SpringerPlus
a SpringerOpen Journal

RESEARCH Open Access

Interim analysis for binary outcome trials with a long fixed follow-up time and repeated outcome assessments at pre-specified times

Sameer Parpia[1*], Jim A Julian[1], Chushu Gu[1], Lehana Thabane[2] and Mark N Levine[1]

Abstract

In trials with binary outcomes, assessed repeatedly at pre-specified times and where the subject is considered to have experienced a failure at the first occurrence of the outcome, interim analyses are performed, generally, after half or more of the subjects have completed follow-up. Depending on the duration of accrual relative to the length of follow-up, this may be inefficient, since there is a possibility that the trial will have completed accrual prior to the interim analysis. An alternative is to plan the interim analysis after subjects have completed follow-up to a time that is less than the fixed full follow-up duration. Using simulations, we evaluated three methods to estimate the event proportion for the interim analysis in terms of type I and II errors and the probability of early stopping. We considered: 1) estimation of the event proportion based on subjects who have been followed for a pre-specified time (less than the full follow-up duration) or who experienced the outcome; 2) estimation of the event proportion based on data from all subjects that have been randomized by the time of the interim analysis; and 3) the Kaplan-Meier approach to estimate the event proportion at the time of the interim analysis. Our results show that all methods preserve and have comparable type I and II errors in certain scenarios. In these cases, we recommend using the Kaplan-Meier method because it incorporates all the available data and has greater probability of early stopping when the treatment effect exists.

Keywords: Interim analysis; Binary outcome; Power; Type I error

Background

Interim analyses that permit early stopping of a randomized controlled trial (RCT) for extremely positive results or for futility are included in the design for ethical and economic reasons. Strategies have been developed for interim analyses such that the overall type I error of the entire trial is preserved at a fixed level (Haybittle 1971; O'Brien and Fleming 1979; Peto et al. 1976; Pocock 1977).

Often, the primary outcome is whether or not a subject experienced an event over a fixed period of time T. In some trials, the outcome is assessed repeatedly at pre-specified times during follow-up, and the subject is considered a failure if the event occurs at any time. For example, in a cardiovascular RCT investigating the effect of an intervention for preventing post-thrombotic syndrome, subjects can be assessed every 6 months for up to 24 months using a disease-specific questionnaire (Enden et al. 2012; Vedantham et al. 2013). A failure has occurred if the questionnaire score exceeds a pre-specified threshold. Another example would be a breast cancer radiotherapy RCT where adverse cosmesis (i.e. a dichotomy), assessed at 1, 3 and 5 years post-randomization, would be the primary safety outcome and the focus of the interim analysis.

Interim analyses are generally performed after half or more of the subjects have completed follow-up (Pedley 2011). Depending on the duration of accrual relative to the length of follow-up, this strategy may be inefficient because it is possible that accrual will have been completed and patients will have finished treatment prior to

* Correspondence: parpia@mcmaster.ca
[1]Ontario Clinical Oncology Group, Department of Oncology, McMaster University, 711 Concession Street – G (60) Wing 1st Floor, Hamilton, ON L8V 1C3, Canada
Full list of author information is available at the end of the article

the interim analysis. If, however, the interim analysis was done earlier and a statistically significant effect was found, the trial may be stopped, and all future subjects would receive the experimental therapy.

In this situation, one alternative is to plan an interim analysis after a smaller percentage of subjects have completed full follow-up. However, there is a low probability of terminating the trial early when the interim analysis is based on so little information, and, therefore, such an analysis would unnecessarily spend alpha (Togo and Iwasaki 2013). A second alternative is to plan the interim analysis after half or more of the subjects have completed a specified portion of the follow-up R, where $R < T$, and T is the fixed full follow-up duration for each subject.

Several researchers have studied methods that combine data from subjects who have completed full follow-up with those who have been followed for duration R in situations where the outcome is reversible (Marschner and Becker 2001; Sooriyarachchi et al. 2006; Whitehead et al. 2008). In our research, however, the situation is different in that the outcome can be ascertained at any of the pre-specified visits during follow-up and is irreversible.

In this paper, we consider 3 methods of estimating the interim event proportion (risk) for each treatment group in an RCT for an interim analysis: 1) estimated event proportion based only on subjects who have been followed for at least duration R or who had an outcome event; 2) the event proportion based on data from subjects that have been randomized by the time of the interim analysis, and 3) the Kaplan-Meier approach to estimate the event proportion. We investigate the effect of each method on the type I and II errors and the probability of early stopping through computer simulation of various trial scenarios.

Methods

Consider a trial designed to detect an absolute risk reduction (ARR) between the standard group (π_0) and the experimental group (π_1) over the time period 0 to T using a normal approximation Z-test with

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_0}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n_0}}}$$

where $\hat{\pi}_0$ and $\hat{\pi}_1$ are the observed proportions, n_0 and n_1 are the group sample sizes, and we are testing the one-sided hypotheses H_0: $\pi_1 \geq \pi_0$ versus H_1: $\pi_1 < \pi_0$. Furthermore, we assume 90% power, an alpha of 0.025 and a 1:1 randomization. Since the normal distribution is symmetric, the p-value for a one-sided test is equivalent to half of the two-sided p-value.

Suppose the trial requires 4 years for enrolment, each subject is followed for 2 years (i.e. $T = 24$ months), and failures are ascertained at any of the four 6-monthly pre-specified visits post-randomization. Let the start of the trial (calendar time) be denoted by τ_0. Following the notation in Table 1, let t_j be the pre-specified visit times in the trial where $t_j \leq T$ and j is the visit number where $j = 0, 1, 2\ldots J$, and J denotes the number of visits (e.g. $J = 4$ and $t_0 = 0$, $t_1 = 6$, $t_2 = 12$, $t_3 = 18$, $t_4 = 24$ months). Suppose an interim analysis is scheduled to occur when 50% of the subjects have completed $R = 12$ months ($t_2 = R$) of follow-up which, assuming a uniform recruitment pattern, corresponds to approximately 36 months after the start of the trial, denoted by τ_1 (Figure 1). At the interim analysis, the proportion of subjects who fail in each group could be estimated using any of the following approaches.

Method 1: event proportion based on subjects followed for at least duration R or who had an event

In RCTs where the length of enrolment relative to follow-up is not an issue, subjects included in the interim analysis are those who have completed their full follow-up T or who have had an event prior to completion (Pedley 2011). A similar approach is used here whereby we include only subjects who have completed at least duration R (where $t_r = R$, r refers to the visit at which follow-up time equals R) of their full follow-up T, or have had an event prior to this point. Since the interim analysis occurs after 50% of the subjects have completed at least follow-up of R, this approach includes the first 50% of enrolled subjects plus those subjects that have experienced an event but have not completed follow-up of R. For each treatment group i (0 = standard, 1 = experimental) at visit time t_j, let m_{ij} be the number of subjects at risk (i.e. have completed visit at t_j without having an event), and let e_{ij} be the number of new events diagnosed. Then the event proportion in treatment group i at the time of interim analysis τ_1 is given by:

$$\hat{\pi}_i(\tau_1) = \frac{\sum_{k=1}^{J} e_{ik}}{m_{ir} + \sum_{k=1}^{J} e_{ik}}$$

The individuals who have experienced an event but have not completed duration R of follow-up are included in the numerator and the denominator.

Method 2: event proportion based on data from subjects that have been randomized by the time of the interim analysis

This simple approach uses data from the subjects randomized by the time of the interim analysis τ_1 (i.e. once 50% of the subjects have been followed for at least time R). Let n_i

Table 1 Notation table for estimation of event proportions

Visit number J	Visit time t_j	Subjects at risk m_j	New events e_j	Incidence at visit j d_j
0	t_0 (<6 m)	m_0	$e_0 = 0$	$d_0 = 0$
1	t_1 (6 m)	m_1	e_1	$d_1 = e_1/m_1$
2	t_2 (12 m)	m_2	e_2	$d_2 = e_2/m_2$
3	t_3 (18 m)	m_3	e_3	$d_3 = e_3/m_3$
4	t_4 (24 m)	m_4	e_4	$d_4 = e_4/m_4$

be the number of subjects who have been randomized to treatment group i. Then the event proportion for each group at the time of interim analysis τ_1 is given by

$$\hat{\pi}_i(\tau_1) = \frac{\sum_{k=1}^{J} e_{ik}}{n_i}$$

which is simply the total number of observed events divided by the number of subjects randomized by time τ_1.

Method 3: Kaplan-Meier approach

This approach also uses all the data available at the time of the interim analysis τ_1 (i.e. once 50% of the subjects have been followed for at least time R). For individuals who have not completed follow-up time T (i.e. the full fixed follow-up duration) and have not had the event, they are simply right-censored at the latest time that they were observed. Then the Kaplan-Meier (KM) estimates can be calculated using all randomized subjects and the event proportion in treatment group i at the time of interim analysis τ_1 is given by

$$\hat{\pi}_i(\tau_1) = 1 - S_i(T)$$

where $S_i(T)$ is the KM survivor function estimate. Following the notation in Table 1, this is equivalent to

$$\hat{\pi}_i(\tau_1) = 1 - \prod_{k=1}^{J}(1 - d_{ik}).$$

We evaluated these methods in terms of overall type I and II errors and the probability of early stopping of the trial for a positive result at the interim. The interim analysis was performed using the Haybittle-Peto (Haybittle 1971; Peto et al. 1976) and O'Brien-Fleming (O'Brien and Fleming 1979) monitoring boundaries for extreme positive results. These boundaries are conservative and require small p-values for early stopping of the trial. Other less conservative boundaries such as the Pocock approach were not evaluated (Freidlin and Korn 2009; Pocock 2005).

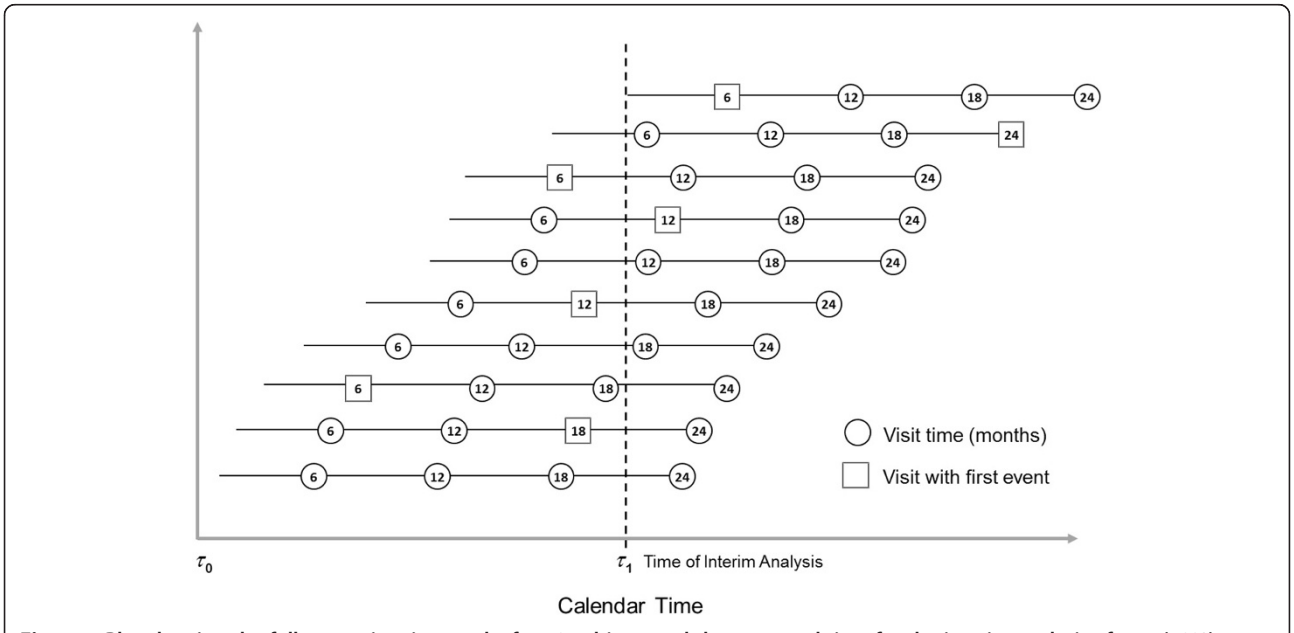

Figure 1 Plot showing the follow-up time in months for 10 subjects and the proposed time for the interim analysis after 5 (50%) subjects have completed 12 months of follow-up.

Simulation

We considered six RCTs similar to the trial described in the Methods section (see Table 2). Data for the binary endpoint were generated using the binomial distribution under the null and alternative hypotheses.

For each subject with an event, the time at which the event occurred was randomly assigned to reflect five clinically-plausible scenarios (Table 3), using the following: 1) events were distributed equally across the four time-points with probabilities (0.25, 0.25, 0.25, 0.25) for both groups; 2) the majority of the events occurred in the first two time-points with probabilities (0.35, 0.30, 0.20, 0.15) for both groups; 3) the majority of the events occurred in the last two time-points with probabilities (0.15, 0.20, 0.30, 0.35) for both groups; 4) the standard group follows distribution (3) and the experimental group follows distribution (2); and 5) the reverse of scenario (4). Entry times for subjects over 48 months were randomly generated from a uniform distribution, and the interim analysis was carried out after 50% of the subjects completed $R = 12$ months of follow-up. We carried out 10,000 replications for each trial. Given that $Z(x)$ and $Z(y)$ are the interim and final test statistics, respectively, the type I error rate, $P_{H_o}(Z(x) > g \text{ or } [Z(x) \leq g \text{ and } Z(y) > f])$, and the type II error, $P_{H_1}(Z(x) \leq g \text{ and } Z(y) \leq f)$, were obtained from data generated under the null and alternative hypotheses, respectively, where g and f are the interim and final critical values of the O'Brien-Fleming ($g = 2.797$, $f = 1.977$) and Haybittle-Peto ($g = 3.0$, $f = 1.967$) monitoring boundaries. The probability of early stopping, $P_{H_1}(Z(x) > g)$, was obtained under the alternative hypotheses. All analysis was performed in R 2.15 (www.r-project.org).

Results

The results of the type I error rates for the three methods are shown graphically in Figure 2. The three methods have comparable type I error rates across each of the trials and event distribution scenarios. The methods in general have nominal or close-to-nominal type I error rates when the event distribution probabilities are equivalent between treatment groups or when the experimental treatment group events occurred earlier in the trial compared with the standard group. However, under these same scenarios, slightly greater-than-nominal type I error rates are seen in the trials where (π_0, π_1) = (0.30, 0.10) and (π_0, π_1) = (0.50, 0.45), where the type I error rates are approximately 0.03. For the scenario where the experimental group events occurred later in the trial compared with the standard group, the type I error was generally inflated for all methods.

Table 2 Summary of six trials considered for simulation with β = 0.10 and a one-sided α = 0.025

Standard group event proportion (π_0)	Experimental group event proportion (π_1)	Absolute risk reduction (π_0-π_1)	N
0.30	0.25	0.05	3342
0.30	0.20	0.10	796
0.30	0.10	0.20	160
0.50	0.45	0.05	4182
0.50	0.40	0.10	1030
0.50	0.30	0.20	242

Table 3 Summary of the event distribution probabilities for the simulated scenarios

Scenario	Event distribution probabilities by visit time	
	t_1, t_2, t_3, t_4	
	Standard group	Experimental group
1	0.25, 0.25, 0.25. 0.25	*same as standard*
2	0.35, 0.30, 0.20, 0.15	*same as standard*
3	0.15, 0.20, 0.30, 0.35	*same as standard*
4	0.15, 0.20, 0.30, 0.35	0.35, 0.30, 0.20, 0.15
5	0.35, 0.30, 0.20, 0.15	0.15, 0.20, 0.30, 0.35

The three methods also have comparable type II error rates (Figure 3). In general, under all event distribution scenarios and trials, the type II error rates are comparable to the nominal value of 0.10 regardless of the interim analysis method or stopping boundary rule. Moreover, in the scenario where the experimental group events occurred later in the trial compared with the standard group, the type II errors rates are much lower than the nominal value for the trials with ARRs of 0.05 and 0.10.

Under the alternative hypothesis, methods 1 and 3 have comparable probabilities for early stopping in scenarios where the treatment groups have equivalent event distributions probabilities over time, specifically in the trials where $\pi_0 = 0.30$ (Figure 4). Method 3 has a slightly greater probability of early stopping than method 1 in the trials where $\pi_0 = 0.50$. Moreover, method 2 has the smallest probability of early stopping in scenarios where the treatment groups had equivalent event distributions probabilities over time. On the other hand, all methods have comparable probabilities of early stopping in the scenarios where the treatment groups had contrasting event distributions over time. The highest probabilities for early stopping are seen in the trials where the experimental group had a smaller proportion of events occur earlier in the trial compared with the standard group, and the lowest probabilities of early stopping are seen in the opposite scenario. In general, the probability for early stopping is greater using the O'Brien-Fleming boundaries compared with the Haybittle-Peto monitoring boundaries.

Discussion

In RCTs with binary endpoints, interim analyses are generally conducted after a considerable percentage of subjects

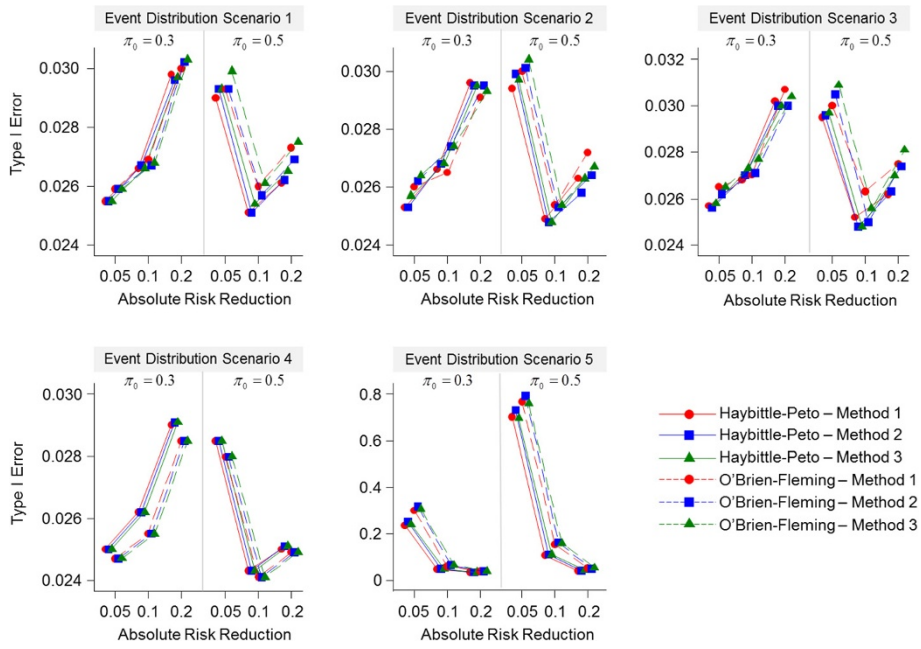

Figure 2 Overall type I error rates for each trial by event distribution scenario.

have completed follow-up. However, under certain situations this approach is not optimal since the trial may have completed accrual and all the subjects will have been treated by that time. We evaluated three approaches for an interim analysis when a considerable percentage of subjects complete a follow-up time that is less than the planned trial follow-up.

We observed that the type I error rates were comparable for all three methods. For most trials simulated, under the scenarios where the event distributions were equivalent between treatment groups or the experimental group had events occur earlier than the standard group, the type I error rates were close to the nominal value. These results concur with those of Pedley (2011),

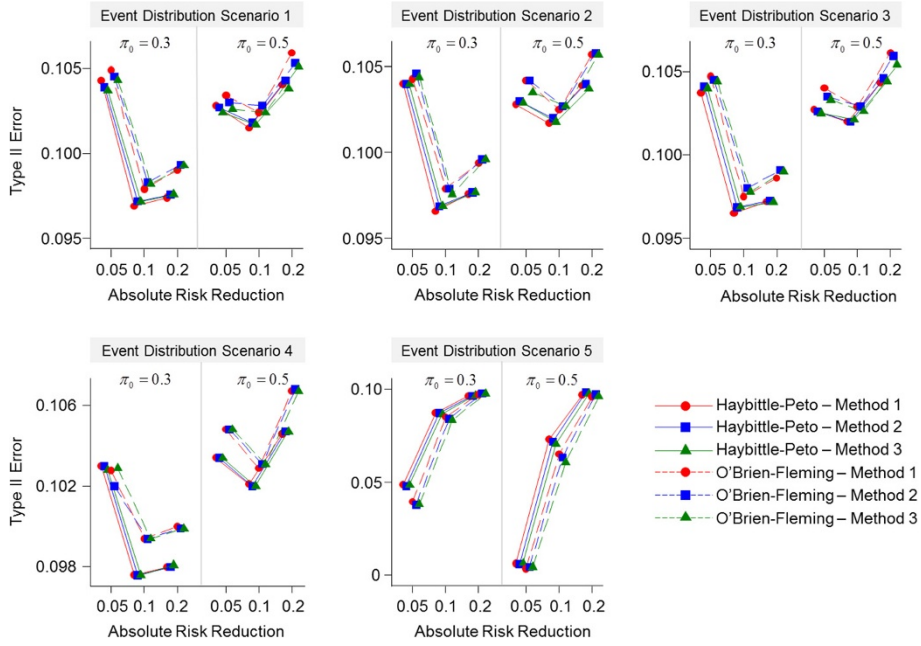

Figure 3 Overall type II error rates for each trial by event distribution scenario.

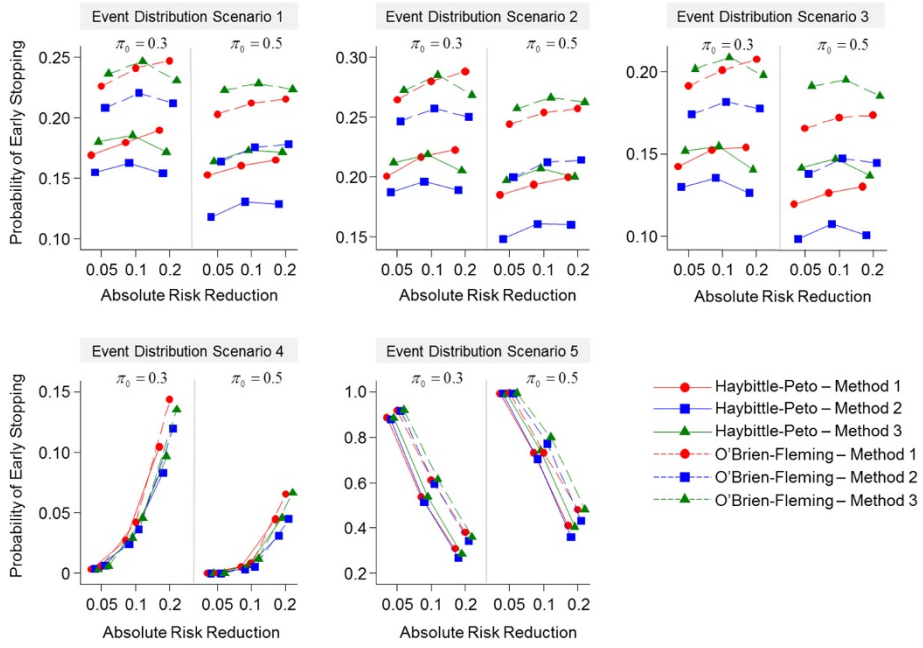

Figure 4 Probabilities for early stopping under the alternative hypothesis for each trial by event distribution scenario.

who showed that conducting the interim analysis after a considerable percentage of subjects had completed full follow-up (using method 2) produced nominal type 1 error rates, albeit in the situation where events could be measured at any time during follow-up and not just at specific time points. However, we also observed that the type I error rate increased with increasing absolute risk reduction for trials with a standard group event proportion of 0.3, thus resulting in slightly higher type I error rates for the trial with ARR to 0.20. In addition, similar slightly higher type I error rates were seen in the trial with a standard group event proportion of 0.5 and the ARR = 0.05. This is perhaps due to a combination of less variability and a small sample size for the former, and a large sample size and small ARR for the latter. Therefore, trialists should be cautious of using either of these methods under these situations.

While there were situations in which the type I errors were slightly inflated with all methods, the methods performed much better with regard to the type II errors under all scenarios, suggesting that these methods will not have a negative effect on the power to detect the hypothesized difference between treatment groups provided the difference exists. Under the scenarios where the experimental group had events occur later compared with the standard group, the methods showed increased overall power because the probability of early stopping was greater in these scenarios. However, under these scenarios, the type I error rates are inflated.

The methods differed on the probability of early stopping under the alternative hypothesis with method 2 having the lowest probability. This is because this approach includes data from all subjects that have been randomized by the time of the interim analysis in the denominator of the estimation of the event proportion even though a subgroup of these patients would not have had any assessment of the outcome since they would not have reached their first time point for outcome assessment. The consequence is the dilution of the interim treatment effect leading to lower interim power. Method 3 also uses all available data from randomized subjects at the time of the interim analysis. However, it employs a conditional probability approach which differentiates between those subjects who have not yet had an assessment visit (i.e. censored) and who are at risk at each assessment visit, thus yielding a greater probability of early stopping. Similarly, since method 1 uses only a subset of randomized subjects at the time of the interim analysis, the estimated interim treatment effect is less diluted and, therefore, has greater probability for early stopping than method 2. Conversely, since it uses a smaller number of subjects compared with method 3, the probability for early stopping is slightly lower than method 3 in trials where the standard group event proportion is 0.5, because the variability is greater for proportions closer to 0.5. Furthermore, we observed that the probabilities for early stopping are greater using the O'Brien-Fleming boundary compared with the Haybittle-Peto boundary since it is less conservative.

Although the largest probabilities of early stopping under the alternative hypothesis and the smallest type II errors were seen under the scenario where the experimental group had events occurring later compared with

the standard group, the type I errors is greatly inflated and, therefore, none of the methods can be recommended in this situation. Since there is a delay in occurrence of the event in the experimental group, this may be perceived as an effect of treatment. However, in situations where investigators are interested in the occurrence of an event over a fixed time period, this scenario, although rare, would still be considered under the null hypothesis.

Our study had some limitations. The generalizability of our findings may be limited since we evaluated six trial scenarios with particular event distributions over time. In diseases where the event distributions over time differ from the ones evaluated in this research, further simulations would be required to evaluate these methods. Secondly, we evaluated trials with one interim analysis after 50% of the subjects completed 12 months of follow-up using the O'Brien-Fleming or Haybittle-Peto approach. These findings may not be applicable to trials in which interim analyses are required at multiple times or when using the alpha spending function approach to monitor the trial. Finally, the biases of the interim event proportions and treatment effects were not evaluated primarily because it is well known that estimators at the interim are biased, especially for estimators that allow for early stopping for positive results. However, further investigation on the estimators is needed.

Conclusion

Nonetheless, we have shown that under certain scenarios, conducting an interim analysis when a considerable number of subjects have some follow-up data, using any of the methods, preserves the type I and II errors. Although all three methods preserve type I and II errors under these scenarios, we recommend using the Kaplan-Meier method because it incorporates all the available data and has greater probability of early stopping when the treatment effect exists. We have also shown that under certain scenarios, none of these methods is suitable for an interim analysis, and trialists should be cautious when using them. Finally, when possible, an interim analysis should be undertaken when data from a considerable number of subjects who have completed full follow-up are available. However, if waiting for a considerable number of subjects to complete full follow-up is not an efficient approach, such as in the examples described, the methods outlines in this paper should be considered and evaluated to fit the specific needs of the trial.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

SP, JAJ, CG, LT and MNL conceived the study. SP conducted literature review, designed and implemented the simulation, and wrote the initial draft of the manuscript. All authors reviewed and revised the draft version of the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

This research was funded in part by funds from the CANNeCTIN Program.

Author details

[1]Ontario Clinical Oncology Group, Department of Oncology, McMaster University, 711 Concession Street – G (60) Wing 1st Floor, Hamilton, ON L8V 1C3, Canada. [2]Biostatistics Unit - FSORC, St Joseph's Healthcare - Hamilton, 50 Charlton Avenue East, Hamilton, ON L8N 4A6, Canada.

Received: 2 April 2014 Accepted: 20 June 2014
Published: 26 June 2014

References

Enden T, Haig Y, Klow NE, Slagsvold CE, Sandvik L, Ghanima W, Hafsahl G, Holme PA, Holmen LO, Njaastad AM, Sandbaek G, Sandset PM, CaVenT Study Group (2012) Long-term outcome after additional catheter-directed thrombolysis versus standard treatment for acute iliofemoral deep vein thrombosis (the CaVenT study): a randomised controlled trial. Lancet 379(9810):31–38

Freidlin B, Korn EL (2009) Stopping clinical trials early for benefit: impact on estimation. Clin Trials 6(2):119–125

Haybittle JL (1971) Repeated assessment of results in clinical trials of cancer treatment. Br J Radiol 44(526):793–797

Marschner IC, Becker SL (2001) Interim monitoring of clinical trials based on long-term binary endpoints. Stat Med 20(2):177–192

O'Brien P, Fleming T (1979) A multiple testing procedure for clinical trials. Biometrics 35:549–556

Pedley A (2011) Applying survival analysis techniques to interim analysis and sample size reassessment of clinical trials with dichotomous endpoint. ProQuest, UMI Dissertations Publishing, Dissertation, Boston University

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 34(6):585–612

Pocock S (1977) Group sequential methods in the design and analysis of clinical trials. Biometrika 64:191–199

Pocock SJ (2005) When (not) to stop a clinical trial for benefit. JAMA 294 (17):2228–2230

Sooriyarachchi MR, Whitehead J, Whitehead A, Bolland K (2006) The sequential analysis of repeated binary responses: a score test for the case of three time points. Stat Med 25(12):2196–2214

Togo K, Iwasaki M (2013) Optimal timing for interim analyses in clinical trials. J Biopharm Stat 23(5):1067–1080

Vedantham S, Goldhaber SZ, Kahn SR, Julian J, Magnuson E, Jaff MR, Murphy TP, Cohen DJ, Comerota AJ, Gornik HL, Razavi MK, Lewis L, Kearon C (2013) Rationale and design of the ATTRACT Study: a multicenter randomized trial to evaluate pharmacomechanical catheter-directed thrombolysis for the prevention of postthrombotic syndrome in patients with proximal deep vein thrombosis. Am Heart J 165(4):530, e3

Whitehead A, Sooriyarachchi MR, Whitehead J, Bolland K (2008) Incorporating intermediate binary responses into interim analyses of clinical trials: a comparison of four methods. Stat Med 27(10):1646–1666

doi:10.1186/2193-1801-3-323
Cite this article as: Parpia *et al.*: **Interim analysis for binary outcome trials with a long fixed follow-up time and repeated outcome assessments at pre-specified times.** *SpringerPlus* 2014 **3**:323.