Interim analysis of binary outcome data in clinical trials: a comparison of five estimators

Qing Shu Lu, Shein-Chung Chow & Siu-Keung Tse

To cite this article: Qing Shu Lu, Shein-Chung Chow & Siu-Keung Tse (2019): Interim analysis of binary outcome data in clinical trials: a comparison of five estimators, Journal of Biopharmaceutical Statistics, DOI: 10.1080/10543406.2018.1559852

To link to this article: https://doi.org/10.1080/10543406.2018.1559852

Published online: 01 Jan 2019.
Interim analysis of binary outcome data in clinical trials: a comparison of five estimators

Qing Shu Lu, Shein-Chung Chow, and Siu-Keung Tse

a Department of biostatistics, Singapore Clinical Research Institute, Singapore; b Center for Quantitative Medicine, Office of Clinical Sciences, Duke-NUS Graduate Medical School, Singapore; c Department of Biostatistics & Bioinformatics, School of Medicine, Duke University, Durham, USA; d Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong

ABSTRACT
In clinical trials, where the outcome of interest is the occurrence of an event over a fixed time period, estimation of the event proportion at interim analysis can form a basis for decision-making such as early trial termination, sample size re-estimation, and/or dropping inferior treatment arms. In addition to derivation of mean squared error under an exponential time-to-event distribution, we performed a simulation study to examine the performance of five estimators of the event proportion when time to the event is assessable. The simulation results showed advantages of the Kaplan–Meier estimator over others in terms of robustness, and the bias and variability of the event proportion estimate. An example was given to illustrate how the estimators affect dropping treatment arms in a multi-arm multi-stage adaptive trial. We recommended the use of the Kaplan–Meier estimator and discourage the use of other estimators that discard the inherent time-to-event information.

1. Introduction
In some clinical trials, the primary outcome could be proportion of subjects who experienced an event over a fixed time period rather than time to occurrence of the event. For example, in a smoking cessation study that evaluated the effectiveness of adding group-based treatment adjuncts to a smoking cessation intervention, smoking status was assessed at the conclusion of intervention, 6, 12 and 24 months post intervention, and the primary outcome was the abstinence rate at 24 months (Gruder et al. 1993). Another example is relapse of disease in cancer trials where the relapse can occur anytime during the follow-up. Such binary outcomes pose a unique issue in the estimation of event proportion at interim analysis as compared to other binary outcomes that are assessed only at the end of follow-up. For the latter, clearly subjects at ongoing follow-up do not contribute data to the estimation of event proportion; for the former, such subjects may or may not have experienced the event and including them in interim analysis will affect the estimate of event proportion.

Interim analysis forms a basis for decision-making during clinical trials. This is of critical importance, especially in long-term clinical trials or those of life-threatening diseases. Decisions based on interim analysis can be early trial termination for: (a) efficacy when it is evidenced that the experimental arm is clearly better than the control; (b) futility when it is unlikely to obtain a positive study outcome even the trial continues to its end; or (c) safety when a serious safety concern arises from adverse event data. Moreover, a planned adaptation could be made pertaining to other aspects of trial design such as sample size and number of treatment arms (FDA 2010).
In clinical trials, where an outcome is measured repeatedly at several fixed time points on each subject with the last one being the primary outcome, incorporating outcome data from ‘pipeline’ subjects who yet to observe the primary outcome is a possible way to increase the efficiency of interim analysis (Galbraith and Marschner 2003, Hampson and Jennison 2013, Stallard 2010). This is desirable especially in circumstances where limited information of the primary outcome is available at interim analysis and there are a large proportion of ‘pipeline’ subjects. The idea was reflected in the data analysis methods for binary outcomes that were reversible – for example, during a long-term treatment period of a disease, a favourable response achieved by a patient may subsequently become unfavourable (Marschner and Becker 2001; Sooriyarachchi et al. 2006; Whitehead et al. 2008), and that were irreversible (Parpia et al. 2014). Our study falls in the secondary category, i.e., the outcome being whether or not an event of interest occurs during a fixed period rather than time to occurrence of the event, e.g., relapse of disease in cancer trials in the example above. This kind of binary outcomes is associated with time-to-event information, and our study focus is on the situation where the binary outcome of the event rather than the time-to-event information is of primary interest. This shall be distinguished from a binary outcome with no inherent time-to-event information, which is not covered in this study. An example of such an outcome is sputum culture status (positive/negative) at the end of a 6-month standard therapy for pulmonary tuberculosis (TB).

Although three methods of estimating the event proportion at interim analysis have been evaluated in terms of type I error rate, power, and the probability of early stopping for efficacy in group sequential trials through extensive simulations (Parpia et al. 2014), there is a gap in knowledge regarding bias and variability of the estimates, which we attempted to fill. In this paper, results of a simulation study comparing five estimators of estimating the event proportion are presented, in addition to expressions of square root of mean squared error (RMSE) for exponentially distributed time-to-event data. The five estimators are: (1) the ‘pre-specified duration’ (PD) estimator – estimation based on only subjects who have been followed up for at least a pre-specified duration or in whom the event has occurred; (2) the ‘all randomised’ (AR) estimator – estimation based on all randomised subjects who have completed their treatment; (3) the Kaplan–Meier (KM) estimator; (4) the ‘all completed’ (AC) estimator – estimation based on all subjects who have completed follow-up; (5) the parametric (PR) estimator – estimation based on a parametric model of time to the occurrence of event, of which estimators 1–3 were considered in Parpia et al. (2014). The five estimators and their expressions of RMSE for exponentially distributed time to the event are described in Section 3. The results of simulations are presented in Section 3. An example is given in Section 4, followed by concluding remarks in Section 5.

2. Methods

Consider a randomised controlled trial where the treatment period is $T_0$ followed by a followed-up period of $T$ for each enrolled subject in the experimental and control arms, regardless of whether the event of interest occurs in the subject. An interim analysis is planned to occur when a pre-specified criterion is met, for example, when half of the target sample size has completed follow-up (Pedley 2011). Subject enrolment is ongoing at the time of interim analysis. The primary objective of the interim analysis is to estimate the event proportion (denoted by $\pi$), which is defined as probability of the event occurring at any time over the post-treatment follow-up period $T$. Without loss of generality, we formulate our research questions for the experimental arm. We denote by $r_1$ the number of events among the $n_1$ subjects in the experimental arm who have completed the treatment by the time of interim analysis, of which $r_2$ events are observed from the $n_2$ subjects who have completed follow-up ($r_2 \leq r_1$, $n_2 \leq n_1$) (We assume events occur only in the follow-up period. If the event of interest may occur during the treatment period, $n_1$ is the number of subjects that have been enrolled). Throughout this paper, we assume interim analysis is planned to occur when follow-up is completed in a pre-specified number of subjects $n_2$. We further assume the event
occurs in a subject no more than once. Therefore, the number of events is the same as the number of subjects in whom the event has occurred. In what follows, we describe five methods of estimating the event proportion \( \pi \).

### 2.1. The ‘pre-specified duration’ estimator

A seemingly plausible estimator of the event proportion is based on subjects who have been followed up for at least a pre-specified duration \( T_1 \) \( (T_1 < T) \) by the time of interim analysis. This estimator will lead to an underestimate as more events are expected if all these subjects had completed follow-up. One way to correct the underestimate is to include subjects who have experienced the event but have been followed up for less than duration \( T_1 \). In other words, the event proportion is estimated as the proportion of subjects with the event observed among those who have been followed up for at least \( T_1 \) or in whom the event has occurred:

\[
\hat{\pi}_{PD} = \frac{r_{12} + r_{11}}{n_{T_1} + r_{11}},
\]

where \( n_{T_1} \) is the number of subjects who have been followed up for at least \( T_1 \), of which there have been \( r_{12} \) events; \( r_{11} = r_1 - r_{12} \) is the number of events observed from the subjects who have been followed up for less than \( T_1 \). We call this the PD estimator. The estimate \( \hat{\pi}_{PD} \) can be formulated as a weighted average of \( r_{12}/n_{T_1} \) and 1:

\[
\hat{\pi}_{PD} = \omega \frac{r_{12}}{n_{T_1}} + (1 - \omega),
\]

where \( \omega = \frac{n_{T_1}}{n_{T_1} + r_{11}} \) is the weight. Since \( \frac{r_{12}}{n_{T_1}} \) underestimates the event proportion for \( T_1 < T \), so does \( \hat{\pi}_{PD} \) if \( r_{11} \) is very small as compared to \( n_{T_1} \) (i.e., \( \omega \approx 1 \)), which could happen when the fixed duration \( T_1 \) is too short. On the hand, \( \hat{\pi}_{PD} \) may overestimate the event proportion for a large value of \( T_1 \). An extreme example is \( T_1 = T \), under which \( \frac{r_{12}}{n_{T_1}} \) becomes an unbiased estimate of the event proportion. However, analytical forms for the bias and variance of \( \hat{\pi}_{PD} \) elude us.

### 2.2. The ‘all randomised’ estimator

The AR estimator bases the estimate of event proportion on the number of all randomised subjects who have started the follow-up (i.e., with treatment completed) by the time of interim analysis. In particular, the event proportion is estimated as:

\[
\hat{\pi}_{AR} = \frac{r_1}{n_1},
\]

Obviously, this underestimates the event proportion unless all randomised subjects have completed their follow-up and should be not used in practice. Nevertheless, it is interesting to evaluate how the extent of underestimate be affected by factors such as recruitment rate and distribution of time to occurrence of the event.

### 2.3. The ‘all completed’ estimator

Only subjects who have completed follow-up are included in the AC estimator, which is also known as the ‘reduced sample’ estimator (Kaplan and Meier 1958). The AC estimator of the event proportion is given as:

\[
\hat{\pi}_{AC} = \frac{r_2}{n_2}.
\]
Note that \( r_2 \) is the number of events in \( n_2 \) subjects for each of whom the probability of event is \( \pi \). That is, \( r_2 \) follows a binomial distribution with parameters \( n_2 \) and \( \pi \). Hence, \( \hat{\pi}_{AC} \) is an unbiased estimator of the event proportion \( \pi \). In some situations, however, the number of subjects who have completed follow-up is small at interim analysis. The implication is that the variability in the estimate could be substantial. Moreover, discarding data from subjects who have not completed follow-up may impair efficiency of the analysis.

### 2.4. The Kaplan–Meier estimator

In analysis of time-to-event data, the KM estimator offers a non-parametric estimate of the survival function. To apply the estimator in the estimation of event proportion, we have to capture time from the end of treatment to the occurrence of event (denoted by \( t \)). For those in whom the event has not occurred, the time to event \( t \) is censored at the time of interim analysis or the end of follow-up, whichever the earlier; for those in whom the event has occurred, \( t \) is uncensored. Let \( t_{e1}, t_{e2}, \ldots, t_{eK} \) be the distinct times to event of the \( r_1 \) events from the \( n_1 \) randomised subjects who have completed the treatment (\( K \leq r_1 \)). Denoted by \( m_i \) the number of subjects ‘at risk’ of the event just prior to time \( t_{ei} \), and by \( d_i \) the number of events at time \( t_{ei} \). Then, the KM estimate of the event proportion is

\[
\hat{\pi}_{KM} = 1 - \prod_{i=1}^{K} \frac{m_i - d_i}{m_i}.
\]

We note that the event proportion defined above carries forward the KM estimate at the maximum observed time to the end of follow-up \( T \) when the maximum observed time is censored and less than \( T \). This issue will be explored in the discussion section. The KM estimate includes data from all randomised subjects and does not rely on an assumption of a parametric distribution for the data. Moreover, the estimate is consistent and asymptotically normal under certain conditions (Breslow and Crowley 1974, Meier 1975).

### 2.5. The parametric estimator

In some trials, there may be a subset of subjects who are free of the event. For example, patients may be cured of a disease and will never have a relapse. Without loss of generality, we denote by \( \theta \) the probability of the event occurring at any time over an unlimited period. Let \( S(t) \) be the survival function at time \( t \) for subjects in whom the event occurs. By definition the event proportion \( \pi = \theta(1 - S(T)) \). Consequently, the estimate of event proportion by the PR estimator is given as:

\[
\hat{\pi}_{PR} = \hat{\theta}(1 - \hat{S}(T)).
\]

where \( \hat{\theta} [\hat{S}(T)] \) is the maximum likelihood estimate (MLE) of \( \theta [S(T)] \). The corresponding likelihood function at the interim analysis is

\[
L = \prod_{i=1}^{n_1} [\theta f(t_i)]^{\delta_i} [1 - \theta + \theta S(t_i)]^{1 - \delta_i},
\]

where \( t_i \) is the observed time to the event for the \( i \)th randomised subject (\( 1 \leq i \leq n_1 \)), \( \delta_i \) is the indicator of an observed event for the subject and \( f(t) \) is the probability density function for the distribution of time to the event. Experience from similar trials on distribution of the time-to-event data shall be useful in the choice of distribution. When the data come exactly from the selected distribution, the PR estimator is expected to perform well in estimating the event proportion. In reality, however, there is always a risk of misspecification of such a distribution.
2.6. RMSE of estimators

We derived the RMSE of the event proportion estimators except for the PD one. We consider a scenario where subject enrolment is at a constant rate of \( s \) subjects per week and the event will occur in all subjects if they are followed up long enough (i.e., \( \theta = 1 \)). To simplify the issue, we further assume there is no loss to follow-up, and time from the end of treatment to the event follows an exponential distribution with a hazard rate of \( \lambda \). Consequently, the event proportion \( \pi = 1 - \exp(-\lambda T) \). Note that interim analysis is planned to occur when a pre-specified number of subjects \( n_2 \) complete follow-up, and the number of subjects at on-going follow-up \( n_1 - n_2 \) has an expectation of \( Ts \). By grouping the \( n_1 \) subjects into those who have completed follow-up \( (n_2) \) and those who are at on-going follow-up \( (n_1 - n_2) \), the expectation and variance of \( r_1 \) are obtained as:

\[
E(r_1) = n_2\pi + s \int_0^T [1 - \exp(-\lambda(T - u))] du = n_2\pi + Ts(1 - \frac{\pi}{\lambda T}) \quad \text{and} \quad \text{Var}(r_1) = n_2\pi(1 - \pi) + \frac{T\pi^2}{2\lambda T}
\]

Hence, we obtained

\[
\text{RMSE}(\hat{\pi}_{AR}) = \frac{1}{n_2 + Ts} \sqrt{(Ts)^2 \left[ \left( \frac{1}{\lambda T} \right) \pi - 1 \right]^2 + n_2\pi(1 - \pi) + \frac{T\pi^2}{2\lambda T}} \quad (1)
\]

Obviously, the RMSE of the AC estimator \( \hat{\pi}_{AC} \) is given as:

\[
\sqrt{\pi(1 - \pi)/n_2} \quad (2)
\]

Based on the formula (2h) in Kaplan and Meier (1958), we obtained the approximate variance of the KM estimator as:

\[
\text{Var}(\hat{\pi}_{KM}) \approx \lambda(1 - \pi)^2 \int_0^T \frac{\exp(\lambda u)}{n_2 + (T - u)s} du \quad (3)
\]

Kaplan and Meier (1958) have shown that the approximate variance of the KM estimator is smaller than that of the AC estimator under more general situations. Moreover, \( \hat{\pi}_{AC} \) is unbiased and \( \hat{\pi}_{KM} \) is a consistent estimator. Consequently, we expect the RMSE of \( \hat{\pi}_{KM} \), approximately equal to its standard error, is smaller than that of \( \hat{\pi}_{AC} \) for a large sample size. Based on the delta method and the approximate variance of MLE of \( \lambda \), we approximate the RMSE of \( \hat{\pi}_{PR} \) as:

\[
\text{RMSE}(\hat{\pi}_{PR}) \approx \lambda T(1 - \pi)/\sqrt{n_2\pi + Ts\left( \frac{1 - \pi}{\lambda T} \right)} \quad (4)
\]

As an example, for \( \lambda = 0.4 \) and \( T = 1 \) year, the event proportion \( \pi = 0.33 \). Suppose recruitment rate is 1 subject per day and interim analysis is planned to occur when \( n_2 = 365 \) subjects have completed their 1-year follow-up. Then, according to Equations 1–4, \( \text{RMSE}(\hat{\pi}_{AR}) = 0.079 \), \( \text{RMSE}(\hat{\pi}_{AC}) = 0.025 \), \( \text{RMSE}(\hat{\pi}_{KM}) \approx 0.021 \), and \( \text{RMSE}(\hat{\pi}_{PR}) \approx 0.020 \).

3. Simulations

In our simulation study, data were generated for a single-arm trial with a sample size of 200 subjects unless specified otherwise. The recruitment rate was \( s \) subjects per week. Each subject was followed up for \( T = 65 \) weeks post-treatment with a treatment period of \( T_0 = 12 \) weeks. The probability of event during the post-treatment follow-up is \( \pi \) for each subject. Interim analysis was performed to estimate the event proportion \( \pi \) when \( n_2 \) subjects have completed follow-up, and their data until the end of follow-up were available for interim analysis. We assumed a time gap of 6 weeks between a visit and the time when data are available for analysis. We compared the performance of the five estimators described in Section 2.1–2.5 in terms of bias and RMSE in estimating the event.
proportion. For the PD estimator, three durations were considered, i.e., $T_1 = 12, 24$ and $48$ weeks. An exponential distribution was assumed for time to the event in the PR estimator.

Simulations have been performed for comparison of the five estimators under various scenarios that were different in recruitment rate $s$, event proportion $\pi$, distribution of time to the event or number of subjects who have completed follow-up $n_2$. For each scenario, 50,000 random samples were generated, which ensured the standard error of the mean of an event proportion estimate was less than 0.07% for $n_2 \geq 10$. In the simulations, the time interval between enrolments of two consecutive subjects was generated from an exponential distribution with a mean of 1/$s$ week. Unless specified, time-to-the-event data were generated from an exponential distribution with a mean of 4.5 months for subjects in whom the event occurred. The event indicator (for unlimited follow-up) was a Bernoulli variable with a parameter $\pi$.

**3.1. Results**

Simulation results were presented in Table 1 for different combination of event proportion and number of subjects who have completed follow-up, in Table 2 for two distributions of time to event, and in Figure 1 for different enrolment rates.

The results indicated that the PD estimator was biased (Tables 1 and 2). The PD estimator could under- or overestimate the event proportion, depending on the pre-specified duration $T_1$ and the distribution of time to event. For a value of $T_1$, the estimator could overestimate the event proportion

<table>
<thead>
<tr>
<th>$n_2$</th>
<th>$T_n$ (month)</th>
<th>$n_1$</th>
<th>$\hat{\pi}_{PD}$ 12 weeks</th>
<th>$\hat{\pi}_{PD}$ 24 weeks</th>
<th>$\hat{\pi}_{PD}$ 48 weeks</th>
<th>$\hat{\pi}_{AR}$</th>
<th>$\hat{\pi}_{AC}$</th>
<th>$\hat{\pi}_{KA}$</th>
<th>$\hat{\pi}_{PR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 5%$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>28.8 (1.5)</td>
<td>123 (9)</td>
<td>-0.3 (2.2)</td>
<td>0.4 (2.5)</td>
<td>1.3 (3.2)</td>
<td>-1.4 (2.2)</td>
<td>0.0 (3.4)</td>
<td>0.0 (2.4)</td>
<td>0.0 (2.4)</td>
</tr>
<tr>
<td>80</td>
<td>38.1 (2.1)</td>
<td>163 (9)</td>
<td>-0.2 (1.9)</td>
<td>0.3 (2.0)</td>
<td>0.8 (2.4)</td>
<td>-1.1 (1.9)</td>
<td>0.0 (2.4)</td>
<td>0.0 (1.9)</td>
<td>0.0 (1.9)</td>
</tr>
<tr>
<td>120</td>
<td>47.4 (2.6)</td>
<td>198 (4)</td>
<td>-0.1 (1.6)</td>
<td>0.2 (1.8)</td>
<td>0.6 (2.0)</td>
<td>-0.7 (1.6)</td>
<td>0.0 (2.0)</td>
<td>0.0 (1.7)</td>
<td>0.0 (1.7)</td>
</tr>
<tr>
<td>$n = 40%$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>28.8 (1.5)</td>
<td>123 (9)</td>
<td>-2.6 (5.6)</td>
<td>1.3 (5.5)</td>
<td>5.6 (8.0)</td>
<td>-11.3 (12.0)</td>
<td>0.0 (7.7)</td>
<td>0.0 (5.5)</td>
<td>0.1 (5.5)</td>
</tr>
<tr>
<td>80</td>
<td>38.1 (2.1)</td>
<td>163 (9)</td>
<td>-1.8 (4.6)</td>
<td>0.9 (4.5)</td>
<td>3.7 (5.9)</td>
<td>-8.5 (9.3)</td>
<td>0.0 (5.5)</td>
<td>0.0 (4.5)</td>
<td>0.0 (4.4)</td>
</tr>
<tr>
<td>120</td>
<td>47.4 (2.6)</td>
<td>198 (4)</td>
<td>-1.4 (3.9)</td>
<td>0.7 (3.9)</td>
<td>2.8 (4.8)</td>
<td>-6.0 (6.9)</td>
<td>0.0 (4.5)</td>
<td>0.0 (3.8)</td>
<td>0.0 (3.8)</td>
</tr>
</tbody>
</table>

$n_2$: number of subjects who have completed follow-up; $n_1$: number of subjects randomised; $T_n$: time point of interim analysis; $\hat{\pi}_{PD}$: the ‘pre-specified duration’ estimator; $\hat{\pi}_{AR}$: the ‘all randomised’ estimator; $\hat{\pi}_{AC}$: the ‘all completed’ estimator; $\hat{\pi}_{KA}$: the KM estimator; $\hat{\pi}_{PR}$: the parametric estimator.

<table>
<thead>
<tr>
<th>Time-to-event distribution</th>
<th>Trigger of interim analysis</th>
<th>$\hat{\pi}_{PD}$ 12 weeks</th>
<th>$\hat{\pi}_{PD}$ 24 weeks</th>
<th>$\hat{\pi}_{PD}$ 48 weeks</th>
<th>$\hat{\pi}_{AR}$</th>
<th>$\hat{\pi}_{AC}$</th>
<th>$\hat{\pi}_{KA}$</th>
<th>$\hat{\pi}_{PR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential distribution</td>
<td>$n_2 = 40$</td>
<td>-10.3 (11.5)</td>
<td>-4.6 (6.9)</td>
<td>0.8 (5.2)</td>
<td>-24.1 (24.5)</td>
<td>0.0 (7.2)</td>
<td>0.0 (5.8)</td>
<td>-0.6 (5.4)</td>
</tr>
<tr>
<td>$n_2 = 80$</td>
<td>-7.3 (8.4)</td>
<td>-3.1 (5.2)</td>
<td>0.6 (4.2)</td>
<td>-18.2 (18.6)</td>
<td>0.0 (5.1)</td>
<td>0.0 (4.5)</td>
<td>-0.3 (4.3)</td>
<td></td>
</tr>
<tr>
<td>$n_2 = 120$</td>
<td>-5.6 (6.7)</td>
<td>-2.4 (4.3)</td>
<td>0.4 (3.6)</td>
<td>-13.2 (13.6)</td>
<td>0.0 (4.2)</td>
<td>0.0 (3.8)</td>
<td>-0.2 (3.7)</td>
<td></td>
</tr>
<tr>
<td>Weibull distribution</td>
<td>$n_2 = 40$</td>
<td>-21.1 (21.7)</td>
<td>-14.3 (15.3)</td>
<td>-6.6 (8.7)</td>
<td>-33.0 (33.2)</td>
<td>0.0 (7.2)</td>
<td>0.0 (6.2)</td>
<td>-15.5 (16.0)</td>
</tr>
<tr>
<td>$n_2 = 80$</td>
<td>-14.8 (15.4)</td>
<td>-9.6 (10.6)</td>
<td>-4.2 (6.2)</td>
<td>-24.9 (25.1)</td>
<td>0.0 (5.1)</td>
<td>0.0 (4.7)</td>
<td>-13.0 (13.4)</td>
<td></td>
</tr>
<tr>
<td>$n_2 = 120$</td>
<td>-11.4 (12.0)</td>
<td>-7.2 (8.2)</td>
<td>-3.1 (4.9)</td>
<td>-18.7 (18.9)</td>
<td>0.0 (4.2)</td>
<td>0.0 (3.9)</td>
<td>-11.8 (12.1)</td>
<td></td>
</tr>
</tbody>
</table>

$n_2$: number of subjects who have completed follow-up; $\hat{\pi}_{PD}$: the ‘pre-specified duration’ estimator; $\hat{\pi}_{AR}$: the ‘all randomised’ estimator; $\hat{\pi}_{AC}$: the ‘all completed’ estimator; $\hat{\pi}_{KA}$: the KM estimator; $\hat{\pi}_{PR}$: the parametric estimator.
at some scenarios and underestimate at others. For example, the PD estimator with $T_1 = 24$ weeks overestimated at $\pi = 5\%$ and $40\%$ (Table 1), and underestimated at $\pi = 70\%$ (Table 2). The magnitude of bias with the PD estimator was up to $30\%$ of the event proportion, which depended substantially on the pre-specified duration $T_1$ and the distribution of time to the event but less on the threshold number $n_2$ and the event proportion $\pi$ (Table 2). The mean of $\hat{\pi}_{PD}$ got closer to the true value $\pi$ for larger values of $n_2$. This could be explained by the following facts. The calculation of $\hat{\pi}_{PD}$ could be viewed as weighted estimates of event proportion from three subsets of subjects: (a) subjects who have completed follow-up; (b) subjects who have been followed up for less than $T_1$ and have experienced the event; (c) subjects who have been followed up for at least $T_1$ but yet to complete follow-up. For a trial with a constant enrolment rate, the expected numbers of subjects in subsets (b) and (c) were relatively stable during the period when enrolment was still ongoing and some subjects have been completed the follow-up. Since the estimate of event proportion was unbiased in the subset (a), increasing number of subjects in this subset would pull $\hat{\pi}_{PD}$ closer to $\pi$. We noted that the bias of $\hat{\pi}_{PD}$ was also affected by the enrolment rate (Figure 1). In particular, the magnitude of bias in the estimates increased with the enrolment rate for the PD and AR estimators.

The AR estimator always underestimated the event proportion unless all subjects have completed the follow-up. For a fixed value of $n_2$, the bias of the AR estimator was more substantial when the recruitment rate increased (Figure 1). The bias was up to $47\%$ of the event proportion in the simulations (Table 2, lower panel). Similar to the PD estimator, the bias became less severe when interim analysis was conducted with a larger number of subjects who have completed the follow-up. The results in Table 2 indicated that the extent of bias by the AR estimator is larger when the event tends to occur later in time (median of the exponential distribution $= 6.9$ months $< 9.9$ months $=$ median of the Weibull distribution).

As shown in Table 1, the AC, KM and the PR estimators all produced unbiased estimates of event proportion (with one exception for the PR estimator). Consequently, the RMSE in Table 1 was essentially the sample standard deviation (SD) of estimators for these three estimators. The KM estimator gained more efficiency over the AC estimator by including all subjects in the estimation, reflected as a smaller RMSE of $\hat{\pi}_{KM}$ than $\hat{\pi}_{AC}$. Such gain in efficiency was more substantial when events occurred earlier (Table 2) or the proportion of subjects who have not completed followed up was bigger. The latter could happen if interim analysis was performed early, i.e., with a smaller $n_2$ (Tables 1 and 2), or if the enrolment was faster (Figure 1, the right graph). Interestingly, the PR

![Figure 1. The mean and RMSE of event proportion estimates as a function of enrolment rate. Event proportion $\pi = 25\%$; interim analysis was performed after $n_2 = 40$ subjects have completed follow-up; upper limit of sample size = 400. The marker of the 'all completed' estimator was enlarged in the left graph for a better presentation of the results.](image-url)
estimator did not lead to an appreciably smaller RMSE in the estimation than the (nonparametric) KM estimator when the assumed model was correct in the PR estimator. Since an exponential distribution was assumed for the time to event in the PR estimator, such an assumption was generally incorrect when the data actually followed a Weibull distribution. At such scenarios, it was no surprise that estimate of event proportion by the PR estimator was substantially biased (Table 2, lower part). The KM estimator, on the other hand, produced unbiased estimates regardless of what the underlying distribution of time to event was. We noted that the RMSE of the PD and AR estimators could be more than triple of that of the KM estimator, although they were comparable in some scenarios.

4. An example

We consider a TB trial with a multiple-arm multiple-stage (MAMS) design. Details of MAMS designs were given by Royston et al. (2003) and Magirr et al. (2012). Briefly, experimental arms that do not show sufficient promise at interim analysis are dropped and recruitment is continued only in the remaining arms. This offers an opportunity to speed up assessments of new treatment regimens and meanwhile increases the chance of identifying at least one successful treatment regimen in a trial. The designs have been implemented in several trials, such as STAMPEDE (STAMPEDE Trial Development Group 2004), and the TAILoR trial (Wason & Jaki, 2012).

In this example, subjects randomised to the experimental arms received treatment for 3 months and then they are followed up until 18 months for monitoring TB relapse. The primary outcome is the sputum culture status at 18 months after randomisation, while dropping experimental arms is based on the estimate of TB relapse rate. An experimental arm with a relapse rate of 20% or higher is considered not acceptable. We derive the criterion for dropping an experimental arm from testing the following hypotheses.

\[ H_0 : \pi = 20\% \text{ versus } H_1 : \pi = 10\% \]

The interim analysis is performed upon completion of the follow-up for \( n_2 = 56 \) subjects in an experimental arm. If estimate of relapse rate \( \hat{\pi} \) is equal to or smaller than a critical value of 16.1%, the null hypothesis is rejected and the corresponding experimental arm continues to the next stage. Otherwise, the experimental arm is dropped from the trial. This criterion ensures a one-sided significance level of 30% and a power of 95% at the interim analysis if the AC estimator is used. That is, the probability of not being dropped was 30% and 95% for an experimental arm with a relapse rate of 20% and 10%, respectively. A significance level greater than its typical value 2.5% is desired to have the interim analysis performed at an early time (Royston et al. 2011). With this criterion of dropping experimental arms, we carried out simulations to compare the performance of the five estimators of the event proportion – the relapse rate, in terms of type I error rate and power. We simulated data for one experimental arm only as there was no comparison between arms, and the arm-dropping criterion applies separately to each experimental arm.

As shown in Table 3, the AC estimator, producing a larger type I error rate and a smaller power, performed worse than the KM estimator. That is to say, if the critical value for dropping an experimental arm is adjusted for the KM estimator such that the type I error rate is 30%, its power would be larger than that of the AC estimator. The power of the AR estimator was greater than that of the KM and the PR estimators, at the cost of a much larger type I error rate. This means the AR estimator was too aggressive in the sense that it allowed an experimental arm with a relapse rate of 20% continuing to the next stage with a probability much greater than 30%, the desired significance level. The performance of the PD estimator was sensitive to its pre-specified duration and the distribution of time to event. We noted that the type I error rate of the PR estimator was inflated when assumption of the distribution of time to the event was incorrect (Table 3, lower part).
5. Discussion

In clinical trials aiming to estimate an event proportion at interim analysis, i.e., the probability of an event occurring any time during a follow-up period, we found from our simulations that the PD estimator generally results in biased estimates of the event proportion. The bias could be as much as 30% of the event proportion, and the extent of bias depends substantially on the ‘pre-specified duration’, the distribution of time to the event and the enrolment rate. It was no surprise that the AR estimator always underestimates the event proportion unless all subjects have completed the follow-up. The AC estimator, although producing unbiased estimates, comes with loss of efficiency as subjects who have not completed follow-up are excluded from the estimation of event proportion. Overall, the KM estimator performed best when considering both unbiasedness and variability of the estimates, and robustness of estimates. The efficiency gain of the KM over the AC estimator was more substantial when event occurred earlier or a larger proportion of subjects were at ongoing follow-up. This can be seen in Figure 1 and all tables. Figure 1 illustrated the situation where the interim analysis was performed with a fixed number of subjects who have completed the follow-up. As the enrolment rate increased, the proportion of subjects who were at ongoing follow-up at the time of interim analysis became bigger, and so was the difference in MSE between the AC and KM estimators. Methods have been proposed to extend the KM estimate from the maximum observed time to the end of follow-up period. Efron (1967) assigned 0 as the KM estimate of the survival function at time points beyond

<table>
<thead>
<tr>
<th>Distribution of time to relapse</th>
<th>Estimator of relapse rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential distribution with a mean = 10 months</td>
<td>PD</td>
</tr>
<tr>
<td>Type I error rate (%)</td>
<td></td>
</tr>
<tr>
<td>s = 0.5 subject per week</td>
<td>40.4</td>
</tr>
<tr>
<td>s = 1 subject per week</td>
<td>52.9</td>
</tr>
<tr>
<td>Power (%)</td>
<td></td>
</tr>
<tr>
<td>s = 0.5 subject per week</td>
<td>98.9</td>
</tr>
<tr>
<td>s = 1 subject per week</td>
<td>99.9</td>
</tr>
</tbody>
</table>

Weibull distribution with a mean = 10 months

| Type I error rate (%)          |                |                |                |            |   |   |   |
| s = 0.5 subject per week       | 58.4 | 47.2 | 25.3 | 72.8 | 29.3 | 21.8 | 31.0 |
| s = 1 subject per week         | 81.6 | 65.9 | 26.8 | 94.5 | 29.2 | 19.9 | 35.0 |
| Power (%)                      |                |                |                |            |   |   |   |
| s = 0.5 subject per week       | 99.6 | 99.1 | 95.8 | 99.9 | 95.2 | 94.6 | 97.2 |
| s = 1 subject per week         | 100.0 | 99.9 | 97.6 | 100.0| 95.2 | 95.5 | 98.8 |

In the situation of fast enrolment and a long follow-up period, it could happen all enrolled subjects’ follow-up is still ongoing at the time of interim analysis. Hence, the maximum observed time to event is smaller than the follow-up period $T$. This is challenging to the AC and the KM estimator as the former is not defined when none of the subjects has completed follow-up and the latter is not defined unless the maximum observed time is not a censored time point. Methods have been proposed to extend the KM estimate from the maximum observed time to the end of follow-up period. Efron (1967) assigned 0 as the KM estimate of the survival function at time points beyond
the maximum observed time to event. Gill (1980) proposed carrying forward the KM estimate at the maximum observed time to its subsequent time period, a method we have adopted in the KM estimator. The procedure by Gelber et al. (1993) projects KM estimates beyond the maximum observed time by fitting an appropriate parametric model to the tail of a survival curve. We note that the methods by both Efron (1967) and Gill (1980) are subject to bias at time points beyond the maximum observed time. Both the procedure by Gelber et al. (1993) and the PR estimator in our simulation rely on the appropriateness of the parametric model.

As an alternative to the methods above for extending KM estimate from the maximum observed time to the end of follow-up period, estimates of event proportion over a shorter period rather than the follow-up period, e.g., half of the follow-up period, may be considered. Although the event proportion over a shorter period may be less informative, it may form a good basis for decision-making at interim analysis. Considering our example of the TB trial, relapse rate at 1 year instead of 18 months may be the basis for dropping unpromising experimental arms. The unacceptable relapse rate of 20% at 18 months is translated into 17% at 1 year after randomisation (9 months after stopping treatment), according to the finding from 15 TB trials that 78% and 91% of relapses occurred within 6 months and 12 months, respectively, of stopping treatment (Nunn et al. 2010). Similarly, a relapse rate of 10% at 18 months is translated into 8.5% at 1 year. Then, the critical value in a criterion for dropping an experimental arm that is based on the estimate of 1-year relapse rate could be chosen from distributions of the estimators by simulation.

Five estimators of event proportion are compared in this paper in terms of bias and variability of the estimates. An example is given to illustrate how the estimators affect dropping arms at interim analysis of a multi-arm and multi-stage trial. These estimators could also affect decision-making at interim analysis when sample size re-estimation is of interest. Given the unbiasedness, variability and robustness of the KM estimator, we expect it performs better than other four estimators in sample size re-estimation as well.

**Acknowledgments**

The authors would like to thank the referees for their insightful comments and suggestions.

**References**


