이학석사 학위논문

다양한 임상 데이터에 대한 repeated time to event 모델 개선 및 NONMEM으로 구현

Refining repeated time-to-event models across different scenarios and implementing them in NONMEM

울산대학교대학원

의과학과

서지영

Refining repeated time-to-event models across different scenarios and implementing them in NONMEM

지도교수 임형석

이 논문을 이학석사 학위 논문으로 제출함

2024년 8월

울산대학교대학원

의과학과

서지영

서지영의 이학석사학위 논문을 인준함

심사위원 배균섭 (인)

심사위원 임형석 (인)

심사위원 정성윤 (인)

울산대학교대학원

2024년 8월

Abstract

Introduction: Non-parametric/parametric time to event (TTE) approaches are widely used in the analysis of clinical trial data for drug development. It is used to assess response to drug treatment, patient prognosis and adverse drug reactions.

The popularity of TTE modelling analysis is due to the ability to understand the progression of specific events over time, which allows simulations to predict long-term outcomes following different drug administration scenarios.

Objectives: Given the variety of ways to analyze the time to event in each case, an accurate analysis is essential to predict the time and frequency of events in different clinical settings over the long term.

This study aims to guide the use of appropriate methods by characterizing events, establishing an appropriate probability distribution function and then implementing it in NONMEM through the refined equations for different time-to-event data.

Methods: Stochastic algebraic equations for different types of time-to-event data were reviewed and re-derived to reflect the characteristics of the data types. Based on these newly derived equations, parametric time-to-event analyses were performed for each typical dataset of single or repeated time-to-event with exact or interval censored time. The results of these parametric analyses were compared with the non-parametric estimates of the mean cumulative function.

Results: We categorized time to event by single or repeated event and exact or interval censored time, based on whether the exact times of the events were known or only the time ranges of the events and the number of events within a patient. Equations for probability distributions (survival function, hazard function, cumulative hazard function) for each type of event dataset were re-derived for the following time-to-event data: repeated exact-time event, repeated interval-censored event.

The newly derived equations were then implemented in NONMEM$^{®}$ to estimate parameters for an exemplary time to event datasets, and they were compared with the results of conventionally used models in NONMEM.

Monte-Carlo simulation showed that the newly derived model better described the exemplary datasets represented by mean cumulative function estimates than conventional model.

NONMEM$^{®}$ simulation results applied to the datasets provided in the existing reference and arbitrary virtual datasets demonstrated the goodness of fit of the refined equation and highlighted the distinct differences from the existing equation.

Conclusion: Using the new equations presented in this research, we expect that we can evaluate time to event data from a variety of sources such as treatment outcomes and adverse drug reactions of investigational drugs with different characteristics. The result of the current research will help to properly evaluate the efficacy and safety of drugs.

Table of Contents

List of Tables

List of Figures

Introduction

Parametric time-to-event (TTE) modelling is widely used in the analysis of clinical trial data for drug development. It is used to assess response to drug therapy, patient prognosis and adverse drug reactions.

The popularity of TTE modelling analysis lies in its ability to show the progression of specific events over time, allowing predictions of long-term drug administration scenarios through simulation. As post-treatment follow-up for relapse or adverse event analysis increases, there is a growing need for methods that analyze repeated events rather than single events.

However, time-to-event modelling has only been established for single events, and methods for repeated events are not well established.

There are only methods that consider each event independently and extend the methods used for single events, but methods that consider the relationship between each event, taking into account the properties of actual repeated events, need to be further explored. Given the variety of methods for analyzing the time to event in each case, an accurate analysis is essential to predict the time and frequency of events in different clinical settings over the long term.

Therefore, in this study, we classified the event type into exact time event and interval censored event and explored the repeated time to event analysis method from the definition of the required function to the code of NONMEM, a commonly used program.

This study aims to guide the use of appropriate methods by characterizing the event characteristics, establishing an accurate probability distribution function, and then implementing the corresponding NONMEM code through the equations of the repeated time-to-event model.

Method

I. Non-parametric estimation of the probability

A popular non-parametric approach to time-to-event analysis is the Kaplan-Meier method. This method, which first appeared in a 1958 paper by Edward L. Kaplan and Paul Meier, is a censored data analysis method. It calculates the time-related probability of an event occurring or not occurring in a population with the same characteristics in a study. The probability that the event will not occur at a given time is obtained from Kaplan-Meier.[1)]

Kaplan-Meier is usually constructed based on five assumptions. [1)] [2)]

The first assumption is that the censored event should be uninformative, i.e. the censored event should have the same probability as the observed event. This means that the underlying prognostic characteristics of subjects lost to follow-up should be the same as those of subjects observed. The second assumption is that subjects in a given trial should have the same probability of survival regardless of whether they were recruited early or late in the trial. The third assumption is that the exact time at which the specific event of interest occurred must be available to provide accurate survival estimates. The fourth assumption is that the probability is constant within the same time interval. Since the probability of an event occurring in the specific time interval is constant, the hazard rate within the specific time interval follows a unit distribution. Here, the probability of survival is the probability that an event will not occur and follows a binomial distribution that considers two cases: either an event occurs for each subject in the entire population, or no event occurs.

The fifth assumption is that all events observed per subject are distinct from each other and from censored events, so that there are no overlapping events. This assumption includes the absence of competing risk events that are distinct from the event of interest. For example, if the event of interest is a recurrence during a subject's follow-up period, Kaplan-Meier could not take into account the other competing risk event called death, which is distinct from censoring. This means that there is one event per patient, and other competing risk events that are distinct from it can be treated in the same way as censoring. Only when these five assumptions are met can the Kaplan-Meier estimator be used appropriately. The probability of survival is the probability that an event will not occur, and follows a binomial distribution that considers the occurrence of one event for each subject in the entire population.

For repeated events, however, the second assumption does not hold. Since each subject has multiple events, the probability of an event occurring is not the same for each subject over time, nor is it the case that the censored events have the same probability as the observed event. This violates the assumption of one event per subject. Therefore, the case of repeated events does not follow a binomial distribution, so the usual Kaplan-Meier equation cannot be applied.

Furthermore, the fifth assumption is also not fulfilled in the case of repeated events. Not surprisingly, all events are distinct and are treated independently. But since there is not one event per subject, this assumption is not satisfied. It is easy to see that repeated events and the presence of competing risks also make it difficult to apply Kaplan-Meier analysis.

As an alternative, time-to-first-event analysis is sometimes used when analyzing repeated events. The idea is to select only the first of several repeated events in a subject and use it to apply Kaplan-Meier estimation. By using this method, we can approximate probabilities that focus on the probability of the event not occurring over time.

However, with repeated events, the more important concept is the risk of a particular event occurring at a particular time, taking into account all cumulative events. Therefore, time-to-first-event analysis, as modified by Kaplan-Meier estimation, has difficulty in finding the cumulative risk of repeated events over time. In conclusion, the Kaplan-Meier method can be considered to be applicable when there is no competing risk event, and one event occurs per subject.

Since the value obtained by the Kaplan-Meier method is a probability, when applied to repeated events it must be transformed into a conditional probability if the previous event has occurred. However, this does not fit the assumption of a binomial distribution, so another distribution is needed.

In other words, whereas the Kaplan-Meier method focuses on the probability of the number of subjects not experiencing an event over time, assuming that there is one event per subject and that each subject's events are independent, with recurrent events there are multiple events per subject, so we need a method that focuses on the risk of an event occurring over time, assuming that the events in the same time interval are all independent, rather than that each subject's events are independent.

In summary, for repeated events it is more appropriate to calculate the cumulative risk of an event occurring over time rather than the probability of an event not occurring over time. This is the mean cumulative function approach.

Parametric methods, on the other hand, calculate probabilities based on assumptions about the distribution of the risk of an event, so it is possible to calculate directly the conditional probability of repeated events. If we want to find the probability of no event occurring at all, over time, for repeated events in a non-parametric way, we can use the relationship between the survival function and the cumulative hazard function, $S(t) = e^{-H(t)}$. The part of H(t) should be a different concept of cumulative risk function from the cumulative hazard calculated by the conventional Kaplan-Meier method. The method used here is the mean cumulative function.

The mean cumulative function is a concept proposed to adequately reflect the cumulative risk at a given point in time. This is the average number of cumulative events experienced by a subject in the study at each time point since the start of follow-up. This function could be considered a "mean curve", as it is the pointwise average of all population curves passing through the vertical line at each time point t. The mean curve is treated as continuous.

The notation of mean cumulative function is MCC(t) or M(t). And the derivative of M(t) is m(t), which is the instantaneous mean population recurrence or intensity function. m(t) means that the number of events per time unit per population unit. Therefore, conversely, the integral of m(t) is MCC(t), [3)]

Although not the same, it can be interpreted as presenting a similar concept to the cumulative hazard obtained from the Kaplan-Meier estimation. The difference between the two concepts is whether the event could be repaired or not.

Thus, for non-repaired events (from life data), the instantaneous failure rate is the hazard function h(t) and the cumulative failure rate is the cumulative hazard function. Correspondingly, for repaired events (from repeated data), the instantaneous mean hazard rate is the mean hazard rate function m(t) and the cumulative mean hazard rate function is M(t).[3)]

In MCF, three of the five assumptions of Kaplan-Meier remain the same, except for the second and fifth assumptions. MCF does not assume that each subject's MCF is identical. MCF estimates are valid whether or not there is serial correlation (or cause and effect) in the histories of events. Instead, it assumes that the occurrence of events in non-overlapping intervals are independent.3) MCF also allows for the occurrence of multiple events per patient and is applicable in the presence of competing risk events. Based on the above assumptions, MCF can be considered to follow a non-homogeneous Poisson process because each interval has the same mean risk rate of event occurrence internally, but the mean risk rate per each time interval is not homogeneous. [4)]

Conceptually, the Mean Cumulative Function (MCF) is a cumulative history function that shows the cumulative number of occurrences of an event, such as repairs, over time. When applied to a clinical setting, a repair can be thought of as a recoverable event. In the context of repair over time, the value of the Mean Cumulative Function (MCF) can be thought of as the average number of post-event repairs that each subject will have undergone after a given period of time. It applies only to subjects who have experienced recoverable events and assumes that each event (recovery) is identical.

The MCF can be applied to recurrent events or adverse events that are characterized by a subject repeating the event many times after full treatment and recovery.[5)] The concepts and notations used to calculate MCF estimates are as follows.[3)]

- N : Total number of subjects
- t_j : Time points with distinct event times, where $1 \leq j \leq m$

All time intervals use either the time of the event occurrence or a follow up period, usually in the form of closed on the left and open on the right, $[t_{j-1}, t_j)$.

- c_j : The number of censored subjects in each time interval
- e_j : The number of events that occurred in each time interval
- r_j : The number of competing-risk events that occurred in each time interval
- n_j : The number of subjects remaining in each time interval, meaning the number of subjects at risk.

The risk set (n_j) means the population of subjects at risk for an event. The risk set in MCF is constructed differently from Kaplan-Meier. Kaplan-Meier does not include competing risk events in the risk set and includes them as censoring events (c_j), but in the mean cumulative function, this part is classified differently. In Kaplan-Meier method, individuals can only experience a censoring outcome only and are removed from the risk set. And individuals can experience the event of interest and are removed from the risk set. Therefore, the risk set is $n_j = n_1 - \sum_{j=1}^{m}\left(e_j + c_j\right)$ and $n_1 = N$.

Otherwise, in the mean cumulative function, individuals can only experience a competing risk event or censoring outcome once and are removed from the risk set. And individuals can experience the event of interest and remain in the risk set without affecting the survival probability. That is, the risk set is $n_j = n_1 - \sum_{j=1}^{m}\left(r_j + c_j\right)$ and $n_1 = N$. If there is no competing risk event, r_j is all zero, so the risk set is $n_j = n_1 - \sum_{j=1}^{m} c_j$ and $n_1 = N$.

- $m_j = \left(\frac{e_j}{n_j}\right) \cdot KM\left(t_j\right),\ KM\left(t_j\right) = \prod_{j=1}^{m}(1 - \frac{r_j}{n_j})$: The immediate occurrence of an event unit per population unit in time interval, which is the mean risk rate per interval.[13)]

The value of m(t) is calculated by dividing the number of events e_j in the time interval of the interest by the remaining risk set n_j, multiplied by the proportion of subjects who have not suffered a competing risk event up to the time interval of the interest $(KM\left(t_j\right))$. If there are no competing risk events, then $KM\left(t_j\right)$ is 1 because r_j is 0 at all time intervals, leaving only the $\frac{e_j}{n_j}$ part. Then, $m_j = \frac{e_j}{n_j}$.

Since we are calculating the average of the risk rates of events occurring in each time interval, we can assume that the proportion of events occurring in each time interval is constant. This is an assumption that can be made for any Poisson distribution with the mean as a parameter, and because the risk rate of events occurring in each time interval is different, we can conclude that the mean cumulative function follows a non-homogeneous Poisson distribution.

Since MCF follows a non-homogeneous Poisson process, m_j are independent of each other, i.e., the mean risk rates for each interval can be treated as independent of each other and for all events in the same time interval, m_j is also the same.

- $M^*(t_{j+1}) = m_{j+1} + M^*(t_j)$: The sum of the mean cumulative event occurrence rates up to time t_{j+1} , which means the cumulative risk rate up to time t_{j+1}

Depending on the definition of m_j, $M^*(t_j)$ is defined differently for the case with and without competing risk events. When competing-risk events exist, $M^*(t_j) = \sum_{j=1}^{m}\left(\frac{e_j}{n_j}\right)\cdot KM(t_j),\ KM(t_j) = \prod_{j=1}^{m}(1-\frac{r_j}{n_j})$ and if there is no competing-risk event, it is defined as $M^*(t_j) = \sum_{j=1}^{m}\frac{e_j}{n_j}$.\

The relationship between M(t) and m(t) in the MCF is that the derivative form of M(t) is m(t). In a staircase curve obtained from observed data, the value of M(t_{i+1}) minus M(t_i) equals m(t_{i+1}). This can be easily seen by applying this relationship to calculate that the integral of m(t) is equal to M(t). Therefore, the differences between adjacent values of the estimated mean cumulative function are the number of events per time per population unit. If we extend the staircase graph to a continuous curve, the value of m(t) between t_i and t_{i+1} is the mean rate of change between M(t_{i+1}) and M(t_i). If we make the time intervals between t_i and t_{i+1} very small and calculate m(t) from the value of M(t) at one point, t_{i+1}, which is the process of setting the time interval, δt to zero, it becomes equal to the slope of the tangent line at M(t_{i+1}), i.e., the derivative of M(t_{i+1}) is obtained as m(t_{i+1}).

This relationship is used in parametric methods to curve H(t) and then find the parameters that fit the observed data. The calculation of h(t) at each time interval provides the basis for calculating the mean rate of change of H(t) using the observed data. A more detailed proof and analysis is described later in the parametric analysis section.

This relationship is similar as the relationship between the cumulative hazard function H(t) and h(t) obtained from the Kaplan-Meier method. A similar relationship between M(t) and m(t) and H(t) and h(t) is shown in Figure 1 and Figure 2 as follows.

Figure 1. The relationship between M(t_{i+1}) and m(t_{i+1})

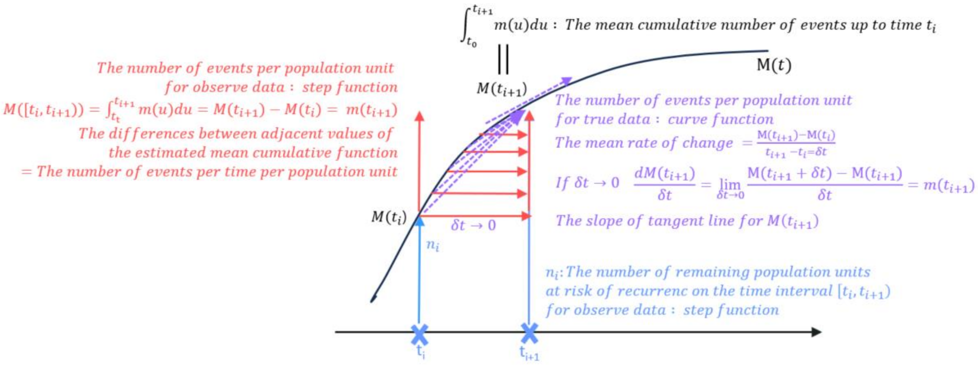
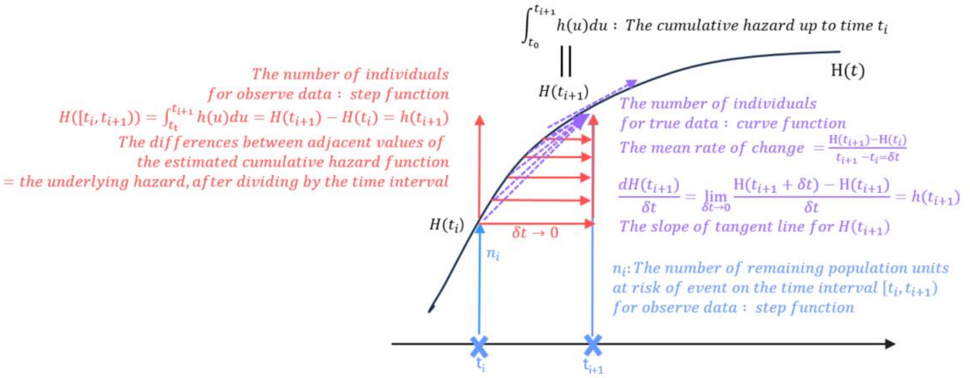

$\int_{t_0}^{t_{i+1}} m(u)du$: The mean cumulative number of events up to time t_i
||
M(t)
$M(t_{i+1})$
The number of events per population unit
for observe data : step function
$M([t_i,t_{i+1})) = \int_{t_t}^{t_{i+1}} m(u)du = M(t_{i+1}) - M(t_i) = m(t_{i+1})$
The differences between adjacent values of
the estimated mean cumulative function
= The number of events per time per population unit
The number of events per population unit
for true data : curve function
The mean rate of change $= \frac{M(t_{i+1})-M(t_i)}{t_{i+1}-t_i=\delta t}$
If $\delta t \to 0$ $\frac{dM(t_{i+1})}{\delta t} = \lim_{\delta t\to 0}\frac{M(t_{i+1}+\delta t)-M(t_{i+1})}{\delta t} = m(t_{i+1})$
The slope of tangent line for $M(t_{i+1})$
$M(t_i)$
$\delta t \to 0$
n_i
n_i: The number of remaining population units
at risk of recurrenc on the time interval $[t_i, t_{i+1})$
for observe data : step function
t_i
t_{i+1}

Figure 2. The relationship between H(t_{i+1}) and h(t_{i+1})


$\int_{t_0}^{t_{i+1}} h(u)du$: The cumulative hazard up to time t_i
||
H(t)
$H(t_{i+1})$
The number of individuals
for observe data : step function
$H([t_i,t_{i+1})) = \int_{t_t}^{t_{i+1}} h(u)du = H(t_{i+1}) - H(t_i) = h(t_{i+1})$
The differences between adjacent values of
the estimated cumulative hazard function
= the underlying hazard, after dividing by the time interval
The number of individuals
for true data : curve function
The mean rate of change $= \frac{H(t_{i+1})-H(t_i)}{t_{i+1}-t_i=\delta t}$
$\frac{dH(t_{i+1})}{\delta t} = \lim_{\delta t\to 0}\frac{H(t_{i+1}+\delta t)-H(t_{i+1})}{\delta t} = h(t_{i+1})$
The slope of tangent line for $H(t_{i+1})$
$H(t_i)$
$\delta t \to 0$
n_i
n_i: The number of remaining population units
at risk of event on the time interval $[t_i, t_{i+1})$
for observe data : step function
t_i
t_{i+1}

- $V(M^*(t_j)) = V(\sum_{k=1}^{j} m_k) = V(m_1) + V(m_2) + \cdots + V(m_j) = V\left(\sum_{k=1}^{j} \frac{m_k}{r_k}\right), m_j\ is\ independent$
- $V(m_j) = E(m_j) = \frac{m_j}{r_j}, m_j \sim Nonhomogeneous\ possion\ process$

Since the number of events(points) follows the non-homogeneous Poisson Process (NHPP) and each point occurs independently of one another in the same time interval, the mean recurrence rate of points depends on the location of the underlying space of the Poisson point process. Applying the concept of a constant mean risk rate at each time interval to a non-homogeneous Poisson process, the number of events on the time interval is a random variable X that follows a Poisson distribution with mean value m_j. Since X is observed as the value e_j for each interval, $m_j = \frac{e_j}{n_j}$, we can see that $\frac{X}{n_j}$ also follows the same distribution as X. Therefore, m_j is a random variable that also follows a Poisson distribution. So, using the property that $E(m_j)$= $Var(m_j)$, we could find the value of $Var(m_j)$ as$\frac{e_j}{n_j^2}$= $\frac{m(t_j)}{n_j}$. A summary of this process is shown as below.

- $X \sim random\ variable: the\ number\ of\ events\ depending\ on\ the\ time\ interval\ [t_j, t_{j+1})$
- $X \sim Pois(m(t)) = \frac{(m(t))^x e^{-m(t)}}{x!}, on\ the\ time\ interval\ [t_j, t_{j+1})$
- $E(X)$= $E(X = e_j) = m(t_{i+1})$= $Var(X)$
  $: the\ expected\ number\ of\ events\ per\ population\ unit\ on\ the\ time\ interval\ [t_j, t_{j+1})$
- $m(t_{j+1}) = \frac{X=e_j}{n_j} \sim random\ variable$
  $: the\ number\ of\ events\ per\ population\ depending\ on\ the\ time\ interval\ [t_j, t_{j+1})$
- $E(m(t_{j+1})) = E(\frac{X=e_j}{n_j}) = \frac{1}{n_j} E(X = e_j) = \frac{1}{n_j} E(X) = \frac{1}{n_j} \cdot m(t_{j+1}) = = \frac{M(t_{j+1}) - M(t_j)}{n_j} = \frac{1}{n_j} \cdot \frac{X=e_j}{n_j} = \frac{e_j}{n_j^2} = Var(m(t_{j+1}))$
- $Var(M(t_{j+1})) = Var(m(t_1)) + Var(m(t_2)) + \cdots + Var(m(t_{j+1})) = \sum_{k=1}^{j+1} \frac{m(t_k)}{n_k}$

The obtained Var(M(t)) as above can be used to obtain a naive confidence interval of M(t).

- $M^*(t_j) \pm Z_{\alpha/2}\sqrt{V(M^*(t_j))}$: Confidence intervals of $M^*(t_j)$

To obtain a confidence interval of $(1-\alpha)\%$, in the case of the two-sided test, the points of the inverse of the standard normal distribution corresponding to $\alpha/2$ can be found as $Z_{\alpha/2}$, multiplied by the standard error of $M^*(t_j)$ to obtain the upper and lower limits. This method is naïve confidence interval, and this could be applied to the discrete data.

While the variance of the mean cumulative function is calculated under the assumption that the number of events in each time interval follows a Poisson distribution, the variance of the Kaplan-Meier function is calculated under the assumption that the number of individuals who experienced the event or who did not experience the event in each time interval follows a Binomial distribution. In Kaplan-Meier, both the shape and variance of the cumulative hazard function are different when the survival function is obtained by the empirical estimation $S(t) = \prod_{j=1}^{m}(\frac{n_j - e_j}{n_j})$ and the Nelson's Aalen estimate $S(t) = \prod_{j=1}^{m}(e^{-\frac{e_j}{n_j}})$ respectively.

The process of finding the variance of the empirical estimate is shown as below.

- $X \sim random\ variable: the\ number\ of\ individuals\ who\ not\ experience\ event\ on\ the\ time\ interval$
- $X = n_j - e_j \sim Binom(n_j, p_j) = \binom{n_j}{x_j} {p_j}^{x_j}(1-p_j)^{n_j - x_j},\ \delta t = [t_j, t_{j+1})$
  $p_j \sim true\ probability\ of\ events\ that\ individuals\ does\ not\ experience\ the\ event\ on\ the\ time\ interval$
- $By\ normal\ approximation,$ if $n_j p_j > 5$
  $X = n_j - e_j \sim Binom(n_j, p_j) \sim N(n_j p_j, n_j p_j q_j) = N\left(n_j \frac{n_j - e_j}{n_j}, n_j \frac{n_j - e_j}{n_j} \frac{e_j}{n_j}\right) = N(n_j - e_j, \frac{n_j(n_j - e_j)}{n_j})$
- $Varp_j = Var\{\frac{n_j - e_j}{n_j}\} = \frac{1}{{n_j}^2} Var(n_j - e_j) = \frac{1}{{n_j}^2} Var(X) = \frac{1}{{n_j}^2}\{\frac{e_j(n_j - e_j)}{n_j}\}$
- $By\ Taylor\ series\ approximation, Var(-logp_j) \approx (\frac{d(-logp_j)}{dp_j})^2 Varp_j = \frac{1}{{p_j}^2} Varp_j = \frac{1}{{p_j}^2}\frac{1}{{n_j}^2}\{\frac{e_j(n_j - e_j)}{n_j}\} = \frac{e_j}{n_j(n_j - e_j)}$
- $H(t) = \sum_{j=1}^{m} \log\left(\frac{n_j}{n_j - e_j}\right) = - \sum_{j=1}^{m} \log\left(\frac{n_j - e_j}{n_j}\right) = - \sum_{j=1}^{m} \log\left(1 - \frac{e_j}{n_j}\right), for\ t_j \le t < t_{j+1}$
- $VarH(t) = Var\{-\sum_{j=1}^{m} \log\left(\frac{n_j - e_j}{n_j}\right)\} = \sum_{j=1}^{m} Var\{-log\left(\frac{n_j - e_j}{n_j}\right)\} = \sum_{j=1}^{m} Var\{-logp_j\} = \sum_{j=1}^{m}\{\frac{e_j}{n_j(n_j - e_j)}\}$

The process of finding the variance of the Nelson’s-Aalen estimate is shown as below.

- $X \sim random\ variable : the\ number\ of\ individuals\ who\ experience\ event\ on\ the\ time\ interval$
- $X = e_{\mathrm{j}} \sim Binom(n_j, p_j) = \binom{n_j}{x_j} {p_j}^{x_j} (1 - p_j)^{n_j - x_j},\ \delta t = [t_j, t_{\mathrm{j}+1}),$

  $p_j \sim true\ probability\ of\ events\ that\ individuals\ experience\ the\ event\ on\ the\ time\ interval$
- $By\ normal\ approximation,$ if $n_j p_j > 5$

  $X = e_{\mathrm{j}} \sim Binom(n_j, p_{\mathrm{j}}) \sim N(n_{\mathrm{j}} p_{\mathrm{j}}, n_j p_{\mathrm{j}} q_j) = N\left(n_j \frac{e_{\mathrm{j}}}{n_{\mathrm{j}}}, n_j \frac{n_j - e_{\mathrm{j}}}{n_j} \frac{e_{\mathrm{j}}}{n_j}\right) = N(e_{\mathrm{j}}, \frac{n_j(n_j - e_j)}{n_j})$
- $By\ Taylor\ series\ approximation,$

  $Var(logp_j) \approx\ (\frac{d(logp_j)}{dp_{\mathrm{j}}})^2 Varp_j{=}\frac{1}{{p_j}^2} Varp_j\ = \frac{1}{{p_j}^2} \frac{1}{{n_j}^2}\ \{\frac{e_{\mathrm{j}}(n_j - e_j)}{n_{\mathrm{j}}}\} = \frac{{n_i}^2}{{e_j}^2} \frac{1}{{n_i}^2}\ \{\frac{e_j(n_j - e_j)}{n_{\mathrm{j}}}\} = \frac{(n_j - e_j)}{n_j e_j}$
- $Var\{logh(t)\} = Var\{log\{\frac{e_j}{\tau_j n_j}\}\} = Var\{log\{\frac{e_j}{n_{\mathrm{j}}}\} + log\{\frac{1}{\tau_j}\}\} {=} Var\{log\{\frac{e_j}{n_j}\}\} = Var\{logp_j\} \approx \frac{(n_{\mathrm{j}} - e_j)}{n_j e_j}$
- $By\ Var\{logh(t)\} = \left(\frac{h(t)'}{h(t)}\right)^2 Var\{h(t)\} = \left(\frac{1}{h(t)}\right)^2 Var\{h(t)\},$

  $Var\{h(t)\} = h(t)^2 Var\{logh(t)\}{=}\ h(t)^2 \frac{(n_j - e_j)}{n_j e_j}$
- $H(t) = \sum_{j=1}^{m} \frac{e_{\mathrm{j}}}{n_j} {=} \sum_{j=1}^{m} \frac{e_{\mathrm{j}}}{\tau_j n_j} \tau_{\mathrm{j}} = \sum_{j=1}^{m}\ h(t)\tau_{\mathrm{j},}\ for\ t_j \le t <\ t_{j+1},\ where\ \ \tau_j = t_{j+1} - t_j$
- $VarH(t) =\ Var\{\sum_{j=1}^{m} \frac{e_{\mathrm{j}}}{n_j}\} {=} \sum_{j=1}^{m} Var\{\frac{e_{\mathrm{j}}}{\tau_j n_j}\ \tau_{\mathrm{j}}\} = \sum_{j=1}^{m} Var\{h(t)\tau_{\mathrm{j}}\} = \sum_{j=1}^{m} {\tau_j}^2 Varh(t) = \sum_{j=1}^{m} {\tau_j}^2 h(t)^2 \frac{(n_j - e_j)}{n_j e_j}$

In Kaplan-Meier, the confidence interval is calculated according to Greenwood's formula.

A comparison of the basic assumptions, definitions, and probability equations for the Kaplan-Meier estimator and the mean cumulative function estimator is shown in Table 1.

Table 1. Summary of difference between Kaplan-Meier estimators and Mean Cumulative Function estimators

Estimators	**Kaplan-Meier estimators**[2]		**Mean Cumulative Function estimators**[3][13]	
Assumption (Common)	I. The observation break should be uninformative, meaning that the observations after the observation break should have the same probability as the event being observed. II. The day, month, and year in which the specific event of interest occurred must be available to provide accurate survival estimates. III. The probability inside a time interval is constant. Because the probability of an event occurring in the interval is constant, the hazard ratio inside the interval follows a unit-distribution.			
Assumption	I. Each subject must have the same probability of survival. II. Each event per subject is independent of each other.		I. Each subject should not have the identical mean cumulative function. II. The events occurrence in non-overlapping intervals are independent of each other.	
Definition	The probability that, at a given point in time, the event will not occur.		The cumulative mean number of events within a given point in time	
Type of event	Single event		Repeated events	
Characteristics	Probability		Not probability	
Competing risk	Not available		Available, not included	Available, included
Risk set	Not include the censoring Not include the event		Not include the censoring Include the event	Not include the censoring Not include the competing risk Include the event
	$n_j = n_1 - \sum_{j=1}^{m}(e_j + c_j)$		$n_j = n_1 - \sum_{j=1}^{m} c_j$	$n_j = n_1 - \sum_{j=1}^{m}(r_j + c_j)$
Distribution in each interval	Binomial distribution		Nonhomogeneous Poisson process	
	The number of subjects W/O events X_j $X_j \sim B(n_j - e_j, \frac{n_j - e_j}{n_j})$,		The number of events X_j $X_j \sim P(m_j)$	
Risk function	Hazard function, $\hat{h}(t)$		Mean risk rate function, $m(t)$	
	Probability of immediate occurrence of an event per unit of time in each interval		The immediate occurrence of an event unit per population unit in time interval $t_j \leq t < t_{j+1}$	
	$\hat{h}(t) = \frac{e_j}{n_j \tau_j}, \tau_j = t_j - t_{j+1}$		$m(t) = \frac{e_j}{n_j}$	$m(t) = \left(\frac{e_j}{n_j}\right) * \prod_{j=1}^{m}(1 - \frac{r_j}{n_j})$
Cumulative risk function	Cumulative hazard function, $\mathrm{H}(t)$		Mean cumulative function, $M^*(t)$	
	Empirical	Nelson-Aalen	$M^*(t_{j+1}) = m_{j+1} + M^*(t_j), t_j \leq t < t_{j+1}$	
	$H(t) = \sum_{j=1}^{m} \log\left(\frac{n_j}{n_j - e_j}\right)$	$H(t) = \sum_{j=1}^{m} \frac{e_j}{n_j}$	$M^*(t_j) = \sum_{j=1}^{m} \frac{e_j}{n_j}$	$M^*(t_j) = \sum_{j=1}^{m}\left(\frac{e_j}{n_j}\right) * \prod_{j=1}^{m}(1 - \frac{r_j}{n_j}),$
No risk function	$S(t) = \prod_{j=1}^{m}\left(\frac{n_j - e_j}{n_j}\right)$	$S(t) = \prod_{j=1}^{m} e^{-(\frac{e_j}{n_j})}$	$e^{-M^*(t_j)} = e^{-\sum_{j=1}^{m}\frac{e_j}{n_j}}$	$e^{-M^*(t_j)} = e^{-\left\{\sum_{j=1}^{m}\left(\frac{e_j}{n_j}\right)*\prod_{j=1}^{m}(1-\frac{r_j}{n_j}),\right\}}$
Variance of Cumulative risk	$\sum_{j=1}^{m}\{\frac{e_j}{n_j(n_j - e_j)}\}$	$\sum_{j=1}^{m} \tau_j{}^2 h(t)^2 \frac{(n_j - e_j)}{n_j e_j}$	$Var\{\sum_{j=1}^{m}\frac{m(t)}{n_j}\}, t_j \leq t < t_{j+1}$	
No risk function	Survival function, S(t)		$e^{-M^*(t_j)} = e^{-\sum_{j=1}^{m}\frac{e_j}{n_j}}$	$e^{-M^*(t_j)} = e^{-\left\{\sum_{j=1}^{m}\left(\frac{e_j}{n_j}\right)*\prod_{j=1}^{m}(1-\frac{r_j}{n_j}),\right\}}$
	$S(t) = \prod_{j=1}^{m}\left(\frac{n_j - e_j}{n_j}\right)$	$S(t) = \prod_{j=1}^{m} e^{-(\frac{e_j}{n_j})}$		
Confidence interval	Greenwood confidence interval		Naïve confidence interval	

When using the mean cumulative function, a non-parametric analysis using R, the dataset was constructed as follows.

Table 2. The rule of assigning the EVENT(DV) values based on the event type in R

Event type	**EVENT(DV)**
Start of study (time = 0)	.
Exact time of adverse event	1
End of study (Censored time of follow up)	0

For repeated interval censored events, we took the median of each interval and replaced it with repeated exact time events for non-parametric analysis.

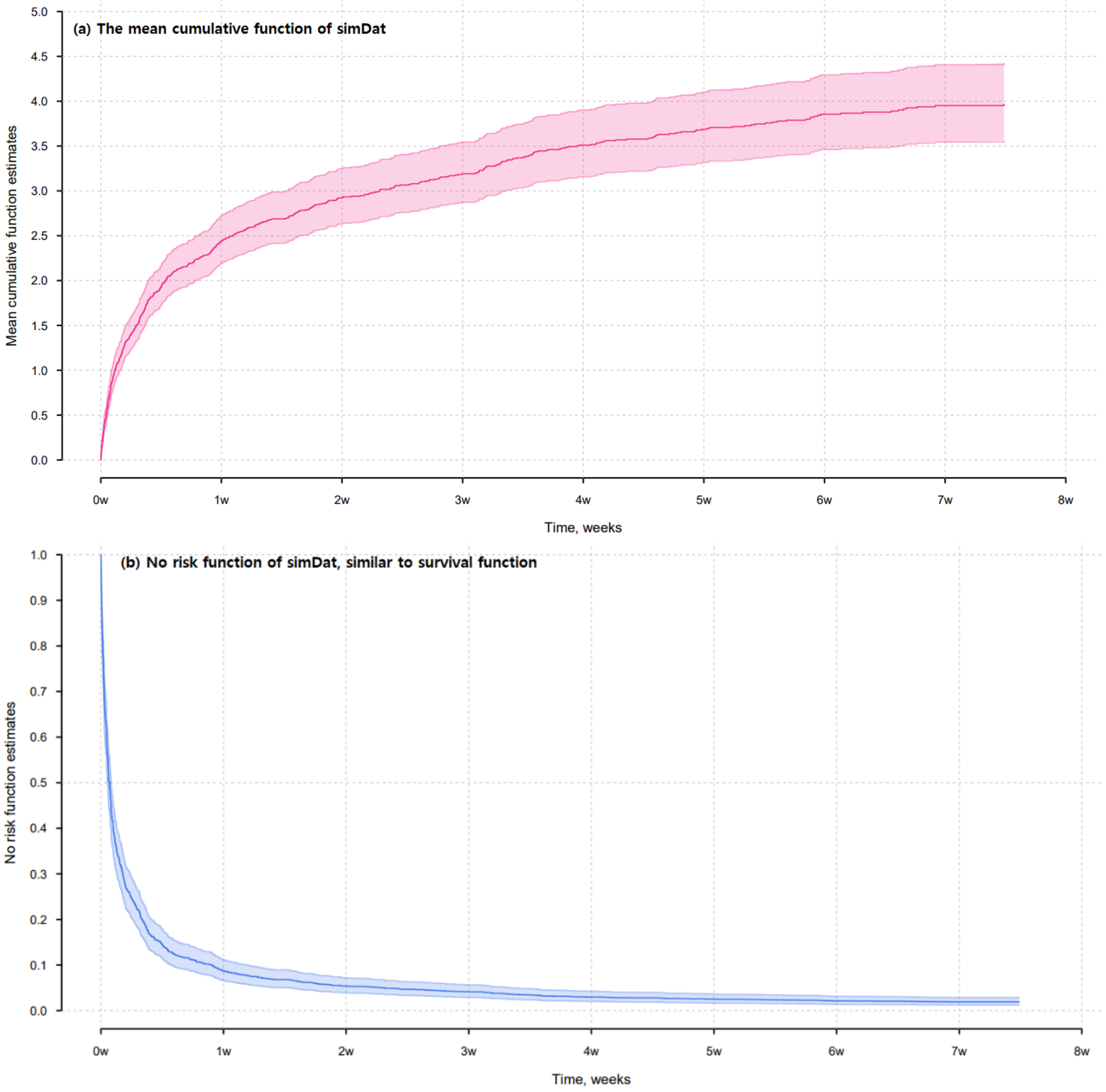
Using MCF, a non-parametric analysis, we ran example analyses for repeated exact times with and without competing risks. Examples for each case are shown in Table 3.

Table 3. The summary of examples for repeated events with/without Competing-risk events

Type	Repeated events with Competing-risk	Repeated events without Competing risk
Data	simDat from reReg R packages	The bladder tumor study data[3]
Total subjects	200	86(placebo = 48, thiotepa = 38)
Total events	674	93(placebo = 87, thiotepa = 45)
Total competing-risk events	118	No competing-risk events
Follow up period	about 60days	About 55months

Figure 3 plots the MCF of simDat, a repeated event with competing risks, and the corresponding no-risk function, like the survival function, as a function of $e^{-M(\mathrm{t})}$.

Figure 3. The mean cumulative function and no-risk function for simDa from reReg R packages

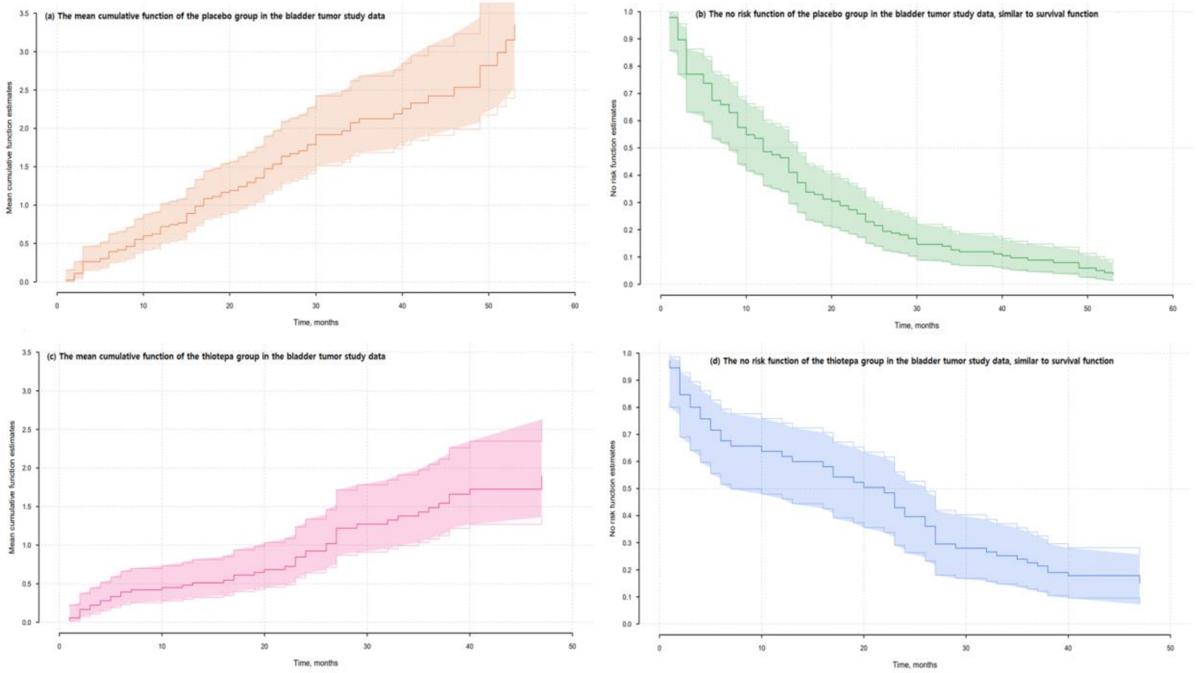


(a) The mean cumulative function of simDat

(b) No risk function of simDat, similar to survival function

Figure 4 plots the MCF of the bladder tumor study data, a repeated event without competing risks, and the corresponding no-risk function, like the survival function, as a function of $e^{-M(\mathrm{t})}$.

Figure 4. The mean cumulative function and no-risk function for the bladder tumor study data



II. Parametric estimation of the probability

The non-parametric methods described in the previous session calculate the risk of an event occurring based on the data without assuming a specific probability distribution. Parametric methods, on the other hand, assume that the pattern of events in the data follows a specific probability distribution and calculate the risk (or survival) of an event based on this assumption. [8)]

If the assumption of a particular probability distribution for the data is valid, then inferences based on that probability distribution will be more accurate. Estimates of numbers, such as relative risk and median survival time, tend to have smaller standard errors than without the distribution assumption.[9)]

In general, models that assume a specific probability distribution for survival times are called parametric models, so it is possible to calculate hazard rates for events other than survival for repeated events, assuming a specific probability distribution, depending on the shape of the cumulative hazard distribution.

In this article, we will focus on parametric analysis of the cumulative hazard for repeated events and discuss the exponential, Weibull, and log-logistic distributions, which are the three most common distributions used in survival analysis.

Before summarizing the equations for representative distributions, we first summarize the definitions and relationships of the five functions used for distributions in parametric models. Table 3 below summarizes the definitions of each function when there is one event per subject.

When there are multiple events per subject, the relationships between the functions remain related, but the descriptions of how to define the probabilities and likelihoods of each function are different, which will be discussed in more detail later. This table is organized to describe the cumulative hazard distribution and the hazard distribution for each distribution based on the relationship between the five functions.

The survival and hazard functions are constructed as a parametric method for the case where there are no competing risks.

The hazard function, which is the conditional probability of an event occurring immediately after the absence of an event up to a certain time t, can be written as $= -\frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{S(t+\Delta t)-S(t)}{\Delta t}$. The cumulative hazard function is the integral of this hazard function, and the value of $e^{(-H(t))}$, exponentiated by the negative cumulative hazard function $H(t)$ is defined as the survival function, which means the cumulative probability of an event occurring after a certain time t. The cumulative likelihood of an event occurring by time t is given by $F(t) = P(T < t) = 1 - S(t)$, which is 1 minus the value of the survival function. Therefore, the probability of an event occurring at a certain time t can be obtained by differentiating $F(t)$ i.e.,

$$f(t) = P(T = t) = -\frac{dS(t)}{dt} = h(t) * S(t).$$ [8)9)10)]

Table 4. The meaning and equation for five functions related to Time-to Single event model

Function Type	The meaning and equation for each function (Single event)
Survival function	The cumulative probability that the event has not yet occurred by time t
	$S(t) = P(T \geq t) = 1 - F(t) = 1 - \int_0^t f(u)du = e^{(-\int_0^t h(u)du)} = e^{(-H(t))}$
Probability density function	The likelihood of observing the event at the time t
	$f(t) = P(T = t) = h(t) * S(t) \ = -\frac{dS(t)}{dt}$
Cumulative density function	The cumulative likelihood of observing the event until the time t
	$F(t) = P(T < t) = \int_0^t f(u)du = 1 - S(t)$
Hazard function	The instantaneous failure rate of event occurrence at the time t_i or The probability that the event has not occurred until the time t_i but will be happened in the next instant of time t_i
	$h(t) = P(T = t \vert T \geq t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\{logS(t)\} = -\frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{S(t+\Delta t)-S(t)}{\Delta t}$
Cumulative hazard function	The cumulative instantaneous failure rate of event occurrence until the time t_i or the cumulative probability that the event has not occurred until the time t but will be happened in the next instant of time
	$H(t) = \int_0^t P(T = u \vert T \geq u)\, du = \int_0^t \left\{-\frac{1}{S(u)} \lim_{\Delta u \to 0} \frac{S(u+\Delta u)-S(u)}{\Delta u}\right\} du$

The following Table 5. are the distributional equations for these three representative distributions. The cumulative hazard distribution and the hazard distribution have been reconstructed based on the relationship between the survival function and the hazard function as described in Table 4.

Table 5. Functions for exponential, Weibull and log-logistic distributions[11)]

Distribution type	Survival function	Cumulative hazard function	Hazard function
Exponential distribution	$S(t) = e^{(-\lambda t)}$	$H(t) = \lambda t$	$h(t) = \lambda$
Weibull distribution	$S(t) = e^{\{-(\lambda t)^{\gamma}\}}$	$H(t) = (\lambda t)^{\gamma}$	$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$
Log-logistic distribution	$S(t) = \frac{1}{1+(\lambda t)^{\gamma}}$	$H(t) = \log\left\{1 + \left(\frac{t}{\lambda}\right)^{\gamma}\right\}$	$h(t) = \frac{\left(\frac{\gamma}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{\gamma}}{1+\left(\frac{t}{\lambda}\right)^{\gamma}}$

In the non-parametric analysis of repeated events described in the previous session, the MCF (Mean Cumulative Function) is the cumulative hazard function, so it is reasonable to compare the pattern with the cumulative hazard distribution of each distribution. Therefore, given each of the three representative distributions, the cumulative hazard function graph obtained from parametric analysis can be compared with the MCF graph obtained from nonparametric analysis, and the distribution like the MCF graph can be considered as the one that best describes the event.

The most commonly used software for parametric analysis is NONMEM. In this paper, we describe the definitions of the survival and hazard functions of the conventional analysis methods used in NONMEM for repeated events, their limitations, and we propose the new method to define more accurate survival and hazard functions and the implementation of the NONMEM code.

A. Repeated exact time events

If n events occur on a subject, let t_i be the time at which each event occurs. The initial time point of the follow up period for each subject is denoted by time = 0 or t_0.

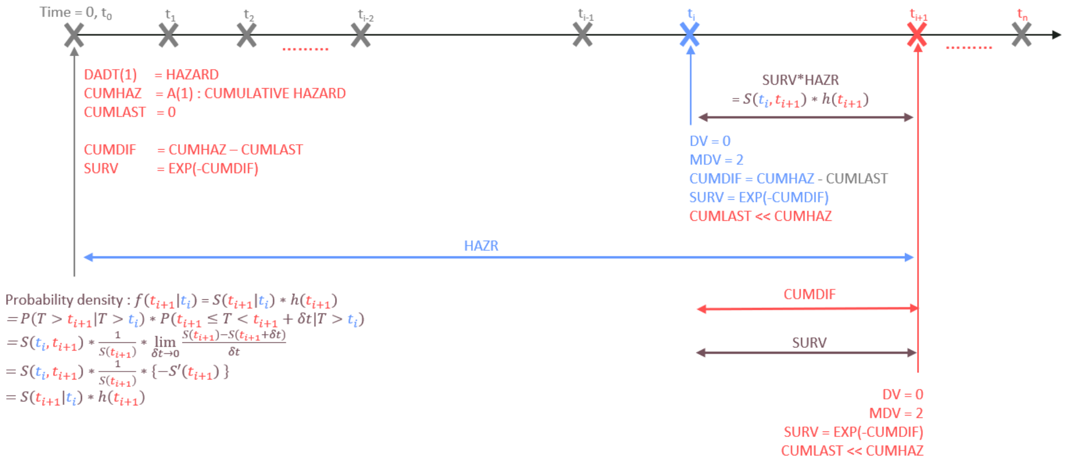
Among the repeated exact events, the likelihood of an event occurring at t_{i+1} can be calculated as the product of the survival function value at t_{i+1} and the hazard function at t_{i+1}, as defined in the table 3. When the survival function in the table 3 is defined as the probability of an event occurring after time = t_{i+1} if the event occurred after time = 0, the survival function in repeated exact events can be redefined as the probability of an event occurring after time = t_{i+1} if the event occurred after time = t_i (when the event occurred just before t_{i+1}). In other words, it is defined as a conditional cumulative probability.

Therefore, the survival function for repeated events is defined as $S(t_{i+1}|t_i)$. The hazard function at time = t_{i+1} can also be redefined as the probability that the event has not occurred by time t_{i+1}, will not occur by t_i., and will occur immediately thereafter. In other words, the original conditional probability, the hazard function, has <u>additional condition: the event must have occurred after t_i .</u> Expressed in an equation, $h(t_{i+1} \mid t_i)$ = $P((t_{i+1} \leq T < t_{i+1} + \delta t | T > t_i) | (T > t_{i+1} | T > t_i))$. Reflecting this, the likelihood of an event occurring at time = t_{i+1} is $f(t_{i+1}|t_i)$=$S(t_{i+1}|t_i) * h(t_{i+1}|t_i)$.

However, in the conventional method of calculating the likelihood, the survival function is reflected as a conditional cumulative probability, but in the case of the hazard function, no additional conditional probability is reflected, and the definition of hazard function is considered as one event occurring per subject after time = 0, $h(t_{i+1}) = P(t_{i+1} \leq T < t_{i+1} + \delta t | T > t_{i+1})$. In other words, the likelihood of an event occurring at t_{i+1} is calculated as $S(t_{i+1}|t_i) * h(t_{i+1})$ instead of $f(t_{i+1}|t_i)$=$S(t_{i+1}|t_i) * h(t_{i+1}|t_i)$.

The figure 5 depicts the equation probability density function and the corresponding NONMEM code for conventional method. SURV means survival function, HAZR means hazard function, and CUMHAZ means cumulative hazard function.

Figure 5. The probability density function and the NONMEM code for conventional method in repeated exact time events



```
DADT(1)   = HAZARD
CUMHAZ    = A(1) : CUMULATIVE HAZARD
CUMLAST   = 0

CUMDIF    = CUMHAZ – CUMLAST
SURV      = EXP(-CUMDIF)
```

```
DV = 0
MDV = 2
CUMDIF = CUMHAZ - CUMLAST
SURV = EXP(-CUMDIF)
CUMLAST << CUMHAZ
```

```
DV = 0
MDV = 2
SURV = EXP(-CUMDIF)
CUMLAST << CUMHAZ
```

Probability density : $f(t_{i+1}|t_i) = S(t_{i+1}|t_i) * h(t_{i+1})$

$$= P(T > t_{i+1} | T > t_i) * P(t_{i+1} \leq T < t_{i+1} + \delta t | T > t_i)$$
$$= S(t_i, t_{i+1}) * \frac{1}{S(t_{i+1})} * \lim_{\delta t \to 0} \frac{S(t_{i+1}) - S(t_{i+1} + \delta t)}{\delta t}$$
$$= S(t_i, t_{i+1}) * \frac{1}{S(t_{i+1})} * \{-S'(t_{i+1})\}$$
$$= S(t_{i+1}|t_i) * h(t_{i+1})$$

- Conditional time pointe of hazard function: Using the initial point t_0 of total follow up period as the conditional time point for the hazard function,
- Definition of hazard function: Although the definition of a hazard function is the probability that the random variable associated with an subject's survival time, T, lies between t and $t + \delta t$, conditional on T being greater than or equal to t, $P(t \leq T < t + \delta t | (T \geq t | T \geq t_0))$, but set the conditional time point as the initial point, t_0 of total follow up period.
- Definition of Probability density function: The product between the cumulative probability of an event occurring after t_{i+1}, given in condition that the event occurs after t_i(SURV), and the probability of an event occurring immediately at time after t_{i+1} given in condition that the event occurs after t_{i+1} (HAZR)

By adding a conditional concept to the hazard function to calculate the likelihood for repeated events, a different value is obtained from the conventional method. When calculating the value of the hazard function at t_{i+1} when an event occurs after time t_i, the probability expression

$$h(t_{i+1}|t_i) = P((t_{i+1} \leq T < t_{i+1} + \delta t|T > t_i)|(T > t_{i+1}|T > t_i))$$

is obtained, and according to the concept of conditional probability, the value of $h(t_{i+1}|t_i)$ is induced as

$$\frac{d}{dt}\left\{\int_{t_i}^{t_{i+1}} \mathrm{h}(t)dt\right\}.$$

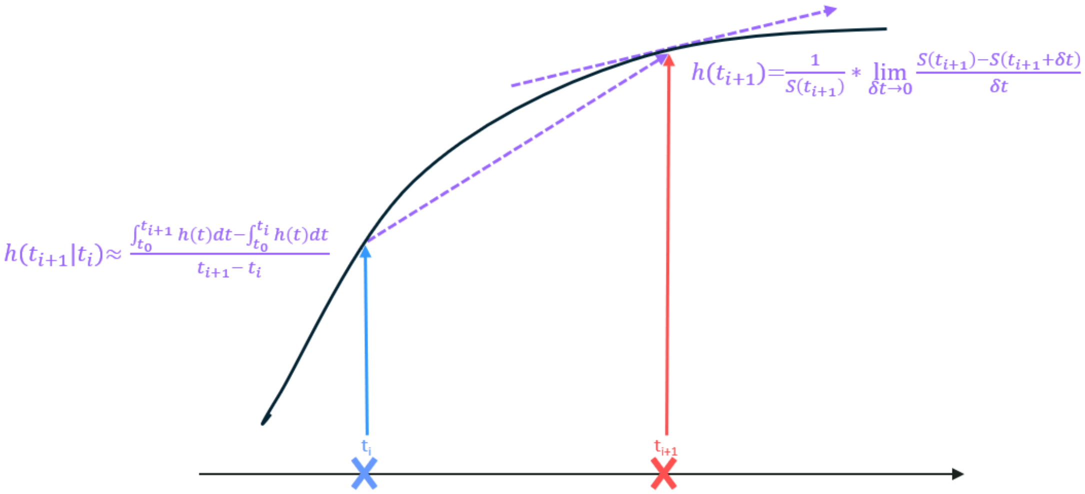
This is the derivative of the value of the cumulative hazard function from t_i to t_{i+1}. Rather than approaching it as the instantaneous rate of change of the hazard function at one time point t_{i+1}, it should be approached as the average rate of change of the cumulative hazard function from t_i to t_{i+1}, which better describes the risk of the event occurring at t_{i+1} after time t_i.

If we approach it as the instantaneous change rate at t_{i+1}, it can be solved in the conventional way as $h(t_{i+1})$, and its geometric meaning is the slope of the tangent line at t_{i+1} of the cumulative hazard function. If we approach it as the average rate of change of the cumulative hazard function from t_i to t_{i+1}, it can be expressed as

$$\frac{\int_{t_0}^{t_{i+1}} \mathrm{h}(t)dt - \int_{t_0}^{t_i} \mathrm{h}(t)dt}{t_{i+1} - t_i}.$$

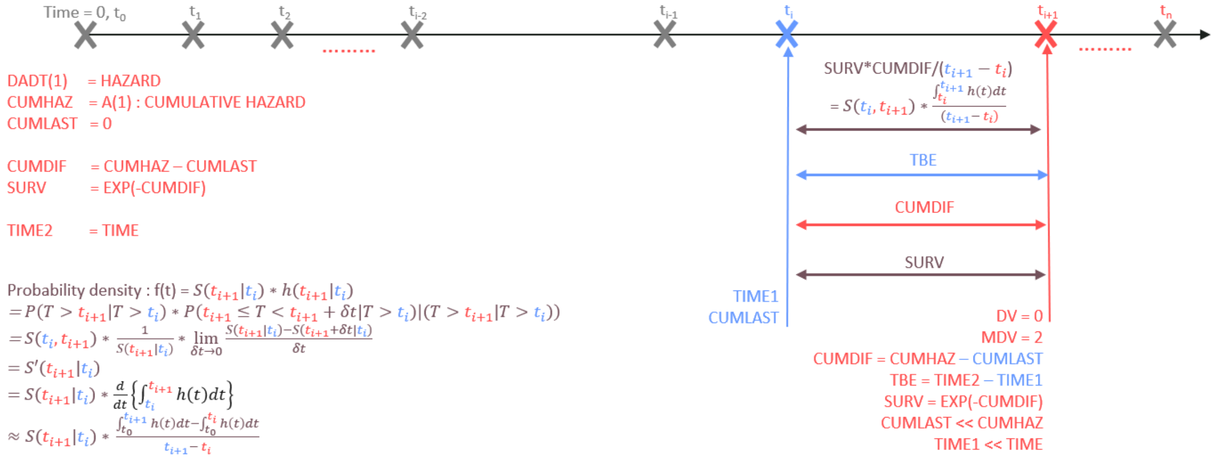
This value is a better representation to the risk of events occurring from t_i to t_{i+1}. The difference in the geometric meaning of the hazard function between the conventional and new methods is shown in the figure below.

Figure 6. The difference in the geometric meaning of the hazard function between the conventional and new methods


$h(t_{i+1}) = \frac{1}{S(t_{i+1})} * \lim_{\delta t \to 0} \frac{S(t_{i+1}) - S(t_{i+1}+\delta t)}{\delta t}$
$h(t_{i+1}|t_i) \approx \frac{\int_{t_0}^{t_{i+1}} h(t)dt - \int_{t_0}^{t_i} h(t)dt}{t_{i+1} - t_i}$
t_i
t_{i+1}

Using the newly calculated conditional hazard function, the obtained likelihood values and the corresponding new NONMEM codes are depicted as Figure 7. SURV means survival function, HAZR means hazard function, and CUMHAZ means cumulative hazard function.

Figure 7. The probability density function and the NONMEM code for new method in repeated exact time events


Time = 0, t_0
t_1
t_2
t_{i-2}
t_{i-1}
t_i
t_{i+1}
t_n
DADT(1) = HAZARD
CUMHAZ = A(1) : CUMULATIVE HAZARD
CUMLAST = 0
CUMDIF = CUMHAZ – CUMLAST
SURV = EXP(-CUMDIF)
TIME2 = TIME
Probability density : f(t) = $S(t_{i+1}|t_i) * h(t_{i+1}|t_i)$
$= P(T > t_{i+1}|T > t_i) * P(t_{i+1} \leq T < t_{i+1} + \delta t|T > t_i)|(T > t_{i+1}|T > t_i))$
$= S(t_i, t_{i+1}) * \frac{1}{S(t_{i+1}|t_i)} * \lim_{\delta t \to 0} \frac{S(t_{i+1}|t_i) - S(t_{i+1} + \delta t|t_i)}{\delta t}$
$= S'(t_{i+1}|t_i)$
$= S(t_{i+1}|t_i) * \frac{d}{dt}\left\{\int_{t_i}^{t_{i+1}} h(t)dt\right\}$
$\approx S(t_{i+1}|t_i) * \frac{\int_{t_0}^{t_{i+1}} h(t)dt - \int_{t_0}^{t_i} h(t)dt}{t_{i+1} - t_i}$
SURV*CUMDIF/$(t_{i+1} - t_i)$
$= S(t_i, t_{i+1}) * \frac{\int_{t_i}^{t_{i+1}} h(t)dt}{(t_{i+1} - t_i)}$
TBE
CUMDIF
SURV
TIME1
CUMLAST
DV = 0
MDV = 2
CUMDIF = CUMHAZ – CUMLAST
TBE = TIME2 – TIME1
SURV = EXP(-CUMDIF)
CUMLAST << CUMHAZ
TIME1 << TIME

- Conditional time point of hazard function: Using the previous event point as the conditional time point, t_i for the conditional hazard function.

- Definition of conditional hazard function: Since the definition of a hazard function is the probability that the random variable associated with an subject's survival time, T, lies between t and $t + \delta t$, conditional on T being greater than or equal to t, when T being greater than or equal to t_i, which the conditional time point is fixed by the previous event time point, t_i, $P(t \ \leq T \ < t + \delta t|(T \geq t|T \geq t_i))$.

- Definition of Probability density function: The product between the cumulative probability of an event occurring after t_{i+1}, given in condition that the event occurs after t_i(SURV), and the probability of an event occurring immediately at time after t_{i+1} given in condition that the event has not yet occurred by time t_i and the event happens after time t_{i+1} (CUMDIF/$(t_{i+1} - t_i)$), which is the average rate of change of cumulative hazard from t_i to t_{i+1}.

The definitions and derivations of the survival function, hazard function, cumulative hazard function, and the resulting probability density function for repeated exact events in the conventional and new methods are summarized in the table 6.

The main difference is whether the hazard function is calculated from the time of the event immediately from preceding the event or from an initial point of the follow up time. The value of hazard function is the instantaneous cumulative hazard change rate (hazard) at time t_{i+1} in the conventional method, while the new method uses the average change rate of the cumulative hazard values from t_i to t_{i+1}.

As a result, the method for calculating the likelihood of an event occurring at each time point has also changed.

Table 6. The meaning and equation for three functions related to Time-to repeated exact time model

Function type	The meaning and equations (repeated exact time)
Conditional Survival function	The probability that the event has not yet occurred by time t_{i+1}, in condition that the event has not yet occurred by time t_i
Conventional/New method	$S(t_{i+1}\|t_i) = P(T > t_{i+1}\|T > t_i) = \frac{S(t_{i+1})}{S(t_i)} = e^{-\int_{t_i}^{t_{i+1}} h(t)dt} = S(t_i,\ t_{i+1})$
Conditional Hazard function	The instantaneous failure rate of event occurrence at the time t_{i+1}
Conventional method	The probability that the event has not occurred until the time t_{i+1} but will be happened in the next instant of time t_{i+1}
	$h(t_{i+1}) = P(t_{i+1} \leq T < t_{i+1} + \delta t\|T > t_{i+1}) = \frac{1}{S(t_{i+1})} * \lim_{\delta t \to 0} \frac{S(t_{i+1}) - S(t_{i+1}+\delta t)}{\delta t}$ $= \frac{1}{S(t_{i+1})} * \{-S'(t_{i+1})\} =- \left[\frac{d}{dt}\left\{-\int_{t_0}^{t} h(t)dt\right\}\right]_{t_{i+1}}$
New method	The probability that the event has not occurred until the time t_i and the event happens after time t_{i+1}, but will be happened in the next instant of time t_{i+1}
	$h(t_{i+1}\|t_i) = P\big((t_{i+1} \leq T < t_{i+1} + \delta t\|T > t_i)\big\|(T > t_{i+1}\|T > t_i)\big)$ $= \frac{1}{S(t_{i+1}\|t_i)} * \left\{\lim_{\delta t \to 0} \frac{S(t_{i+1}\|t_i) - S(t_{i+1} + \delta t\|t_i)}{\delta t}\right\} = -\frac{S'(t_{i+1}\|t_i)}{S(t_{i+1}\|t_i)}$ $= \frac{S(t_{i+1}\|t_i) * \left[-\left[\frac{d}{dt}\left\{-\int_{t_i}^{t_{i+1}} h(t)dt\right\}\right]\right]}{S(t_{i+1}\|t_i)} = \frac{d}{dt}\left\{\int_{t_i}^{t_{i+1}} h(t)dt\right\} \approx \frac{\int_{t_0}^{t_{i+1}} h(t)dt - \int_{t_0}^{t_i} h(t)dt}{t_{i+1} - t_i}$
Conditional Cumulative hazard function	The cumulative instantaneous failure rate of event occurrence until the time t_{i+1}
Conventional method	The cumulative probability that the event has not occurred until the time t_i and the event happens after time $t_i < t < t_{i+1}$, but will be happened in the next instant of time $t_i < t < t_{i+1}$
	$H(t_i, t_{i+1}) =\int_{t_i}^{t_{i+1}} P(t \leq T < t + \delta t\|T > t_i)dt=\int_{t_i}^{t_{i+1}} \left[\lim_{\delta t \to 0} \frac{S(t\|t_i) - S(t+\delta t\|t_i)}{\delta t}\right] dt$ $= \frac{1}{S(t_i)} * \{-\int_{t_i}^{t_{i+1}} S'(t)\, dt\} = \frac{S\,(t_i) - S\,(t_{i+1})}{S(t_i)}$
New method	The cumulative probability that the event has not occurred until the time t_i but will be happened in the next instant of time $t_i < t < t_{i+1}$
	$H(t_i, t_{i+1}\|t_i) =\int_{t_i}^{t_{i+1}} P\big((t \ \leq T < t + \delta t\|T > t_i)\big\|(T \geq t\|T > t_i)\big)dt$ $=\int_{t_i}^{t_{i+1}} \left[\left\{\lim_{\delta t \to 0} \frac{s(t\|t_i) - s(t + \delta t\|t_i)}{\delta t}\right\} * \frac{1}{s(t\|t_i)}\right] dt = -\int_{t_i}^{t_{i+1}} \left[\frac{S'(t\|t_i)}{S(t\|t_i)}\right] dt$ $= -[logS(t\|t_i)]_{t_i}^{t_{i+1}} = -\left[-\int_{t_i}^{t_{i+1}} h(t)dt\right]=\int_{t_i}^{t_{i+1}} h(t)dt=H(t_i, t_{i+1})$

The new method presented in this paper may be more accurate than the conventional method in conceptually defining the hazard function. The table 7 shows the summary of the non-parametric and parametric analysis methods for repeated exact time events.

Table 7. The summary of the non-parametric/parametric analysis methods for repeated exact time events

Event type	Type of analysis		Method	
Exact time event	**Non-parametric analysis**		Kaplan-Meier estimates	Mean cumulative function estimates
	Parametric analysis			
	Def	**Survival function**	$S(t_{i+1}\|t_i) = e^{-\int_{t_i}^{t_{i+1}} h(t)dt}=S(t_i,\ t_{i+1})$	
		Hazard function	$h(t_{i+1})$	$h(t_{i+1}\|t_i) =\frac{\int_{t_0}^{t_{i+1}} h(t)dt - \int_{t_0}^{t_i} h(t)dt}{t_{i+1} - t_i}$
		Cumulative hazard function	$H(t_i, t_{i+1}) =\frac{S\,(t_i) - S\,(t_{i+1})}{S(t_i)}$	$H(t_i, t_{i+1}\|t_i) =\int_{t_i}^{t_{i+1}} h(t)dt=H(t_i, t_{i+1})$

	NM	**Initial condition**	CUMLAST = 0	CUMLAST = 0
				TEVENT1 = 0
				TEVENT2 = 0

		Differential equation	DATA(1) = HAZ
			CUMHAZ = A(1)
			CUMDIF = CUMHAZ – CUMLAST
			SURV = EXP(-CUMDIF)
				DATA(1) = HAZ
				CUMHAZ = A(1)
				CUMDIF = CUMHAZ – CUMLAST
				SURV = EXP(-CUMDIF)
				TIME2 = TIME
				TBE = TIME2 – TIME1

		Exact time event	Y = SURV*HAZR
			CUMLAST = CUMHAZ
				Y = SURV*(CUMDIF/TBE)
				TIME1 = TIME
				CUMLAST = CUMHAZ

		Else	CUMLAST = CUMLAST	TIME1 = TIME1
				CUMLAST = CUMLAST

		Censored event	Y = SURV = EXP(-CUMDIF)	

B. Repeated interval censored events

Next, we'll see how to calculate the likelihood of an event occurring in a repeated interval censoring event. Let $(t_i, t'_i]$ be the start and end time of each interval-censored event, given that n events occurred in a subject. The initial point of the follow up period for each subject is denoted by time = 0 or t_0.

For repeated interval censored events, the likelihood of the event occurring has the same basic idea and format as for repeated exact time events. The probability of an event occurring in an interval with a single event can be represented by the product of the survival function and the hazard function, as in the case of repeated exact time events described above. This is the basic definition of event probability. Since the survival function is the probability of the event occurring later at a given time, it makes sense to express the event probability at a given time as the product of the survival function and the hazard function. Therefore, the hazard function can be defined as the probability of the event occurring immediately at that time by using the survival function as a condition.

In repeated interval censored events, the time of event occurrence is extended to an interval, so we can extend the concept to the cumulative likelihood of event occurrence in the censored interval rather than the likelihood event occurrence at an exact time point. Furthermore, in case of the repeated exact time events, since the likelihood of event occurrence is calculated with the condition that the event occurs after the previous event, it makes sense to calculate the cumulative conditional likelihood of occurrence based on the end of interval for the previous censored event in repeated interval censored events.[11)]

Thus, for a given interval of repeated interval censored events, the likelihood of an event occurring in the interval $(t_i, t'_i]$ is the likelihood that the event occurring after t'_{i-1}, which is the end of the censored interval for the previous event, given that the event occurred within the interval $(t_i, t'_i]$ can be defined as the cumulative likelihood that an event will occur at each time point t within the interval $(t_i, t'_i]$.

When the cumulative conditional likelihood is applied into the survival and hazard functions as the above extended definition of the likelihood, the survival function within an interval $(t_i, t'_i]$ is similarly defined by adding the condition that the event occurred after t'_{i-1}, which is the endpoint of the censored interval for the previous event. Expressed as an equation, this can be written as

$$S(t|t'_{i-1}) = P((T > t|T > t'_{i-1}),\ t_i \leq t \leq t'_i.$$

For the hazard function, under the condition that the event occurred after t'_{i-1}, the endpoint of the censored interval for the previous event, we can write the hazard function within an interval $(t_i, t'_i]$ as the cumulative probability that the event will occur immediately after each point t within $(t_i, t'_i]$, given that the event has not occurred up to that time point t. The conditional hazard function is, by definition, the cumulative probability that each time point t within $(t_i, t'_i]$ can be calculated by integrating the probability of the event occurring at each time point t within $(t_i, t'_i]$. Thus, the equation can be written as follows.

$$h(t|t'_{i-1}) = P\big((t \leq T < t + \delta t\big|T > t'_{i-1})|(T > t|T > t'_{i-1})\big),\ t_i \leq t \leq t'_i.$$

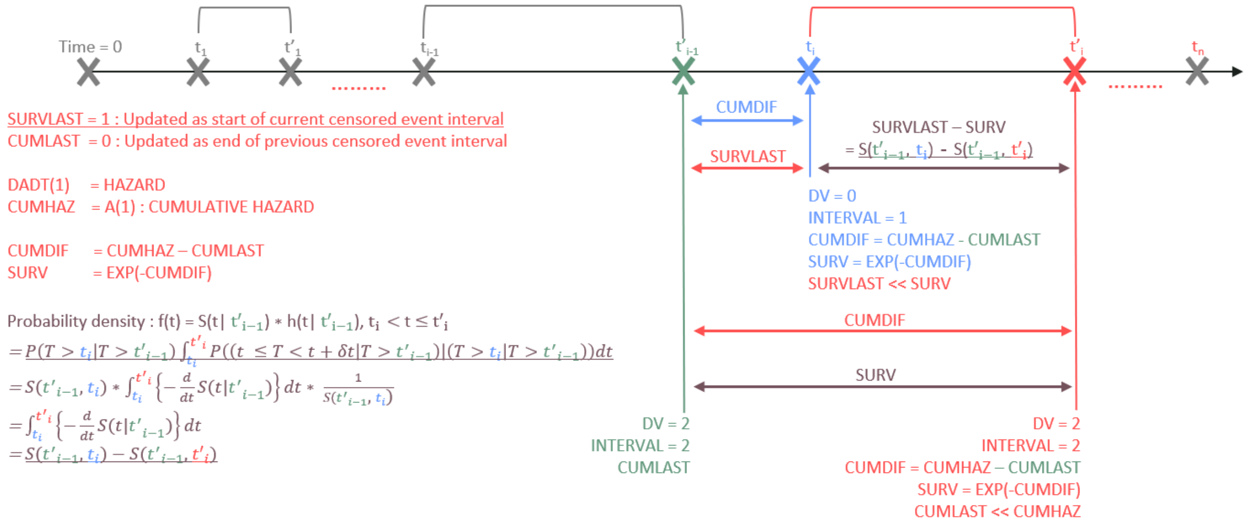
However, in the conventional method of calculating the probability of an interval-censored event, although the survival function reflects the conditional probability based on the endpoint of the interval for the previous censored event, and the additional condition of the hazard function that the event occurs after the endpoint of the interval for the previous censored event is reflected in the same way as in the new method described above, the hazard function reflects the conditional probability using the different time point as a given condition.

In case of the conventional method, when calculating the probability of event occurrence at each time point t within the interval $(t_i, t'_i]$, the concept of hazard function is not applied to each time point t within the interval $(t_i, t'_i]$ as the condition of hazard function, but rather, it uses the time point, t_i, that is the starting time point of the interval $(t_i, t'_i]$. This can be written as

$$h(t|t'_{i-1}) = P\big((t \leq T < t + \delta t\big|T > t_{i-1})|(T > t_i|T > t'_{i-1})\big),\ t_i \leq t \leq t'_i.$$

Note that the new conditional hazard function and the conventional hazard function are conditioned at different points in the interval, $(t_i, t'_i]$. The equations and corresponding NONMEM codes are summarized as Figure 8. SURV means survival function, HAZR means hazard function, and CUMHAZ means cumulative hazard function.

Figure 8. The probability density function and the NONMEM code for conventional method in repeated exact time events


Time = 0
t_1
t'_1
t_{i-1}
t'_{i-1}
t_i
t'_i
t_n
SURVLAST = 1 : Updated as start of current censored event interval
CUMLAST = 0 : Updated as end of previous censored event interval
DADT(1) = HAZARD
CUMHAZ = A(1) : CUMULATIVE HAZARD
CUMDIF = CUMHAZ – CUMLAST
SURV = EXP(-CUMDIF)
Probability density : $f(t) = S(t|\, t'_{i-1}) * h(t|\, t'_{i-1}),\ t_i < t \le t'_i$
$= \underline{P(T > t_i | T > t'_{i-1}) \int_{t_i}^{t'_i} P((t \le T < t + \delta t | T > t'_{i-1}) | (T > t_i | T > t'_{i-1}))dt}$
$= S(t'_{i-1}, t_i) * \int_{t_i}^{t'_i} \left\{ -\frac{d}{dt} S(t|t'_{i-1}) \right\} dt * \frac{1}{S(t'_{i-1}, t_i)}$
$= \int_{t_i}^{t'_i} \left\{ -\frac{d}{dt} S(t|t'_{i-1}) \right\} dt$
$= \underline{S(t'_{i-1}, t_i) - S(t'_{i-1}, t'_i)}$
CUMDIF
SURVLAST
SURVLAST – SURV
$= \underline{S(t'_{i-1}, t_i)} - \underline{S(t'_{i-1}, t'_i)}$
DV = 0
INTERVAL = 1
CUMDIF = CUMHAZ - CUMLAST
SURV = EXP(-CUMDIF)
SURVLAST << SURV
CUMDIF
SURV
DV = 2
INTERVAL = 2
CUMLAST
DV = 2
INTERVAL = 2
CUMDIF = CUMHAZ – CUMLAST
SURV = EXP(-CUMDIF)
CUMLAST << CUMHAZ

- Reference point of conditional hazard function: Using the last point, t'_{i-1} in the interval of the previous censored event as the conditional time point for the conditional hazard function in $(t_i, t'_i]$

- Definition of conditional hazard function: Although the definition of a hazard function is the probability that the random variable associated with an subject's survival time, T, lies between t and $t + \delta t$, conditional on T being greater than or equal to t, $P(t \le T < t + \delta t | (T \ge t_i | T \ge t'_{i-1}))$, but fix the condition t_i as the starting point in the interval, $(t_i, t'_i]$ of the censored event, where the probability will be calculated. And the conditional hazard obtained at each point was integrated to obtain the cumulative conditional hazard.

- Definition of Probability density function: The cumulative probability of an event occurring after t_i given in condition that the event occurs after t'_{i-1}(SURVLAST) minus the cumulative probability of an event occurring after t'_i given in condition that the event occurring after t_i (SURV)

If we define the conditional hazard function in the conventional method, when calculating the probability of event occurrence at each point in the interval $(t_i, t'_i]$, the two conditions are reflected as fixed conditions that do not move. The first condition is that the event occurs after the end of the interval at previous censored event t'_{i-1}, and the second condition is that the event occurs after the starting point of the current interval t_i.

With the two conditions fixed, the two conditions are not included in the integral process while the cumulative probability of the event occurring immediately at each point within $(t_i, t'_i]$ is calculated.

On the other hand, in the new method, one of the two conditions reflected in the conditional hazard function is used as a fixed condition, while the other condition is applied differently at each time point t in $(t_i, t'_i]$, according to t . The first condition is if the event occurs after the end of the interval at the previous censored event t'_{i-1}, and the second condition is if the event has not occurred by each time point t in $(t_i, t'_i]$, according to t.

The first condition is fixed, while the second condition is not fixed and the variable t is included in the integral process that yields the cumulative probability in $(t_i, t'_i]$. The difference in how the conditional hazard function is calculated between the conventional and new methods is illustrated in a more intuitive way in the figure 9 and figure 10.

Figure 9. How the conditional hazard function is calculated in the conventional method

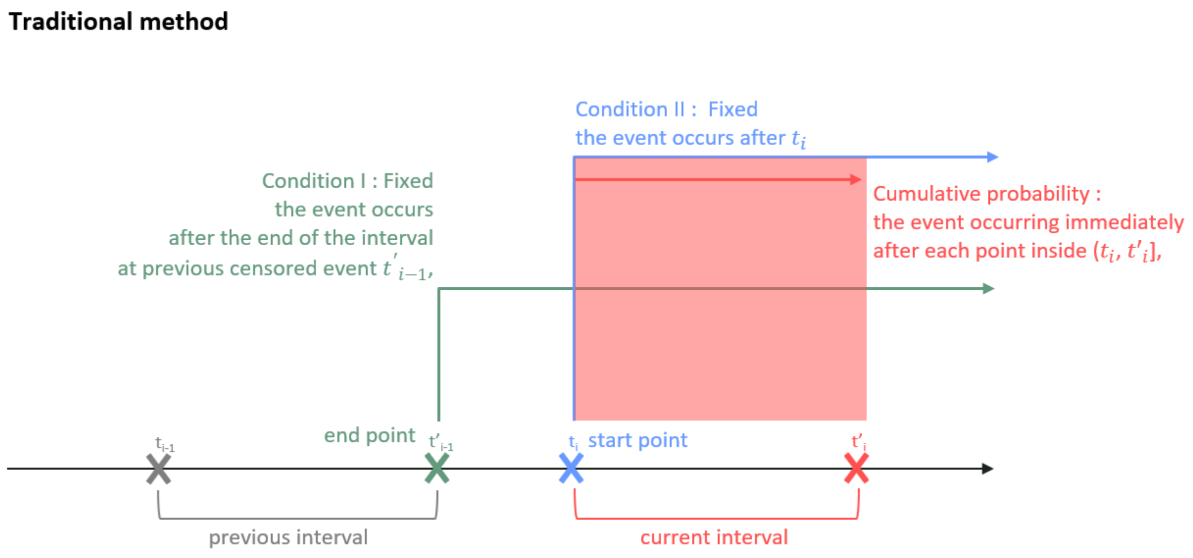
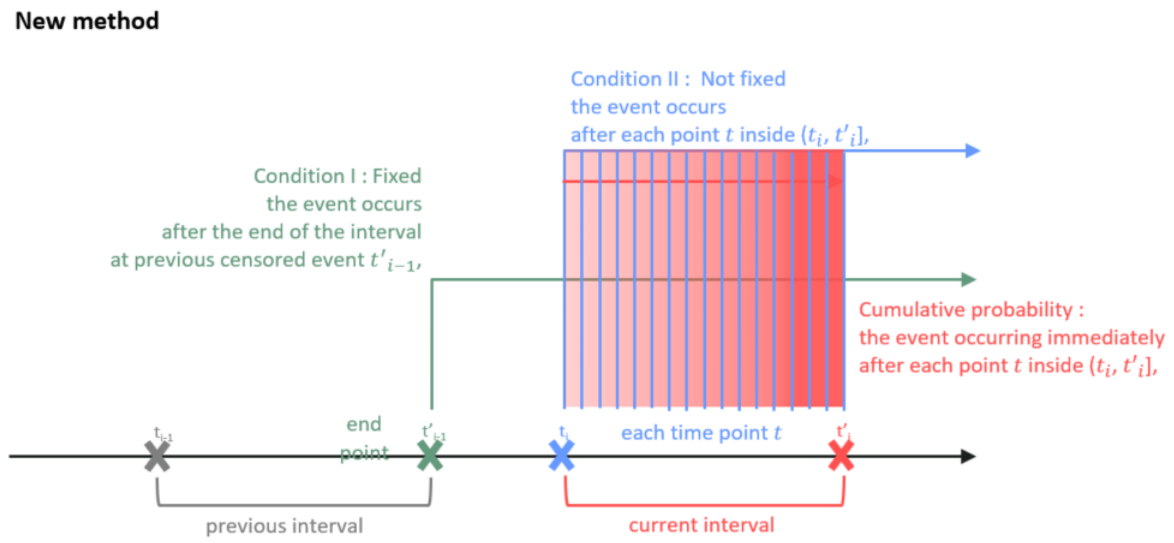


Figure 10. How the conditional hazard function is calculated in the new method



The definitions and derivations of the survival function, hazard function, cumulative hazard function, and the resulting probability density function for repeated interval censored events in the conventional and new methods are summarized in the table 8.

The main difference is whether the conditional hazard function is calculated from the time of the event immediately from the starting point t_i(fixed) of the current interval $(t_i, t'_i]$ or from the interval time point t(moved) of the current interval $(t_i, t'_i]$. As a result, the method of calculating the likelihood of an event occurring at the interval time point of the current interval $(t_i, t'_i]$ has also changed.

Table 8. The meaning and equation for three functions related to Time-to repeated interval censored events model

Function type	**The meaning and equations (Repeated interval censored events)**
Conditional Survival function	The probability that the event has not yet occurred by time t_i, in condition that the event has not yet occurred by time t'_{i-1}
Conventional/New method	$S(t\|t'_{i-1}) = P((T > t\|T > t'_{i-1}) = \frac{S(t)}{S(t'_{i-1})} = e^{-\int_{t'_{i-1}}^{t} h(t)dt} = S(t'_{i-1}, t)\ ,\ t_i \leq t \leq t'_i$
Conditional Hazard function	The instantaneous failure rate of event occurrence at the time t inside $(t_i,\ t'_i]$
Conventional method	The probability that the event has not occurred until the time t_i and the event happens after time t'_{i-1}, but will be happened in the next instant of time $t_i \leq t \leq t'_i$
	$h(t\|t'_{i-1}) = P((t \leq T < t + \delta t \| T > t_{i-1}) \| (T > t_i \| T > t'_{i-1})) = \left\{-\frac{d}{dt} S(t\|t'_{i-1})\right\} * \frac{1}{S(t'_{i-1}, t_i)}$ $= \left[-\frac{d}{dt}\left\{-\int_{t'_{i-1}}^{t} h(t)dt\right\}\right] * e^{-\int_{t'_{i-1}}^{t} h(t)dt} * \frac{1}{S(t'_{i-1}, t_i)} = \frac{d}{dt}\left\{\int_{t'_{i-1}}^{t} h(t)dt\right\},\ t_i \leq t \leq t'_i$
New method	The probability that the event has not occurred until the time t inside $(t_i,\ t'_i]$ and the event happens after time t'_{i-1}, but will be happened in the next instant of time $t_i \leq t \leq t'_i$
	$h(t\|t'_{i-1}) = P((t \leq T < t + \delta t \| T > t_{i-1}) \| (T > t \| T > t'_{i-1}))$ $= \left\{\lim_{\delta t \to 0} \frac{S(t\|t'_{i-1}) - S(t + \delta t\|t'_{i-1})}{\delta t}\right\} * \frac{1}{S(t\|t'_{i-1})} = \frac{S\prime(t\|t'_{i-1})}{S(t\|t'_{i-1})},\ t_i \leq t \leq t'_i$
Conditional Cumulative hazard function	The cumulative instantaneous failure rate of event occurrence until the time t_i
Conventional method	The cumulative probability that the event has not occurred until the time t_i and the event happens after time t'_{i-1}, but will be happened in the next instant of time $t_i \leq t \leq t'_i$
	$H(t_i, t'_i\|t'_{i-1}) = \int_{t_i}^{t'_i} P((t\ \leq T < t + \delta t \| T > t'_{i-1}) \| (T > t_i \| T > t'_{i-1}))dt$ $= \int_{t_i}^{t'_i} \left\{-\frac{d}{dt} S(t\|t'_{i-1})\right\} dt * \frac{1}{S(t'_{i-1}, t_i)} = \frac{S(t'_{i-1}, t'_i) - S(t'_{i-1}, t_i)}{S(t'_{i-1}, t_i)}$
New method	The cumulative probability that the event has not occurred until the time t inside $(t_i,\ t'_i]$ and the event happens after time t'_{i-1}, but will be happened in the next instant of time $t_i \leq t \leq t'_i$
	$H(t_i, t'_i\|t'_{i-1}) = \int_{t_i}^{t'_i} P((t\ \leq T < t + \delta t \| T > t'_{i-1}) \| (T \geq t \| T > t'_{i-1}))dt$ $= \int_{t_i}^{t'_i} \left[\left\{\lim_{\delta t \to 0} \frac{S(t\|t'_{i-1}) - S(t + \delta t\|t'_{i-1})}{\delta t}\right\} * \frac{1}{S(t\|t'_{i-1})}\right] dt$ $= \left[-logS(t'_i\|t'_{i-1}) + logS(t_i\|t'_{i-1})\right]$ $= H(t'_{i-1},\ t'_i) - H(t'_{i-1}, t_i) = H(t_i, t'_i)$

The table 9 shows the summary of the non-parametric and parametric analysis methods for repeated interval censored events.

Table 9. The summary of the non-parametric and parametric analysis methods for repeated interval censored events

Type of event	**Type of analysis**		**Method**	
Interval censored event	**Non-parametric analysis**		Kaplan-Meier estimates	Mean cumulative function estimates
	Parametric analysis		where $t_i \leq t \leq t'_i$	
	Def	**Survival function**	$S(t\|t'_{i-1}) = e^{-\int_{t'_{i-1}}^{t} h(t)dt} = S(t'_{i-1},\ t)$	
		Hazard function	$h(t\|t'_{i-1}) = \frac{d}{dt}\left\{\int_{t'_{i-1}}^{t} h(t)dt\right\}$	$h(t\|t'_{i-1}) = \frac{S\prime(t\|t'_{i-1})}{S(t\|t'_{i-1})}$
		Cumulative hazard function	$H(t_i, t'_i\|t'_{i-1}) = \frac{S(t'_{i-1},\ t'_i) - S(t'_{i-1},\ t_i)}{S(t'_{i-1}, t_i)}$	$H(t_i, t'_i\|t'_{i-1}) = \int_{t_i}^{t_{i+1}} h(t)dt = H(t_i, t_{i+1})$

	NM	**Initial condition**	SURVLAST = 1
			CUMLAST = 1
				SURV1 = 1
				CUMLAST = 1

		Differential equation	DATA(1) = HAZ
			CUMHAZ = A(1)
			CUMDIF = CUMHAZ – CUMLAST	

		Start of interval censored event	SURVLAST = SURV	SURV1 = SURV
				CUMLAST = CUMHAZ

		End of interval censored event	Y = SURVLAST – SURV
			CUMLAST = CUMHAZ
				Y = SURV1*CUMDIF
				CUMLAST = CUMHAZ

		Censored event	Y = SURV = EXP(-CUMDIF)	

The two example datasets used in the non-parametric analysis were applied to the parametric analysis. When using the two examples' data sets for parametric analysis using NONMEM, the dataset was structured as follows.

Table 10. The rule of assigning the EVENT(DV) values based on the event type in NONMEM

Event Type	EVID	EVENT(DV)	MDV	INTERVAL
Start of study(time = 0)	2	.	1	.
Exact time of adverse event	0	0	0	.
Start time of interval censored event	2	.	1	1
End time of interval censored event	0	2	0	2
End of study(Censored time of follow up)	0	1	0	.

Result

We applied the new methodology described in method session to two examples regarding repeated events. Different examples were used for Repeated exact time events and Repeated interval censored events, depending on the event occurrence time. The data used in analysis is as following table 11.

Table 11. The summary of two examples for both repeated exact time event and repeated interval censored event

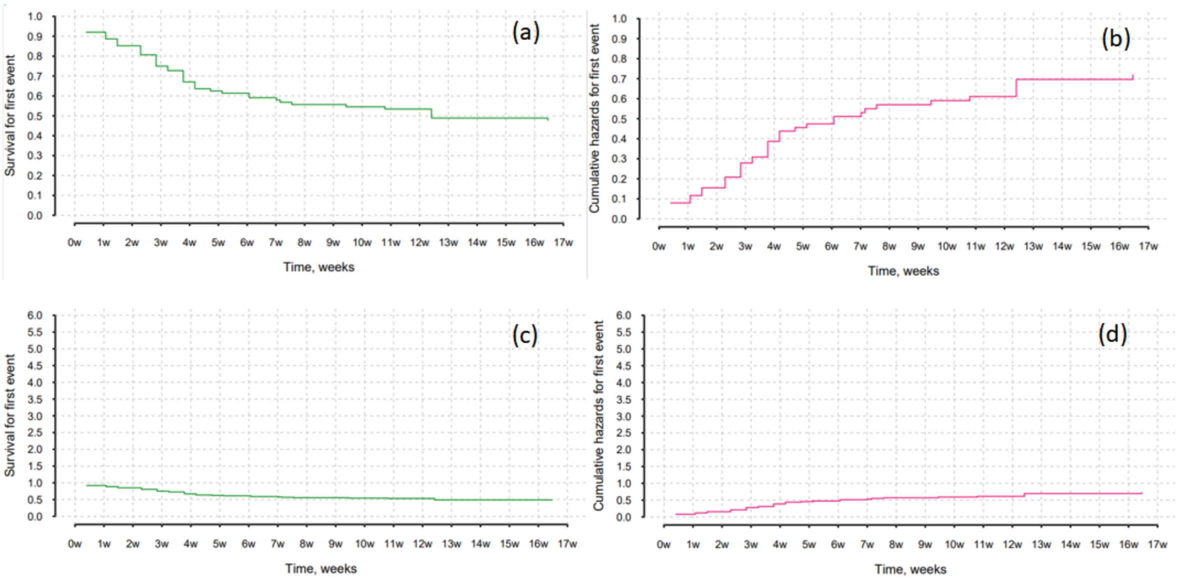
Type	Repeated exact time event	Repeated interval censored event
Data	Female rat data [9)]	Curated and jittered of real adverse events data
Total subjects	48(23rats, 25controls)	56
Total events	212	94
Follow up period	about 122days, 18weeks	about 10weeks, 168days, 1680hrs
Treatment	Retionid prophylaxis	Obesity drug
Dosing time	first dosing at 0time	first dosing at 0time
Characteristics	recurrent mammary tumors after treatment	Not overlapped repeated adverse events in each subject

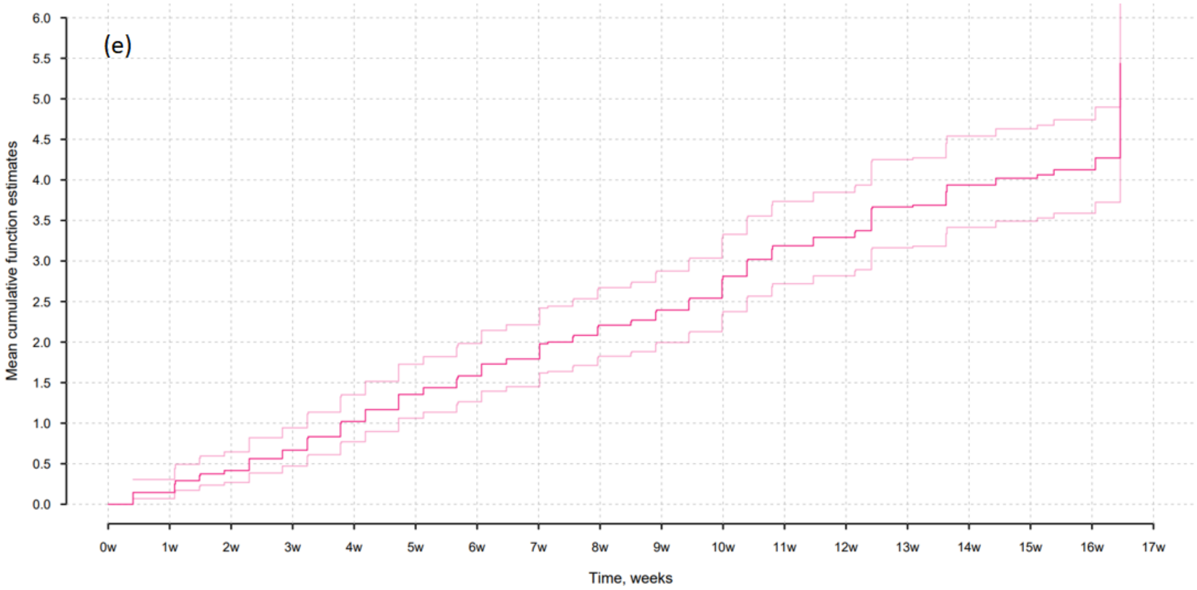
I. Non-parametric estimation of the probability

A. Repeated exact time events

Here is a plot of Kaplan-Meier estimates (First event per subject) and mean cumulative function estimates for repeated recurrent events in female rat data. Kaplan-Meier estimates were analyzed using only the first event of each rat.

Figure 11. Kaplan-Meier estimates of time to first event model for repeated exact time event data (rat data)





(a) is a survival plot using Kaplan-Meier estimates and (b) is a cumulative hazard plot. (c) and (d) are plotted by changing the y-axis of (a) and (b) to 0-6 instead of 0-1. In fact, changing the y-axis to 6 is not significant for (a), which is a survival function. (e) shows the mean cumulative function estimates.

Figure (e), which uses all events instead of just the first event for each rat, is quite different from (b).

Comparing (d) with (e), where the y-axis in (b) is scaled from 0 to 6 as in (e), the time to first event analysis using conventional Kaplan-Meier estimates underestimates the proportion of events that occur.

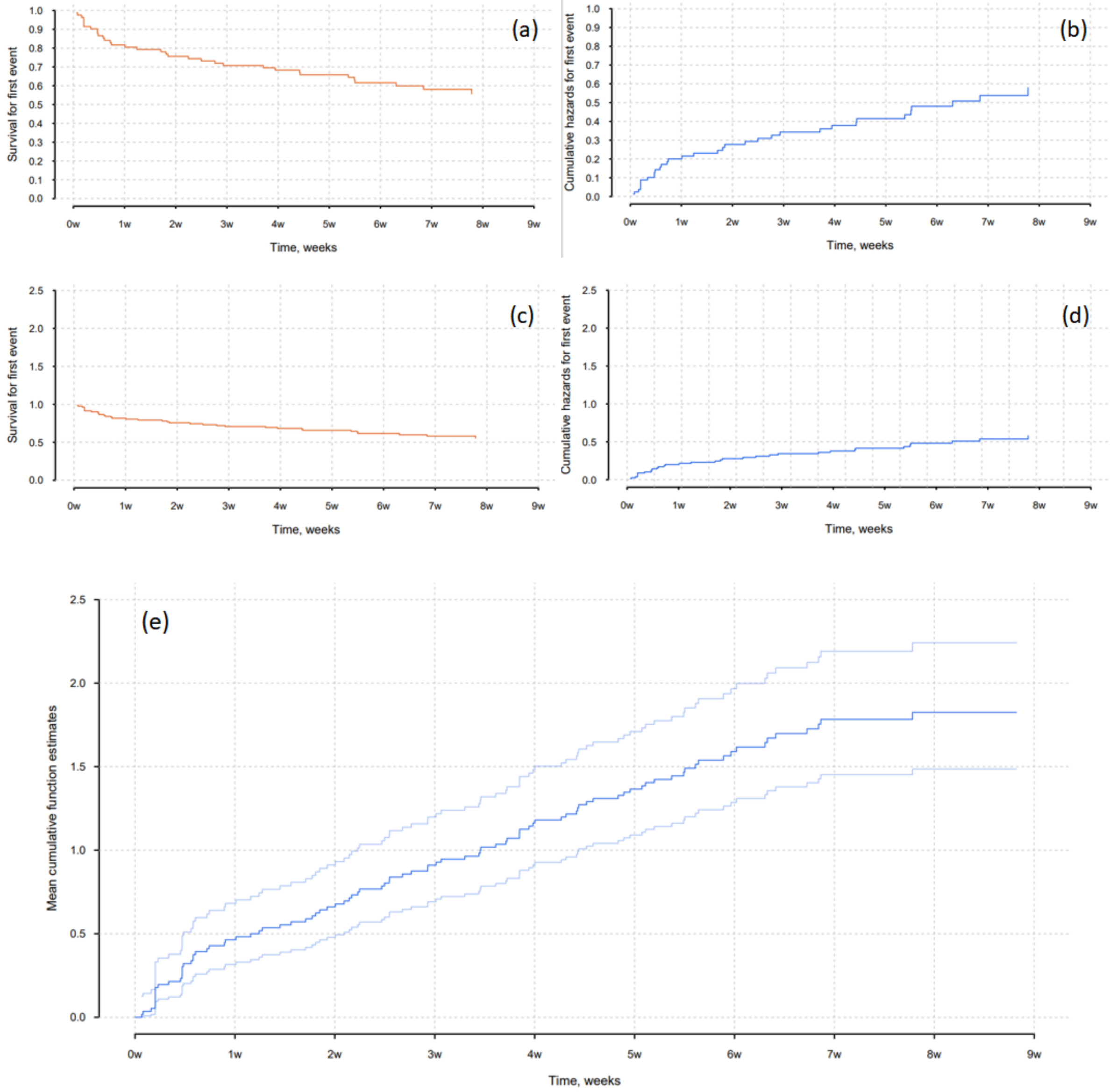
The accuracy is therefore much lower. In addition, in the case of repeated events, the cumulative event rate at each time point has important clinical implications, whereas looking at the first event only focuses on the probability of no event occurring, which is difficult to find meaningful clinical implications.

Figure (e) shows that the mean cumulative function estimates rise in a gentle curve. At the last time point, around week 17, censoring has largely occurred, causing the mean cumulative function estimates to rise dramatically. If we were to take censoring into account, we would expect the value of 4.5 found at week 16 to be closer to the mean cumulative function estimates at the last event.

B. Repeated interval censored events

Here is a plot of Kaplan-Meier estimates (first event per subject) and mean cumulative function estimates for repeated recurrent events in curated and jittered real-world adverse event data. Non-parametric analysis was performed using the midpoint of the censored interval. Kaplan-Meier estimates were analyzed using only the first event of each subject.

Figure 12. Kaplan-Meier estimates of time to first event model for repeated interval censored event data (curated and jittered of real adverse events data)



(a) is a survival plot using Kaplan-Meier estimates, (b) is a cumulative hazard plot, (c) and (d) are plots by changing the y-axis in (a) and (b) to 0-2.5 instead of 0-1, and (e) shows the mean cumulative function estimate. It shows a similar tendency to the Repeated exact time events results analyzed earlier.

Figure (e) shows that the mean cumulative function estimate rises in a slightly convex curve. At the end, around week 8, we see that the mean cumulative function estimate approaches around 2, while the cumulative hazard value of the Kaplan-Meier estimate stays around 0.6.

When Kaplan-Meier analyses repeated exact time events and repeated interval censored events, the underestimation is not surprising as only one occurrence of an event is counted. Therefore it is more appropriate to use the mean cumulative function estimate to check the exact rate of repeated events at each time point, thus validating the parametric analysis value.

II. Parametric estimation of the probability

A. Repeated exact time events

The conventional method and the new method were applied to NONMEM to compare the results of the implementation of the rat data, which are repeated exact time event data. The parameter estimates and cumulative hazard plots of each of the three distributions were compared with the mean cumulative function estimate plots obtained from non-parametric analysis to find the most appropriate distribution.

Figure 13. Simulation plots for three distribution model using the conventional method in NONMEM for repeated exact time event data (rat data)

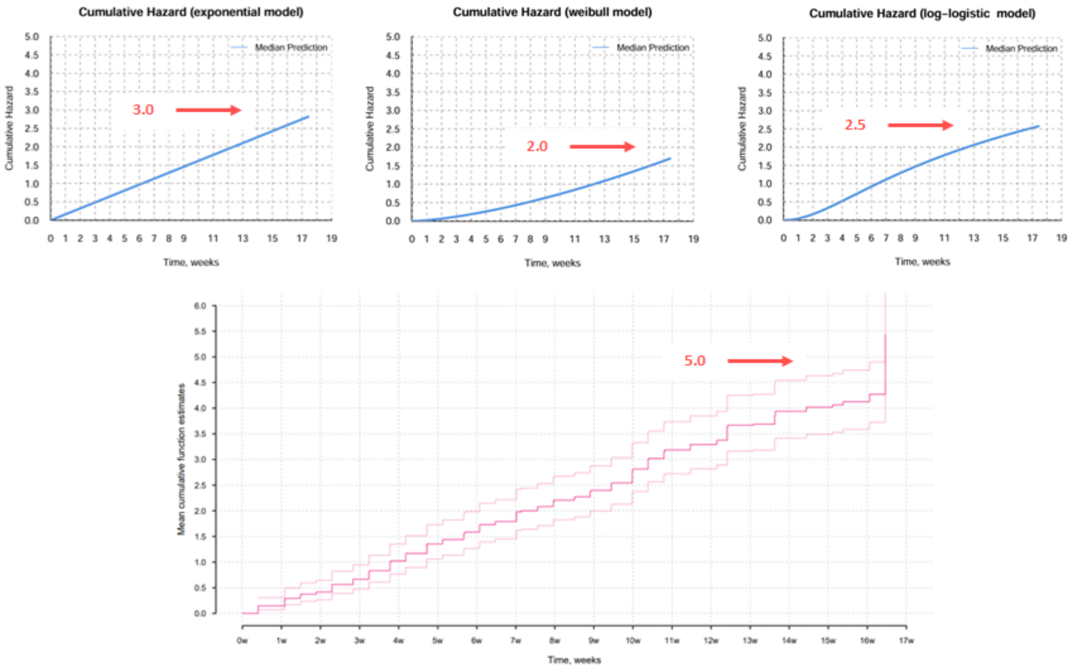


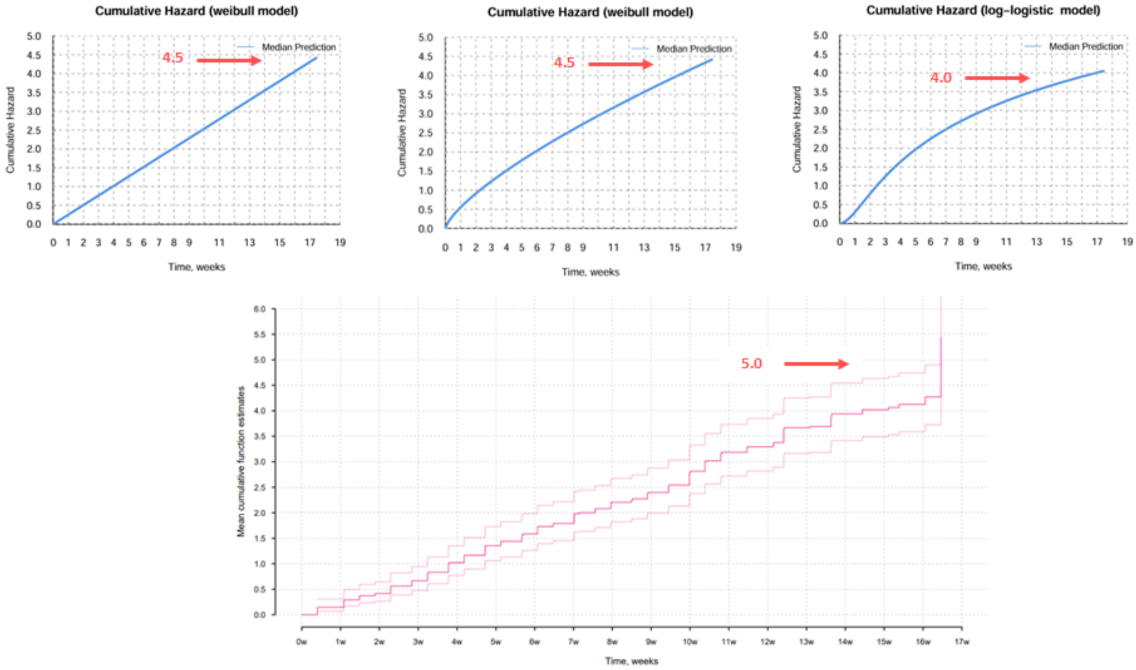
Figure 13 shows the cumulative hazard graphs obtained by applying the conventional method to the three distributions in repeated exact time event data. Compared with the graph obtained from the non-parametric analysis, we can see that the graphs obtained from the three distributions do not fit well.

The slope is concave for the Weibull distribution and convex for the log-logistic distribution, with the exception of the exponential distribution which has a straight increasing slope. Of the three distributions, the Weibull distribution has a similar slope shape to the non-parametric analysis.

However, the final value of the cumulative hazard estimates, which ranged from 4.0 to 5.0 in the non-parametric analysis, showed very different values in the plots obtained from the three distributions, with an estimate of 3.0 for the exponential distribution, 2.0 for the Weibull distribution and 2.5 for the log-logistic distribution.

The following figure 14 is the cumulative hazard graphs implemented by applying the new method to three distributions in repeated exact time events data.

Figure 14. Simulation plots for three distribution model using the new method in NONMEM for repeated exact time event data (rat data)



Compared to the graphs obtained by non-parametric analysis in Figure 10, we can see that the graphs obtained from the three distributions have a better fit compared to the conventional method.

In the case of slope, except for the exponential distribution, which increases in a straight line, the Weibull distribution increases relatively smoothly, while the log-logistic distribution is convex.

Of the three distributions, the Weibull distribution has a similar slope shape to the non-parametric analysis.

In the non-parametric analysis, we can see that the final cumulative hazard estimates obtained with values between 4.0 and 5.0 are similar to the plots obtained from each of the three distributions. The exponential distribution showed very close values with an estimate of 4.5, the Weibull distribution 4.5 and the log-logistic distribution 4.0.

Comparing the slope and the final estimates, we can conclude that the Weibull distribution is the best fitting model.

The comparison between the estimates of the parameters obtained by NONMEM and the value of the objective function using the conventional method and the new method in repeated exact time data, rat data, is as follows Table 12.

Table 12. The summary of results in both conventional method and new method in NONMEM for repeated exact time event data (rat data)

Distribution	Estimator		NONMEM	
			Conventional method	New method
Exponential	Objective function value		2021.554	1831.109
	Lamda(λ)	Estimate	2.31E-02	3.62E-02
		S.E	2.42E-03	4.16E-03
Weibull	Objective function value		2015.917	1812.881
	Lamda(λ)	Estimate	2.01E-02	6.30E-02
		S.E	2.00E-03	9.24E-03
	Gamma(γ)	Estimate	1.18E+00	7.28E-01
		S.E	8.19E-02	6.39E-02
Log-logistic	Objective function value		2034.910	1837.811
	Lamda(λ)	Estimate	3.40E+01	1.26E+01
		S.E	3.75E+00	2.15E+00
	Gamma(γ)	Estimate	1.95E+00	1.77E+00
		S.E	1.81E-01	2.30E-01

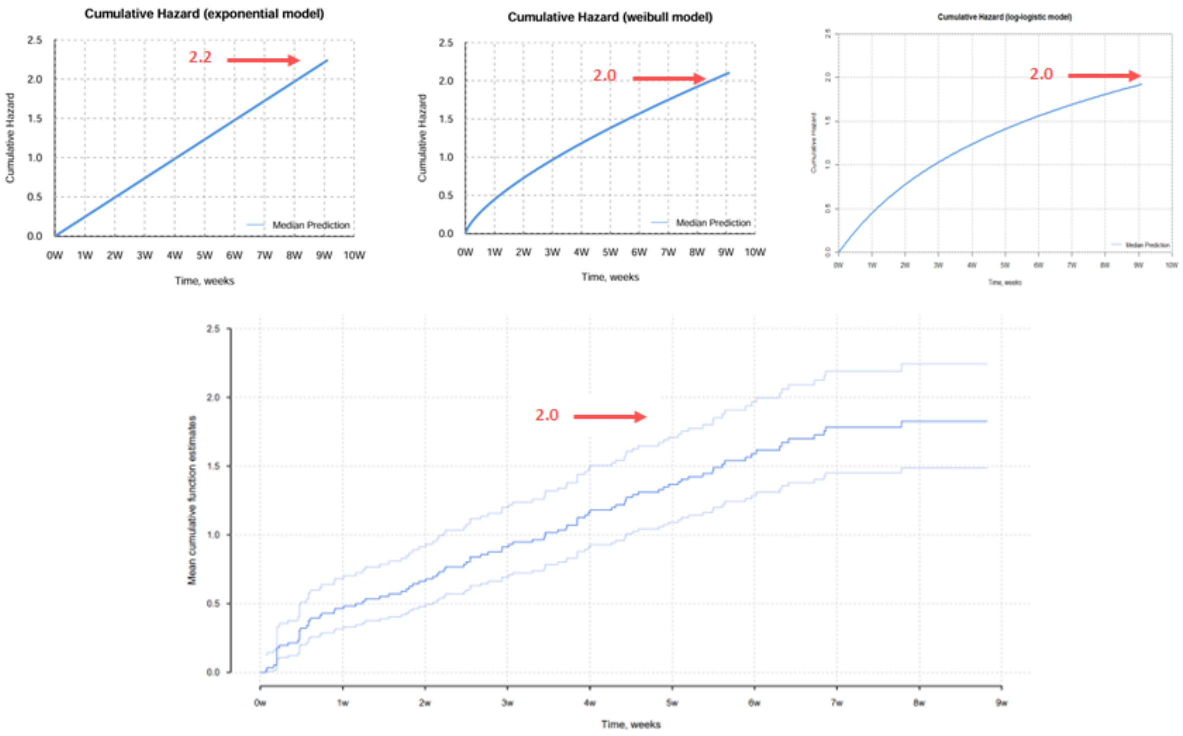
As the implementation of the NONMEM code is slightly different and no parameters have been added, the numerical comparison of the objective function values in both methods is not significant. The objective function values obtained by the new method are smaller than those obtained by the conventional method for all distributions.

When all distributions are run in both methods, minimization successes are obtained and when a Wald test is performed on the estimates of each parameter, they all fall within 2S.E (standard errors), indicating that the estimates are valid. When all the estimates are valid, validation by graph is possible, and when comparing the graphs implemented with the estimates obtained by the two methods, it can be seen that the graph shape obtained by the new method better describes the rat data.

B. Repeated interval censored events

The conventional method and the new method were applied to NONMEM to compare the results of implementing curated and jittered of real adverse events data, which are repeated interval censored time events data. The parameter estimates and cumulative hazard plots of each of the three distributions were compared with the mean cumulative function estimate plots obtained from non-parametric analysis to find the most appropriate distribution.

Figure 15. Simulation plots for three distribution model using the conventional method in NONMEM for repeated interval censored event data (curated and jittered of real adverse events data)



The graph above Figure 15 shows the estimates from three distributions using the conventional method on repeated interval censored event data. Compared with the plot obtained by non-parametric analysis, we can see that the three distributions fit well.

The slope is concave for the Weibull distribution and convex for the log-logistic distribution, except for the exponential distribution, which increases in a straight line.

Therefore, the log-logistic distribution has the most similar slope shape to non-parametric analysis. When checking the final value of the cumulative hazard estimates, the non-parametric analysis showed values between 1.5 and 2.0, and the plots obtained from the three distributions showed that the log-logistic distribution was the most similar,

The exponential distribution estimated a value of 2.2, the Weibull distribution estimated a value of 2.0, and the log-logistic distribution estimated a value slightly below 2.0.

Figure 16. Simulation plots for three distribution model using the new method in NONMEM for repeated interval censored event data (curated and jittered of real adverse events data)

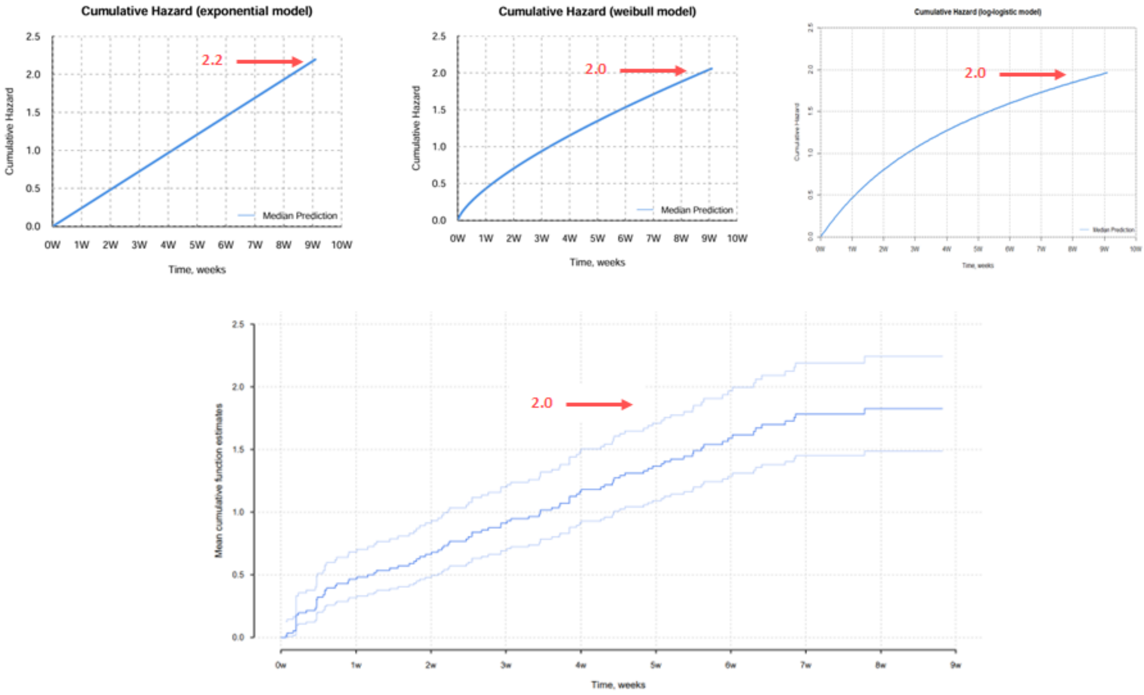


Figure 16 above shows the cumulative hazard plots obtained by applying the new method to three distributions in repeated interval censored event data. Compared to the graphs obtained by non-parametric analysis, the graphs obtained from the three distributions show a similar good fit.

Looking at the slope and final values of the cumulative hazard estimates, we can infer that the log-logistic distribution, which increases gently and has a final value close to 2.0, is the best-fitting model, as are the results from the conventional method.

The comparison between the estimates of the parameters obtained by NONMEM and the value of the objective function by applying the conventional method and the new method in repeated interval censored events, curated and jittered from real adverse event data, is shown in Table 13.

Table 13. The summary of results in both conventional method and new method in NONMEM for repeated interval censored events data (curated and jittered of real adverse events data)

Distribution	Estimator		NONMEM	
			Conventional method	New method
Exponential	Objective function value		805.975	802.334
	Lamda(λ)	Estimate	1.31E-03	1.33E-03
		S.E	1.72E-04	1.79E-04
Weibull	Objective function value		793.754	789.517
	Lamda(λ)	Estimate	1.65E-03	1.71E-03
		S.E	3.28E-04	3.58E-04
	Gamma(γ)	Estimate	7.09E-01	7.02E-01
		S.E	6.90E-02	7.07E-02
Log-logistic	Objective function value		793.934	789.684
	Lamda(λ)	Estimate	3.18E+02	3.04E+02
		S.E	7.95E+01	7.70E+01
	Gamma(γ)	Estimate	1.06E+00	1.06E+00
		S.E	1.09E-01	1.02E-01

The implementation of the NONMEM code is slightly different, but since no parameters have been added, the numerical comparison of the objective function values for both methods is not significant. The objective function values obtained by the new method are smaller than those obtained by the conventional method for all distributions.

When all distributions are run in both methods, minimization is successful and when a Wald test is performed on the estimates of each parameter, they all fall within 2 S.E. (standard error), indicating that the estimates are valid.

If all the estimates are valid, they can be validated by the graph, and the comparison of the implemented graphs with the estimates obtained from the two methods shows that both methods provide a good fit. The distribution with the best graph behavior was the log-logistic distribution for both methods.

As the fitted plots obtained from the conventional method and the new method were similarly good fits, the standard errors of the estimates were used to compare the accuracy. When comparing the standard error (S.E.) values of the estimated parameter values, the S.E. of the new method estimates is slightly smaller than that of the conventional method.

The meaning of the S.E. is the accuracy of each estimated parameter. Therefore, we can see that the parameter estimates obtained by the new method are slightly more accurate.

Discussion

TTE modelling is a popular method in survival analysis. The single event per subject method of TTE is widely accepted in both parametric and non-parametric analysis. Kaplan-Meier analysis is widely used as a non-parametric method and parametric models with different distribution assumptions such as exponential, Weibull and log-logistic distribution as a parametric method. In contrast to single TTE analysis, there is a need to improve the analysis method for repeated TTE.

There are a lot of repeated TTE data, such as adverse events and disease recurrence, and it is not reasonable to apply the method of single TTE analysis to them, as is often done. For these repeated events, the risk of occurrence over time is often clinically important. In this paper we consider how to analyse repeated events using non-parametric and parametric methods when the exact time of the event is known and when only the time range of the event is known (interval censored).

For non-parametric methods, we have shown how to use mean cumulative function estimates, which determine the average risk of an event occurring in a population over time, instead of Kaplan-Meier estimates, which determine the probability that an event has never occurred in a population over time. We believe this is more appropriate for repeated events because it is more important to know the risk of an event occurring at a given time than the probability that the event has never occurred up to that time.

A common approach is to calculate Kaplan-Meier estimates by considering only the first event in a series of repeated events. However, this tends to underestimate the overall cumulative hazard estimates because it does not correctly predict the rate of occurrence of cumulative events.

Kaplan-Meier estimates assume that all subjects have the same probability of survival and that the number of subjects without an event is a random variable, following a binomial distribution. In addition, the hazard function of the Kaplan-Meier estimates calculates the probability of an immediate event occurring per unit time in each interval.

On the other hand, for the mean cumulative function estimates, the number of events occurs in each interval is assumed to be independent of each other, and a random variable, following a non-homogeneous Poisson process. Thus, we calculated the probability of an event occurring instantaneously per unit subject in each interval, not per unit time.

Kaplan-Meier estimates focus on the occurrence of an event at unit time, and thus on the probability of no event occurring l in the population of subjects with the same probability of event occurrence. However, mean cumulative function estimates focus on the occurrence of events per unit subject in each time interval, and are concerned with how many events occur at each time point within each independent time interval in the overall population.

Thus, Kaplan-Meier estimates could be applied to single event per subject data, such as survival data or treatment response, etc, whereas mean cumulative function estimates are recommended for repeated event per subject data that predict the risk rate over time, such as adverse events or recurrence analysis, etc.

For parametric methods, we present a comparison between current commonly available methods and a new method that redefines each survival and hazard function to reflect the nature of repeated events for three representative distribution models: exponential, Weibull, and log-logistic.

The main difference between the current method and the newly proposed method is how the hazard function is defined for both repeated exact time and repeated interval censored events. While the survival function, calculated as a conditional probability based on the occurrence of a previous event or the endpoint of a previous event interval, was the same in both cases between the current and new methods, the hazard function was different.

Repeated exact time events should have been calculated as a conditional hazard function with the condition that the event occurs after the time of the previous event, but in the current, existing method it was defined as a general hazard function without adding such a condition. As a result, we compared the results of non-parametric analysis using rat recurrence data as an example with the estimates obtained from each of the three distributions, and found that the new method fit better than the conventional method.

In the case of repeated interval censored events, the current method also added the condition that the event occur after the end of the interval in which the previous event occurred, but the condition of the original hazard function in the conventional method was fixed as the starting point of the interval to calculate the cumulative probability of the event occurring.

The new method reflects this by modifying the original hazard function condition to reflect the time of each event, rather than the starting point of the interval, to calculate the cumulative probability of event occurrence. Thus, when integrating the probability of an event occurring immediately at each time point in the time interval of interest, the conventional method fixes both conditions of the hazard function, while the new method fixes only one condition (that an event occurs after the end of the interval in which the previous event occurred) and doesn't fix the other condition (that an event has not yet occurred at each time point in the current interval). We compared the results of the non-parametric analysis of the curated and jittered real adverse event data with the estimates obtained from each of the three distributions and found that the new method was broadly similar to the conventional method, but the estimates from the new method had smaller standard errors, indicating higher precision.

The current study has the following limitations. First, the study was not accompanied by a sensitivity analysis to further investigate the differences between the two methods. In the future, we will conduct a sensitivity analysis of the new method to see in which cases it can be recommended. We also plan to validate the proposed values of the new method by applying it to more diverse examples.

Secondly, in the case of repeated interval censored events, we only analysed the non-overlapping intervals and did not analyse the overlapping cases. Considering that most of the data obtained in clinical settings are non-overlapping interval censored data, the method proposed in this paper may be sufficient, but such a method may be necessary when analysing multiple adverse events in an integrated manner. In the future, we will further investigate how to perform non-parametric and parametric analyses on overlapping interval censored data.

Third, the application of the new method needs to try more repeated events data. Since we have investigated the new method and proved it by deriving the exact probability and likelihood expressions through algebraic process literally, we need to apply the more repeated events real data for validation and find more other insights.

In terms of number of events per subject, it is generally accepted that outcomes related to treatment response, such as death or relapse, are identified as single events per subject, whereas recurrence at multiple sites and adverse events are often identified as recurrent events with regular interval follow-up.

Also, from the perspective of time of event occurrence, in clinical trials conducted in hospitals with 24-hour monitoring, the exact time of the event can be confirmed, but in most clinical trials and real-world data, the exact time of the event is often not available to the researcher/physician unless the subject records or reports the time of the event during the follow-up interval, which is categorized as an interval-censored event.

Despite the large number of repeated events in clinical settings, the methods currently used reflect the repeated application of a single event or probability formulas that do not properly reflect the propensity for repeated events.

We expect that the new method proposed in this study will help to properly evaluate the efficacy and safety of new drugs by accurately assessing treatment outcomes and adverse events of new drugs from different data sources with different characteristics as described above, and by enabling accurate predictions in different scenarios.

References

1. Graziella D'Arrigo, Daniela Leonardis. Methods to Analyse Time-to-Event Data: The Kaplan-Meier Survival Curve. Oxid Med Cell Longev. 2021; 2021: 2290120.
2. E. L. Kaplan, Paul Meier. Nonparametric Estimation from Incomplete Observations, Journal of the ASA, 53:282, 457-481
3. Nelson, Wayne. Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications, ASA-SIAM, 2003.
4. Shane G. Henderson. Estimation for Nonhomogeneous Poisson Processes from Aggregate Data, 2022.
5. Wei Yang, Christopher Jepson. Statistical Methods for Recurrent Event Analysis in Cohort Studies in CKD, Clin J Am Soc Nephrol. 2017 Dec 7;12(12):2066-2073.
6. Ørnulf Borgan. Nelson′s-Aalen Estimator, Encyclopedia of Biostatistics, 15 July 2005
7. Jennifer Rogers. The Analysis of Recurrent Events : A summary of Methodology, Department of Statistics, University of OXFORD, 13th September 2016
8. Nick holford. A Time to Event Tutorial for Pharmacometricians, CPT: Pharmacometrics & Systems Pharmacology (2013) 2, e43
9. David Collett. Modelling Survival Data in Medical Research, Third Edition, 2015; 13: 978-1-4398-5678-9
10. Quyen Thi Tran, Jung-woo Chae, Kyun-Seop Bae. A simple time-to-event model with NONMEM featuring rigth-censoring, Transl Clin Pharmacol. 2022 Jun;30(2):75-82
11. Hyeong-Seock Lim. Brief indtrodunction to parametric time to event model. Transl Clin Pharmacol. 2021 Mar;29(1):1-5
12. Marc Lavielle. Mixed Effects Models for the Population Approach, May 2014
13. Huiru Dong. Estimating the Burden of Recurrent Events in the Presence of Competing Risks: The Method of Mean Cumulative Count, Am J Epidemiol. 2015;181(7):532–540

국문 요약

소개: 비모수적/모수적 사건 발생 시간(TTE) 접근법은 약물 개발을 위한 임상시험 데이터를 분석할 때 널리 사용됩니다. 약물 치료에 대한 반응, 환자 예후 및 약물 부작용을 평가하는 데 사용됩니다.

TTE 모델링 분석의 인기는 시간에 따른 특정 사건의 진행을 이해할 수 있기 때문에 시뮬레이션을 통해 다양한 약물 투여 시나리오에 따른 장기적인 결과를 예측할 수 있기 때문입니다.

목표: 각 사례에서 사건 발생까지의 시간을 분석하는 다양한 방법을 고려할 때, 장기적으로 다양한 임상 환경에서 사건의 발생 시간과 빈도를 예측하려면 정확한 분석이 필수적입니다.

이 연구는 사건을 특성화하고 적절한 확률 분포 함수를 설정한 다음 다양한 사건 발생 시간 데이터에 대한 개선된 방정식을 통해 NONMEM 에서 구현함으로써 적절한 방법의 사용을 안내하는 것을 목표로 합니다.

방법: 다양한 유형의 사건 시간 데이터에 대한 확률 대수 방정식을 검토하고 데이터 유형의 특성을 반영하여 새로운 수식을 도출했습니다. 이렇게 새롭게 도출된 방정식을 기반으로 단일 또는 반복 사건까지의 시간, 정확한 시간 또는 간격 검열 시간 등 각각의 일반적인 데이터 세트에 대해 모수적 사건 발생 시간 분석을 수행했습니다. 이러한 모수적 분석 결과를 비모수적 평균 누적 함수 추정치와 비교했습니다.

결과: 사건의 정확한 시간을 알고 있는지, 아니면 사건의 시간 범위만 알고 있는지, 환자 내 발생 횟수를 기준으로 단일 또는 반복 사건과 정확한 시간 또는 간격 검열 시간으로 사건까지의 시간을 분류했습니다. 사건 데이터 세트의 각 유형에 대한 확률 분포(생존 함수, 위험 함수, 누적 위험 함수)에 대한 방정식은 반복된 정확한 시간 사건, 반복된 간격 검열 사건과 같은 사건 데이터에 대해 새롭게 도출되었습니다.

그런 다음 새로 도출된 방정식을 NONMEM®에서 구현하여 임의의 가상 데이터 세트에 맞는 모수들을 추정하여, NONMEM®에서 기존에 사용하던 모델의 결과와 비교했습니다.

몬테카를로 시뮬레이션 결과, 새로 도출된 모델이 기존 모델보다 평균 누적 함수 추정치로 대표되는 예시 데이터 세트를 더 잘 설명하는 것으로 나타났습니다.

임의의 가상 데이터 세트에서 제공된 데이터 세트에 적용한 NONMEM®시뮬레이션 결과는 개선된 방정식의 적합성을 입증하고 기존 방정식과의 뚜렷한 차이점을 강조했습니다.

결론 본 연구에서 제시된 새로운 방정식을 사용하면 임상시험용 의약품의 치료 결과 및 이상 약물 반응과 같이 다양한 특성을 가진 다양한 출처의 사건 발생까지의 시간 데이터를 평가할 수 있을 것으로 기대합니다. 본 연구 결과는 의약품의 효능과 안전성을 적절히 평가하는 데 도움이 될 것입니다.