MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1611

Published online 11 February 2014 in Wiley Online Library

Analysis of safety data in clinical trials using a recurrent event approach

Qi Gong,[a] Barbara Tong,[b] Alexander Strasak,[c] and Liang Fang[d]*

As an important aspect of the clinical evaluation of an investigational therapy, safety data are routinely collected in clinical trials. To date, the analysis of safety data has largely been limited to descriptive summaries of incidence rates or contingency tables aiming to compare simple rates between treatment arms. Many have argued that this traditional approach failed to take into account important information including severity, onset time, and multiple occurrences of a safety event. In addition, premature treatment discontinuation due to excessive toxicity causes informative censoring and may lead to potential bias in the interpretation of safety events. In this article, we propose a framework to summarize safety data with mean frequency function and compare safety events of interest between treatments with a generalized log-rank test, taking into account the aforementioned characteristics ignored in traditional analysis approaches. In addition, a multivariate generalized log-rank test to compare the overall safety profile of different treatments is proposed. In the proposed method, safety events are considered to follow a recurrent event process with a terminal event for each patient. The terminal event is modeled by a process of two types of competing risks: safety events of interest and other terminal events. Statistical properties of the proposed method are investigated via simulations. An application is presented with data from a phase II oncology trial. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: log-rank test; mean frequency function; informative censoring; terminal event; benefit–risk assessment

1. INTRODUCTION

In contemporary clinical trials of therapeutic agents, safety data are routinely collected, and frequently, safety evaluation constitutes a critical aspect of the clinical trial. It is estimated that 70–80% of information collected in clinical trials are related to safety measures [1]. In contrast, research on safety data analysis methods remains quite limited to date. As stated in the 'ICH E9: Statistical Principles for Clinical Trials' (www.ich.org), the most common method to analyze safety data is to calculate and compare the crude incidence rate of patients who experience a specific safety event. In some cases, univariate contingency table analysis, including Chi-square or Fisher's exact test, is additionally performed for statistical inference. Although appealing for simplicity and ease of communication, this approach ignores several aspects of safety data and thus may result in inconclusive, limited information regarding the safety profile of an intervention.

In a clinical trial, safety data are usually collected during a prespecified time period, for example, from enrollment until 30 days after the last treatment dose, with different safety parameters including adverse events (AEs) and laboratory tests. One major challenge in safety data analysis thus is to appropriately reduce data dimensionality in analyses while still providing an accurate and relevant characterization of the safety profile for a therapeutic agent. In addition, as every safety event is multifaceted and measured with onset time, severity, and frequency, it is important to consider not only its rate of incidence but also information on timing, severity, and frequency, in order to comprehensively evaluate the impact on patients receiving the investigational therapy. Finally, informative censoring due to patients discontinuing treatment early may introduce an additional source of bias if not handled properly in analyses [2]. The aforementioned challenges in safety data analysis were recognized but only partially addressed in the literature. Amit *et al*. [3] proposed graphic approaches to analyzing multifaceted safety data. Berry and Berry [4] addressed the multiplicity issue in assessing drug safety with a three-level hierarchical mixture model. Herein, we propose a new framework to analyze safety data with a recurrent event approach. This article primarily focuses on analyzing cumulative incidence rate rather than duration of safety events.

In safety data analysis, one can consider safety events as multiple categories of recurrent events with a terminal event. Figure 1 displays an example of formulating safety data into a recurrent event analysis framework in an oncology clinical trial. Specifically, safety data may be classified into a set of prespecified categories of events of interest, such as grades 1–4 peripheral neuropathy, grades 2–4 fatigue, and so on, each following a recurrent event process. The terminal event is the discontinuation of study treatment, which typically marks the end of main safety data collection period in a clinical trial. In oncology studies where patients are treated until disease progression, the reason for treatment discontinuation may be one of the following: disease progression, death, patient's or physician's decision to withdraw from

[a] *Amgen Inc., South San Francisco, CA, USA*

[b] *Genentech Inc., South San Francisco, CA, USA*

[c] *F. Hoffmann-La Roche Ltd, Basel, Switzerland*

[d] *Gilead Sciences Inc., Foster City, CA, USA*

**Correspondence to: Liang Fang, Gilead Sciences, Inc., 333 Lakeside Dr., Foster City, CA 94404, USA.*
E-mail: liang.fang@gilead.com

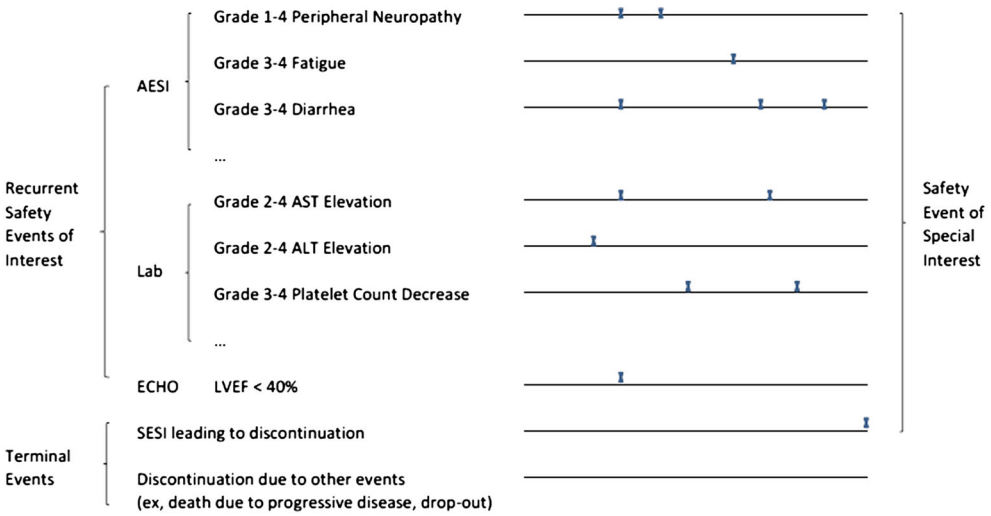


Figure 1. An oncology example of analyzing safety data under recurrent event framework. AESI, adverse events of special interest; ECHO, echocardiogram; SESI, safety event of special interest.

treatment because of nontoxicity reasons, or an AE that leads to treatment discontinuation. Based on the reason for treatment discontinuation, terminal events can be separated into safety events of special interest and other terminal events, and should be treated differently in the recurrent event analysis. Safety data from non-oncology trials can be formulated into this framework in a similar fashion.

Statistical methods for recurrent event analysis have been continuously developed in the past two decades [5–7] and, in more recent years, accommodate informative censoring due to terminal events. Cook and Lawless [8] developed a nonparametric estimate of mean frequency function of recurrent events. Ghosh and Lin [9] derived generalized log-rank tests to compare two mean frequency functions. Chen and Cook [10] extended the generalized log-rank test to analyze multiple types of recurrent events with a multivariate test. Nishikawa *et al*. [3] applied competing risk analysis to safety data to address informative censoring, although only considering the first event of a sequence of possible recurrent events. The fact that a safety event can occur repeatedly was recently addressed by Guttner *et al*. [11] in a stratified Cox regression model for recurrent events.

While previous work in the field of recurrent event analysis mostly treated all terminal events equally in correcting for bias due to informative censoring, we believe it is necessary in safety data analysis to differentiate two types of terminal events: safety events of interest that not only terminate the process but also are included in the comparison of safety profiles between treatment groups versus other events that are only included to correct bias due to informative censoring. For example, if the main objective of safety analysis is to compare all hepatotoxicity related events between two treatment groups, the hepatotoxicity related events leading to treatment discontinuation or death are included in the analysis for both treatment comparison and bias correction of informative censoring purposes, while other events leading to treatment discontinuation or death are only included to correct bias due to informative censoring.

In this article, we extend the methods in [8–10] to handle the two types of terminal events differently when analyzing safety data with a recurrent event analysis approach. The specific objectives of this article are (1) to introduce mean frequency function as a viable measure of safety profile under the framework of recurrent event analysis; (2) to formulate a generalized log-rank test to compare the mean frequency function of safety events of interest between two groups; and (3) to investigate a multivariate test for comparison of the global safety profile between two treatment groups. Section 2 presents the notations and methods of recurrent event analysis in safety data analysis. In Section 3, a simulation study is performed to investigate the inferential properties of the proposed statistics. Finally, the proposed methods are applied to a real dataset in Section 4. Discussion and summary are provided in Section 5.

2. NOTATIONS AND METHODS

2.1. Mean frequency function and cumulative incidence rate

Let D_i^{SI} and D_i^{O} denote the time of termination (e.g., treatment discontinuation) caused by a safety event of interest and other reasons, respectively. These two types of terminal events are often correlated with recurrent safety events, and only one type of terminal events can occur for subject i. Let $D_i = \min(D_i^{SI}, D_i^{O})$ be the time of treatment discontinuation for subject i. Denote $N_{ki}^*(t)(k = 1, \ldots, K)$ as the number of safety events of the k^{th} category that occur up to time t for subject i. Let the first K-1 categories of events be the recurrent events and $N_{Ki}^*(t)$ be the number of terminal event for subject i, up to time t, for which the value can only be 0 or 1. In Figure 1, safety events in category 1 to K-1 are the AEs of special interest, lab events of special interest, and ECHO events, while the terminal events of interest (category K) are the events leading to treatment discontinuation that are of interest for treatment evaluation. Denote censoring time as C_i. The observed data are $\{X_i, N_{ki}(\cdot), \delta_i\}$, where $X_i = \min(D_i, C_i)$, $N_{ki}(t) = N_{ki}^*(t)I(X_i > t)$, and $\delta_i = I(D_i \leqslant C_i)$. Let $Y_i(t) = I(X_i \geqslant t)$ be the at-risk indicator. We also assumed that there exists a time τ such that $P(X_i \geqslant \tau) > 0$.

For the recurrent safety events of interest (i.e., $k = 1, 2, \ldots, K$-1 event category), let

$$\mu_k(t) = \int_o^t S(u) dR_k(u), \tag{1}$$

be the mean frequency function, where $S(t)$ is the survival function of time to D_i, and $R_k(t)$ is the cumulative incident rate function for the safety events of category k. $\mu_k(t)$ can be estimated by $\hat{\mu}_k(t) = \int_o^t \hat{S}(u)d\hat{R}_k(u)$, where $\hat{S}(t)$ is the Kaplan–Meier estimator of terminal event, and $\hat{R}_k(t)$ is the Nelson–Aalen estimator for $R_k(t)$. The Nelson–Aalen estimator is a nonparametric estimator of the cumulative hazard rate function for censored or incomplete data. The Nelson–Aalen estimator is given by $\sum_{t_i \leqslant t} \frac{d_i}{n_i}$, with d_i as the number of events at t_i and n_i the total number of subjects at risk at t_i. The mean frequency function can be interpreted as the mean cumulative number of recurrent events up to a time point t.

For the terminal event of interest, similarly,

$$\mu_K(t) = \int_o^t S(u)dR_K(u), \tag{2}$$

where $R_K(t)$ is the cumulative incident rate function for the terminal event of interest. Similarly $\mu_K(t)$ can be estimated by $\hat{\mu}_K(t) = \int_o^t \hat{S}(u)d\hat{R}_K(u)$, where $\hat{R}_K(t)$ is the Nelson–Aalen estimator for $R_K(t)$.

Following Ghosh and Lin [9], we have

$$\sqrt{n}\left\{\hat{\mu}_k(t) - \mu_k(t)\right\} = n^{-1/2}\sum_{i=1}^{n}\psi_{ki}(t) + o_p(1),$$

where n is the sample size, $\psi_{ki}(t)$ are independent and identically distributed (*i.i.d.*) terms ($i = 1, \ldots, n$) with

$$\psi_{ki}(t) = \int_0^t \frac{S(u)dM_{ki}(u)}{\pi(u)} - \int_0^t \frac{\{\mu_k(t) - \mu_k(u)\}\, dM_i^D(u)}{\pi(u)}, \quad (k = 1, \ldots, K),$$

$$\pi(t) = P(Y_i(t) = 1),$$

$$M_{ki}(t) = N_{ki}(t) - \int_0^t Y_i(u)dR_k(u), \quad (k = 1, \ldots, K),$$

$$M_i^D(t) = N_i^D(t) - \int_0^t Y_i(u)d\Lambda^D(u).$$

Herein, $N_i^D(t)$ is the observed indicator of treatment discontinuation (due to a safety event of interest or other reasons) D_i, and $\Lambda^D(t)$ is the cumulative hazard function of D_i. Then, $\sqrt{n}\left\{\hat{\mu}_k(t) - \mu_k(t)\right\}, (k = 1, \ldots, K)$ converges weakly to a mean-zero Gaussian process with a covariance estimated by

$$\hat{\xi}_k(s,t) = n^{-1}\sum_{i=1}^{n}\hat{\psi}_{ki}(s)\hat{\psi}_{ki}(t),$$

where $\hat{\psi}_{ki}(t)$ is the empirical counterpart of $\psi_{ki}(t)$. To construct the pointwise confidence interval for $\mu_k(t)$, we chose the logarithm transform $\log\{\mu_k(t)\}$, because $\mu_k(t)$ is nonnegative. The asymptotic distribution of $n^{1/2}[\log\{\hat{\mu}_k(t)\} - \log\{\mu_k(t)\}]$ is approximately equal to the one of $n^{1/2}\{\hat{\mu}_k(t) - \mu_k(t)\} / \mu_k(t)$ when $\mu_k(t) > 0$. Then, an approximate $(1-\alpha)$ confidence interval for $\mu_k(t)$ is given by

$$\hat{\mu}_k(t)\exp\left\{\pm n^{-1/2}z_{\alpha/2}\hat{\xi}_k^{1/2}(t,t)/\hat{\mu}_k(t)\right\}.$$

2.2. Generalized log-rank test

To compare the mean frequency functions between two groups, let $\mu_{kl}(t)$ be the mean frequency function of the k^{th} category of safety events ($k = 1, \ldots, K$) in the l^{th} group ($l = 0, 1$). Define covariate $z_i = 1$ for subjects in the treatment group and 0 for those in the control group. A log-rank test statistic to test the difference in the mean frequency function of the k^{th} category of safety events between the two groups can be constructed based on

$$Q_k = \sqrt{\frac{n_0 n_1}{n}}\int_0^\tau W(t)d\left\{\mu_{k1}(t) - \mu_{k0}(t)\right\}, \tag{3}$$

where n_l is the sample size of group l,

$$W(t) = \frac{\pi_0(t)\pi_1(t)}{\pi(t)},$$

$$\pi_l(t) = P(Y_i(t) = 1, z_i = l),$$

$$\pi(t) = \pi_0(t) + \pi_1(t).$$

By replacing the quantities with their empirical counterparts, we have

$$\hat{Q}_k = \sqrt{\frac{n_0 n_1}{n}}\int_0^\tau \hat{W}(t)d\left\{\hat{\mu}_{k1}(t) - \hat{\mu}_{k0}(t)\right\},$$

where $\hat{W}(t) = \frac{n_0(t)n_1(t)}{n(t)}\frac{n}{n_0 n_1}$ and $n_0(t) = \sum_{i=1}^{n} I(z_i = l)Y_i(t)$. Following Ghosh and Lin (2000), we can show that

$$\sqrt{n_l}\left\{\hat{\mu}_{kl}(t) - \mu_{kl}(t)\right\} = n_l^{-1/2}\sum_{i=1}^{n}\psi_{kli}(t)I(z_i = l) + o_p(1),$$

where $\psi_{kli}(t)$ are i.i.d. terms ($i = 1, \ldots, n$) with

$$\psi_{kli}(t) = \int_0^t \frac{S_l(u)dM_{kli}(u)}{\pi(u)} - \int_0^t \frac{\{\mu_{kl}(t) - \mu_{kl}(u)\}\, dM_i^D(u)}{\pi(u)}, \quad (k = 1, \ldots, K),$$

$$M_{kli}(t) = N_{kli}(t) - \int_0^t Y_i(u)dR_{kl}(u), \quad (k = 1, \ldots, K),$$

where N_{kli} is the count function for the k^{th} adverse event (AE) in the i^{th} patient of the l^{th} group, and R_{kl} is the cumulative incidence rate function.

The null hypothesis for testing the overall treatment effect on k^{th} types of safety events is

$$H_0 : \mu_{k0}(t) = \mu_{k1}(t), \quad 0 < t \leqslant \tau.$$

Under the null hypothesis, following Chen and Cook [10], $\hat{Q}_k$ asymptotically follows a normal distribution with mean zero and variance $\mathrm{var}(\hat{Q}_k)$, which is consistently estimated by

$$\mathrm{var}(\hat{Q}_k) = \frac{1}{n}\sum_{l=0}^{1}\frac{n_{1-l}}{n_l}\left[\sum_{i=1}^{n}\left\{\int_0^\tau \hat{W}(t)d\hat{\psi}_{kli}(t)\right\}^2 I(z_i = l)\right].$$

2.3. Multivariate log-rank test for global assessment

The null hypothesis for testing the overall treatment effect on all K types of safety events is

$$H_0 : \mu_{k0}(t) = \mu_{k1}(t), \quad k = 1, \ldots, K, \quad 0 < t \leqslant \tau. \tag{4}$$

Under the null hypothesis, following Chen and Cook [10], $\hat{Q} = (\hat{Q}_1, \ldots, \hat{Q}_K)'$ asymptotically follows a multivariate distribution with mean zero and covariance matrix Σ, where the (j, k) entry of Σ is consistently estimated by

$$\mathrm{cov}(\hat{Q}_j, \hat{Q}_k) = \frac{1}{n} \sum_{l=0}^{1} \frac{n_{1-l}}{n_l} \left[\sum_{i=1}^{n} \left\{ \int_0^\tau \hat{W}(t) d\hat{\psi}_{jli}(t) \int_0^\tau \hat{W}(t) d\hat{\psi}_{kli}(t) \right\} I(z_i = l) \right].$$

Let $\bar{Q}_k = \hat{Q}_k / \sqrt{\mathrm{var}(\hat{Q}_k)}$ be the standardized form of the test statistics, $\hat{Q}_k$. A weighted test statistics could be obtained by

$$\hat{Q}_\omega = \omega_1 \bar{Q}_1 + \omega_2 \bar{Q}_2 + \ldots + \omega_K \bar{Q}_K, \tag{5}$$

where the weights satisfy $\omega_1 + \omega_2 + \cdots + \omega_K = 1$. Let $\omega = (\omega_1, \omega_2, \cdots, \omega_K)'$. The variance estimate can be obtained by

$$\mathrm{var}(\hat{Q}_\omega) = \omega' \left\{ diag(\hat{\Sigma}) \right\}^{-1/2} \hat{\Sigma} \left\{ diag(\hat{\Sigma}) \right\}^{-1/2} \omega.$$

The weighs, ω, can be determined in various ways based on clinical context. Often times, safety events are assessed with a scale of severity. For example, in an oncology clinical trial setting, the AEs and laboratory events are classified into grades 1–5 per National Cancer Institute-Common Terminology Criteria for Adverse Events (NCI-CTCAE) grading system. Grade 1 represents mildest in severity, and Grade 5 means highest severity, that is, death due to toxicity. In non-oncology clinical trials, similar grading systems exist. Therefore, to account for severity of the safety events in the evaluation of a treatment effect, one can weight different grades of events according to their numeric grade. Importantly, because the determination to weights involves a level of subjectivity that reflects the researcher's belief on different types of safety events, the outcomes of the multivariate log-rank test should be confirmed with sensitivity analyses, for example, by varying weight assignments.

3. SIMULATION STUDY

We investigated the statistical performance of the proposed methods in a simulation study. Without loss of generality, we simulated four categories of recurrent safety events of interest and two categories of terminal events: treatment discontinuation due to one of the four safety events and treatment discontinuation due to other reasons. In other words, we set K as 5 in (1) and (2).

The recurrent events were assumed to follow a Poisson process with rate

$$\lambda_{ki}(t|v_i, z_i) = \lambda_{k0}(t) \exp\left(z_i' \beta_k + v_i\right), \quad i = 1, \ldots, n, \quad k = 1, \ldots, 4,$$

where $\lambda_{k0}(t)$ was a baseline hazard function. v_i modeled the subject-level heterogeneity in the rate of safety events among subjects within the same treatment group as well as the correlations among the four categories of recurrent safety events. Time to terminal event of interest, denoted as D_i^5, was generated from an exponential distribution with hazard function

$$\lambda_{5i}(t|u_i, z_i) = (\lambda_{50} + u_i) e^{z_i \beta_5} / 2$$

and survival function

$$S^5(t|u_i, z_i) = \exp\left\{ -(\lambda_{50} + u_i) e^{z_i \beta_5} t / 2 \right\},$$

where z_i was 1 for treatment and 0 for control, and β_5 was the treatment effect on D_i^5. Without loss of generality, we also generated time to treatment discontinuation due to other reasons, D_i^O, from the same distribution, with hazard function

$$\lambda_i^O(t|u_i, z_i) = (\lambda_{50} + u_i) e^{z_i \beta_5} / 2$$

and survival function

$$S^O(t|u_i, z_i) = \exp\left\{ -(\lambda_{50} + u_i) e^{z_i \beta_5} t / 2 \right\}.$$

Therefore, the time to treatment discontinuation for subject i $D_i = \min(D_i^5, D_i^O)$ followed an exponential distribution with survival function

$$S(t|u_i, z_i) = \exp\left\{ -(\lambda_{50} + u_i) e^{z_i \beta_5} t \right\}.$$

u_i represented the heterogeneity in the treatment discontinuation rate among subjects in the same group and also the correlation between two types of terminal events. The association between recurrent events and terminal event was described by $\rho = corr(u_i, v_i)$.

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right).$$

Because of the shared frailties, the death event is an informative censoring to safety events, which mimics the real data. One of the advantages of our approach is that the proposed estimator accounts for informative censoring by including the survival function of the terminal events. The independent censoring time C_i was generated from an exponential distribution with mean a, the value of which varied to achieve different percentages of censoring. The earlier parameter setting allowed us to derive the true values of $\mu_{kl}(t)$ analytically as the following:

$$\mu_{kl}(t) = \int_{-\infty}^{-\infty} \int_{-\infty}^{-\infty} \mu_{kl}(t|u_i, v_i) f(u_i, v_i) du_i dv_i, k = 1, \ldots, 5, l = 0, 1,$$

with respect to (u_i, v_i), where

$$\mu_{kl}(t|u_i, v_i) = \int_o^t S_l(u|u_i) dR_{kl}(u|v_i)$$
$$S_l(t|u_i) = S(t|u_i, z_i = l)$$
$$dR_{kl}(t|v_i) = \lambda_{ki}(t|v_i, z_i = l) dt.$$

In the simulation study, 200 patients were randomly assigned to treatment and control groups with an allocation ratio of 1:1. Without loss of generality, all times were capped at $\tau = 1$. Other parameter settings were as follows:

Table I. Percent bias (%) of mean frequency estimates in the simulation study.

				AE category, $\rho = 0.25$					AE category, $\rho = 0.75$				
β_1, β_2	β_3, β_4	β_5	Censor %	1	2	3	4	5	1	2	3	4	5
0	0	0	25	−0.1	0.1	−0.2	−0.4	0.5	0.1	−0.1	0.1	0.3	−0.1
ln(0.5)	0	0	25	0.3	0.2	−0.1	−0.3	−0.1	−0.1	−0.2	−0.2	0.1	−0.4
0	ln(0.5)	ln(0.5)	25	0.2	0.3	−0.1	0.5	−0.1	−0.1	0.2	−0.1	0.1	0.2
ln(0.5)	ln(0.5)	ln(0.5)	25	0.2	0.1	−0.6	0.3	−0.4	0.3	−0.1	0.1	−0.4	−0.2
0	0	0	50	0.3	0.1	−0.1	0.1	−0.3	0.2	−0.3	0.3	−0.3	0.2
ln(0.5)	0	0	50	−0.1	0.1	−0.1	−0.1	−0.1	−0.4	0.1	−0.4	0.5	−0.5
0	ln(0.5)	ln(0.5)	50	−0.1	−0.2	0.1	0.1	−0.2	0.1	0.3	−0.3	−0.2	−0.4
ln(0.5)	ln(0.5)	ln(0.5)	50	0.7	0.2	1.0	0.3	0.5	−0.1	−0.7	−0.1	−0.1	−0.1

Bias = 100% ∗ (estimated value − true value)/true value.
AE, adverse event.

Table II. Coverage probability (%) of the 95% confidence interval of the mean frequency estimates in the simulation study.

				AE category, $\rho = 0.25$					AE category, $\rho = 0.75$				
β_1, β_2	β_3, β_4	β_5	Censor %	1	2	3	4	5	1	2	3	4	5
0	0	0	25	94	93	93	93	95	94	95	94	94	96
ln(0.5)	0	0	25	93	94	93	94	93	95	94	95	94	95
0	ln(0.5)	ln(0.5)	25	98	98	97	97	97	98	97	97	97	98
ln(0.5)	ln(0.5)	ln(0.5)	25	98	97	97	98	98	98	98	97	98	98
0	0	0	50	93	93	92	93	95	89	89	89	88	91
ln(0.5)	0	0	50	89	91	91	89	91	91	88	91	90	92
0	ln(0.5)	ln(0.5)	50	96	96	98	97	98	97	97	97	97	97
ln(0.5)	ln(0.5)	ln(0.5)	50	96	98	98	97	97	97	97	97	96	98

AE, adverse event.

$$\sigma_1 = \sigma_2 = 0.3,$$
$$\lambda_{10}(t) = \lambda_{20}(t) = 8,$$
$$\lambda_{30}(t) = \lambda_{40}(t) = 4,$$
$$\lambda_{50} = 2.$$

In addition, we set $\beta_k = \ln(0.5),\ k = 1, \ldots, 4$ for treatment effect of recurrent events, $\beta_5 = \ln(0.5)$ for treatment effect for terminal event, and $\rho = 0.25$ or 0.75 for the correlation between the terminal and recurrent events. For the multivariate test statistics (5), two sets of weight were chosen to be

$$\omega_I = (\,1,\ 2,\ 3,\ 4,\ 5\,)'/15$$

$$\omega_{II} = (\,e,\ e^2,\ e^3,\ e^4,\ e^5\,)'/(e + e^2 + e^3 + e^4 + e^5).$$

For each set of simulation parameters, 1000 simulations are run.

Thus, the bias is calculated by $\hat{\mu}_{kl}(\tau) - \mu_{kl}(\tau)$. The $(1-\alpha)$% coverage probability is the percentage of true value $\mu_{kl}(\tau)$ falling in interval $\hat{\mu}_{kl}(\tau)\exp\{\pm n_l^{-1/2} z_{\alpha/2}\hat{\xi}_{kl}^{1/2}(\tau,\tau)/\hat{\mu}_{kl}(\tau)\}$. The type I error rate is calculated as the percentage of rejecting the null hypothesis based on the Z-test. That is, rejecting H_0 if $|\hat{Q}_k/\sqrt{\text{var}(\hat{Q}_k)}| > z_{\alpha/2}(k = 1, \ldots, 5)$ or $|\hat{Q}_\omega/\sqrt{\text{var}(\hat{Q}_\omega)}| > z_{\alpha/2}$. Again, the $(1-\alpha)$ confidence interval of $\mu_{kl}(\tau)$ is based on logarithm transform and is given by $\hat{\mu}_{kl}(\tau)\exp\{\pm n_l^{-1/2} z_{\alpha/2}\hat{\xi}_{kl}^{1/2}(\tau,\tau)/\hat{\mu}_{kl}(\tau)\}$. α is chosen to be 0.05.

Table I shows the bias percentage, defined as bias divided by true value, of the mean frequency estimate of categories 1–5 AE in the treatment arm under difference scenarios. The mean frequency estimate was in general unbiased, with estimated percent biases close to 0 (−0.7–1.0%) in the simulation study. Table II provides the coverage probability of the mean frequency estimate of categories 1–5 AEs in the treatment arm. In most cases, the coverage probabilities were close to the desired level 95%. When the censoring rate is high (50%), the estimated coverage probabilities were lower than the targeted 95% level by a range of 1–7% for a few scenarios. This is likely due to high censoring rate. When doubling the sample size, the coverage probability was very close to 95% (data not shown). Type I error rates of the generalized log-rank tests for both univariate and multivariate tests are provided in Table III. The type I error rates were controlled at 5% level for all cases with a highest value of 6.5%. Power of the generalized log-rank tests was also calculated and in line with our expectation (data not shown). In practice, the power can be obtained through a simulation study similar to ours because it is a factor of multiple factors, including size of treatment effect, number of events, mean frequency of recurrent and terminal events, and weighs of different types of safety events in the multivariate generalized log-rank test. In summary, we concluded that the derived asymptotic was valid.

Table III. Type I error rate of the generalized log-rank tests in the simulation study.

	$\rho = 0.25$		$\rho = 0.75$	
Generalized log-rank test	Censor, 25%	Censor, 50%	Censor, 25%	Censor, 50%
Univariate, category 1	4.2	4.4	5.2	5.0
Univariate, category 2	5.1	5.2	4.4	4.8
Univariate, category 3	4.7	5.2	5.6	6.5
Univariate, category 4	4.9	6.2	3.4	4.7
Univariate, category 5	4.8	4.4	6.4	5.5
Multivariate, weight 1	4.6	5.2	3.1	6.0
Multivariate, weight 2	5.1	5.9	3.2	6.2

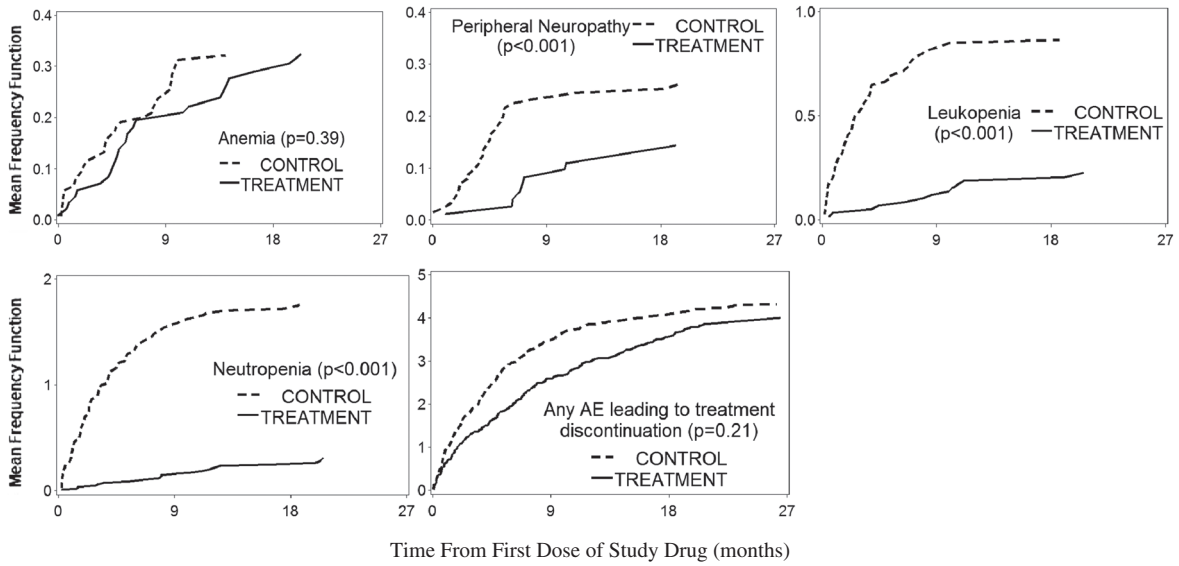


Figure 2. Mean frequency functions of the five types of safety events of interest in the example.

4. EXAMPLE

We applied the proposed methods to analyze a real dataset from an oncology clinical trial. This is a randomized two-arm study to compare an investigational agent to a standard-of-care therapy in patients with metastatic breast cancer. In this study, 130 patients were randomly assigned to receive treatment or control at a 1:1 allocation ratio. Safety data, including AEs, laboratory tests, and other safety parameters such as Electrocardiogram (ECG) and Left ventricular ejection fraction (LVEF) evaluations, were collected on a regular basis until 30 days after the last dose of study treatment. Reasons for treatment discontinuation could be death, disease progression, patient or physician decision, or excessive toxicity, whichever occurs first. Because of the trial sponsor's confidentiality policy, we masked the AE dataset used for this example by swapping treatment assignment for 10% of the study patients.

4.1. Primary analysis: chemotherapy-induced adverse events and adverse events leading to treatment discontinuation

In the safety analysis of this study, the primary questions of interest were whether the treatment arm had (1) less chemotherapy-induced AEs or (2) less AEs leading to treatment discontinuation, compared with the control arm. The key chemotherapy-induced AEs are as follows: anemia, peripheral neuropathy, neutropenia, and leukopenia. The AE collection period is from randomization date to 30 days after drug discontinuation. Treatment discontinuation due to AEs and that due to other reasons are the two types of terminal events.

Figure 2 and Table IV summarize the analysis results based on the proposed method. The treatment arm significantly reduced the frequencies of peripheral neuropathy, neutropenia, and leukopenia, while had similar frequencies of anemia and AEs leading to treatment discontinuation. Figure 2 displayed the pattern of the occurrence of these AEs. For patients in the control arm, most chemotherapy-induced AEs occurred within 9 months of treatment, which was approximately equal to the median duration of chemotherapy. The occurrence of chemotherapy-induced AEs in the treatment arm did not reveal any notable pattern and was generally low.

This example demonstrated that the mean frequency function and its graphic display as in Figure 2 were useful to reveal the multifaceted information and pattern of safety data, and the proposed generalized log-rank tests provided a way to compare the safety data between two groups, adjusting for informative censoring.

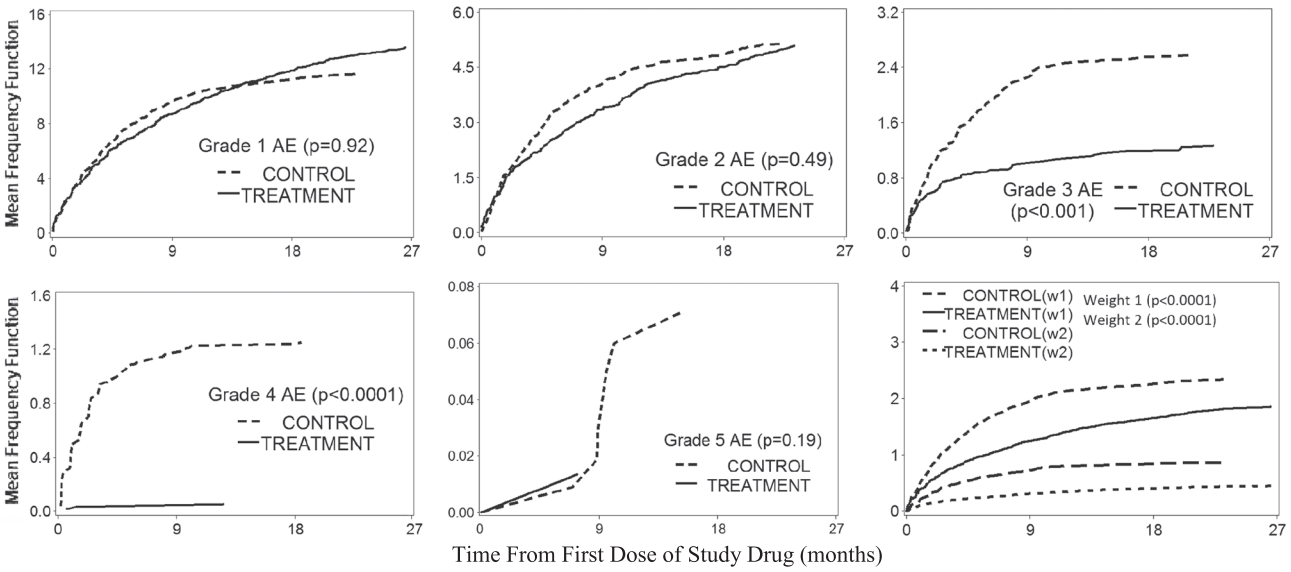
The multivariate test was not preformed since assigning weights to different chemotherapies was not clinically interpretable, and the primary interest was to investigate each type of chemotherapies.

Table IV. Summary of analysis results in the example.

	Control arm		Treatment arm		Difference		*p*-value	
AE category	Mean frequency	95% CI	Mean frequency	95% CI	Mean frequency	95% CI	Mean frequency	95% CI
Anemia	0.3	(0.3, 0.3)	0.3	(0.3, 0.3)	0.0	(−0.1, 0.1)	0.39	Not performed
Peripheral neuropathy	0.3	(0.3, 0.3)	0.1	(0.1, 0.2)	−0.1	(−0.2, −0.0)	< .001	
Leukopenia	0.9	(0.8, 0.9)	0.2	(0.2, 0.2)	−0.7	(−0.9, −0.4)	< .001	
Neutropenia	1.8	(1.7, 1.8)	0.3	(0.3, 0.3)	−1.5	(−1.9, −1.0)	< .001	
AE leading to treatment discontinuation	4.3	(4.2, 4.4)	4.0	(3.8, 4.2)	−0.3	(−1.9, 1.3)	0.21	
Grade 1	11.7	(11.3, 12.0)	13.6	(13.0, 14.2)	1.9	(−3.4, 7.2)	0.92	< 0.001*
Grade 2	5.1	(5.0, 5.3)	5.1	(4.9, 5.3)	0.0	(−2.0, 2.0)	0.49	< 0.001†
Grade 3	2.6	(2.5, 2.6)	1.3	(1.2, 1.3)	−1.3	(−2.0, −0.6)	< .001	
Grade 4	1.3	(1.2, 1.3)	0.1	(0.1, 0.1)	−1.2	(−1.5, −0.9)	< .001	
Grade 5	0.07	(0.06,0.08)	0.01	(0.01, 0.02)	−0.06	(−0.1, 0.0)	0.19	

AE, adverse event.

*Based on weight set 1: $\omega_I = (1,\ 2,\ 3,\ 4,\ 5)'/15$.
†Based on weight set 2: $\omega_{II} = (e,\ e^2,\ e^3,\ e^4,\ e^5)'/(e + e^2 + e^3 + e^4 + e^5)$.



Weight 1: $\omega_I = (1,\ 2,\ 3,\ 4,\ 5)'/15$; Weight 2: $\omega_{II} = (e,\ e^2,\ e^3,\ e^4,\ e^5)'/(e + e^2 + e^3 + e^4 + e^5)$

Figure 3. Mean frequency functions of grades 1–5 adverse events (AEs) in the example.

4.2. Exploratory analysis: overall profile of adverse events and adverse events by severity

An exploratory objective of the safety analysis was to investigate whether the overall safety profile measured by the AE data was superior in the treatment arm. To answer this question, we applied the multivariate log-rank test as described in Section 2.3. All reported AEs were classified into one of the five categories per its NCI-CTCAE grade. Subsequently, increasing weights were assigned to the five AE groups with the lowest weight assigned to the group with NCI-CTCAE grade one, and the multivariate test comparing the two treatment groups was performed. To verify the robustness of the analysis outcomes, two sets of weights were applied. The analyses indicated that patients in the treatment arm experienced significantly less frequent AEs ($p < 0.001$, Table IV), regardless of the choices of weights. Thus, it was concluded that the overall profile of AEs was significantly better in the treatment arm.

In addition, other questions of interest included whether the treatment arm had significant reduction in different grades of AEs, how quickly the AEs occurred, and what was the general pattern of the AEs. Mean frequency functions were estimated as described in Section 2.1 for grades 1–5 events and are plotted in Figure 3. The generalized log-rank test as described in Section 2.2 was applied to each grade group to compare AEs of a particular grade between the two treatment arms. Patients in the treatment arm experienced significantly less frequent graded 3 and 4 AEs ($p < 0.001$) but similar frequency of grades 1, 2, and 5

AEs (Table IV, $p = 0.92, 0.49.0.19$, respectively). Patients in the treatment arm experienced, on average, 13.6 grade 1, 5.1 grade 2, 1.3 grade 3, 0.1 grade 4, and 0.01 grade 5 AEs, when they were treated with the drug. In the control arm, patients experienced an average of 11.7 grade 1, 5.1 grade 2, 2.6 grade 3, 1.3 grade 4, and 0.07 grade 5 AEs. This is as expected because most patients in oncology studies experienced at least one low grade AE given the underlying disease they had. The goal of new cancer therapy development is to reduce more severe AEs (i.e., grades 3–5). Even though, the difference of grade 5 AEs between the two arms was not statistically significant, and this is probably due to the fact that there were very few such fatal events in the study. The numerical difference (0.01 vs. 0.07) was observed. Another interesting pattern as shown in Figure 3 was that there seemed to be a jump on the number of grade 5 AEs in the control arm at the end of the study. Although it might be a random variability due to the small number of events, this observation warrants future investigations. In addition, Figure 3 also revealed the general pattern of the AEs. Most grades 1 and 2 AEs occurred within the first 12 months, while most grades 3–5 AEs occurred earlier. Important findings revealed by Figure 3 and Table IV highlight the advantages of the proposed method over the traditional and simplified approach of summarizing AE data with incidence rates.

5. CONCLUSION

Safety data are routinely collected in clinical trials and usually can be formulated into event data, such as AEs and laboratory abnormalities. Routine safety analysis often report frequency of patients who had at least one particular safety event. This analysis ignores many statistical issues such as informative censoring and recurrence of safety events. In recent years, noninferiority trials have become well accepted for drugs with comparable efficacy but more favorable safety profiles relative to the standard-of-care treatment. However, very few statistical testing procedures have been developed to compare the safety profiles of two groups. The need of a statistical test for safety data becomes apparent. Recently, a joint effort from health authorities, academia, and industry has been made to take the multifaceted features of safety data into account in benefit–risk assessment [12]. All aforementioned challenges highlight the need of statistical methods and applications in the analysis of safety data.

In this article, we attempted to address the aforementioned issues by introducing recurrent event analysis methods into safety data analysis. Methods developed in [8,9], and [10] were extended to accommodate two distinct types of terminal events: terminal events of interest and other terminal events. The first type of terminal events was counted in the calculation and hypothesis testing of mean frequency function of safety events, while the second type was also considered in the calculation but only to correct bias due to informative censoring. We proposed summarizing safety events with mean frequency function and comparing them with a generalized log-rank test. In addition, one can compare the global safety profile with the proposed multivariate log-rank test. The generalized univariate and multivariate log-rank tests allow researchers to compare the safety profiles between groups by formal hypothesis testing. Graphic display of the mean frequency function is very useful to reveal the multifaceted nature of safety data and allows researchers to investigate not only the frequency of the safety events but also the general pattern, such as onset time and rate of occurrence of the events.

The implementation of the proposed method is not trivial. Unfortunately, there is not readily available macro or program to overcome this hurdle. Thus, coding was performed in SAS® (SAS institute, Cary, North Carolina) proc iml to realize the estimation procedure explicitly. Simulation results showed that the mean frequency function estimator was generally unbiased with adequate coverage probability. The univariate and multivariate generalized log-rank tests maintained type I error rates at the desired level, and the power of the tests depended on several factors, including size of treatment effect, number of events, mean frequency of recurrent and terminal events, and weighs of different types of safety events in the multivariate generalized log-rank test. We also demonstrated our methods with a dataset from a real clinical trial.

A similar but simpler approach was taken to analyze cardiac risk of Herceptin® (Genentech Inc, South San Francisco, CA), and the analysis results are included in the package insert. In the Herceptin package insert, time to first cardiac event was analyzed taking into consideration of competing risk of deaths due to noncardiac events. Mean frequency function was presented for the two treatment arms, without statistical testing for treatment comparisons. This approach can be considered as a special case of our method.

Our proposed method provides a general framework of analysis of safety data with recurrent event methods. It also provides a way to conduct statistical testing of treatment effect on safety data as a whole. However, this approach can only be of practical value if the safety events of interest can be prespecified and well defined. This is somewhat a challenging task for drug development, especially when the safety profile of the drug is still under investigation. Knowledge on the mechanism of action of the drug, prior data from earlier phase trials or from drugs in a similar class, and recommendations and discussions with experts in the field and health authorities are helpful to define the safety events of interest. It is also noted that different methods to analyze data of recurrent and terminal events exist in the literatures [13–15]. Future work includes extending these methods to the analysis of safety data following our proposed framework. Although it is possible to formulate and conduct a test to compare the overall safety profile between two treatments, using the multivariate generalized log-rank test proposed, we only recommend such tests being pursued when there are clinical justifications for the choice of weights in the test and clinical meaningful interpretations of the test results.

Our method primarily analyzes cumulative incidence rate rather than duration of safety events. If the duration of safety events is of interest, one should refer to the nonparametric cumulative duration developed in [16]. We developed the method under the scenario that the safety reporting period ends at the same time for all safety data collections. In practice, there could be an additional follow-up period to collect important safety data, such as serious AEs or death, beyond the typical safety data collection period. The extension of our methods to accommodate different lengths of data collection periods for different types of safety data is straightforward and can be achieved by further classifying terminal events into smaller groups by different lengths of data collection periods. Furthermore, future applications on real clinical trial data and engagement with medical community are needed to further evaluate the practical value of the proposed method from a medical practitioner's perspective.

REFERENCES

[1] Chuang-Stein C, Mohberg NR, Musselman DM. Organization and analysis of safety data using a multivariate approach. *Statistics in Medicine* 1992; **11**:1075–89.
[2] Nishikawa M, Tango T, Ogawa M. Non-parametric inference of adverse events under informative censoring. *Statistics in Medicine* 2006; **25**:3981–4003.
[3] Amit O, Heiberger RM, Lane PW. Graphical approaches to the analysis of safety data from clinical trials. *Pharmaceutical Statistics* 2008; **7**:20–35.
[4] Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics* 2004; **60**:418–26.
[5] Anderson PK, Gill RD. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* 1982; **4**:1100–20.
[6] Prentice RL, Williams BJ. On the regression analysis of multivariate failure time data. *Biometrika* 1981; **68**:373–79.
[7] Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* 1989; **84**:1065–73.
[8] Cook RJ, Lawless JF. Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine* 1997; **16**:911–24.
[9] Ghosh D, Lin DY. Nonparametric analysis of recurrent events and death. *Biometrics* 2000; **56**:554–62.
[10] Chen BE, Cook R. Tests for multivariate recurrent events in the presence of a terminal event. *Biostatistics* 2004; **5**:129–43.
[11] Guttner A, Kubler J, Pigeot I. Multivariate time-to-event analysis of multiple adverse events of drugs in integrated analyses. *Statistics in Medicine* 2007; **26**:1518–31.
[12] Guo JJ, Pandey S, Doyle J, Bian B, Lis Y, Raisch D. A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy—report of the ISPOR risk-benefit management working group. *Value in Health* 2010; **13**:657–66.
[13] Liu L, Wolfe RA, Huang X. Shared frailty models for recurrent events and a terminal event. *Biometrics* 2004; **60**:747–56.
[14] Li QH, Lagakos SW. Use of the Wei-Lin_Weissfeld method for the analysis of recurring and a terminating event. *Statistics in Medicine* 1997; **16**:925–40.
[15] Ye Y, Kalbfleisch JD, Schaubel E. Semiparametric analysis of correlated recurrent and terminal events. *Biometrics* 2007; **63**:78–87.
[16] Wang J, Quartey G. Nonparametric estimation for cumulative duration of adverse events. *Biometrical Journal* 2012; **54**: 61–74.