}(u)} \right]$

(also, naturally, $\int_0^t \{1 - \chi^Z(u|\mathcal{H}_u^n)\} du = \int_0^t \left[\frac{dN_u^Z}{\sum_{k=1}^n Z^{(k)}(u)C^{(k)}(u)} \right]$).

Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.conctc.2018.03.002>.

References

- [1] J.F. Lawless, Regression models for Poisson process data, *J. Am. Stat. Assoc.* 82 (1987) 808–815.
- [2] D.Y. Lin, L.J. Wei, I. Yang, Z. Ying, Semiparametric regression for the mean and rate function of recurrent events, *J. Roy. Stat. Soc. B* 62 (part-4) (2000) 711–730.
- [3] D.Y. Lin, L.J. Wei, Z. Ying, Semiparametric transformation models for point processes, *J. Am. Stat. Assoc.* 96 (2001) 620–628 454.
- [4] M. Miloslavsky, S. Keles, M.J. van der Laan, S. Butler, Recurrent events analysis in the presence of time-dependent covariates and dependent censoring, *J. Roy. Stat. Soc. B* 66 (part 1) (2004) 239–257.
- [5] P.K. Andersen, R.D. Gill, Cox's regression model for counting processes, A large sample study, *Ann. Stat.* 10 (1982) 1100–1120.
- [6] R.J. Cook, J.F. Lawless, *The Statistical Analysis of Recurrent Events*, Springer, 2007 pp. 175–176, 121–160, 216–218, 232–236, 10, 68–69, 68.
- [7] O.O. Aalen, *Nonparametric Inference for a Family of Counting Process*, Springer Lecture Notes in Statistics, vol. 2, Springer-Verlag, New York, 1978, pp. 1–25.
- [8] R.J. Cook, J.F. Lawless, L. Lakhali-Chaieb, K. Lee, Robust estimation of mean functions and treatment effects for recurrent event under event dependent censoring and termination: application to skeletal complications in cancer metastatic to bone, *J. Am. Stat. Assoc.* 485 (2009) 60–75.
- [9] D.R. Cox, Regression models and life-tables (with discussion), *J. Roy. Stat. Soc. B* 34 (1972) 187–220.

- [10] R.L. Prentice, B.J. Williams, A.V. Peterson, On the regression analysis of multivariate failure time data, *Biometrika* 68 (1981) 373–379.
- [11] L.J. Wei, D.Y. Lin, L. Weisfeld, Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *J. Am. Stat. Assoc.* 84 (1989) 1065–1073.
- [12] C. Metcalfe, S.G. Thompson, I.R. White, Analyzing the duration of recurrent events in clinical trials: a comparison of approaches using data from the UK700 trial of psychiatric case management, *Contemp. Clin. Trials* 26 (4) (2005) 443–458.
- [13] X. Joan Hu, M. Lorenzi, J.J. Spinelli, S. Celes Ying, M.L. McBride, Analysis of recurrent events with non-negligible event duration, with application to assessing hospital utilization, *Lifetime Data Anal.* 17 (Issue 2) (2011) 215–233.
- [14] P.K. Andersen, O. Borgan, R.D. Gill, N. Keiding, *Statistical Modeling Based on Counting Processes*, Springer-Verlag, New York, 1993, pp. 48–56.