Increasing the power of randomized trials comparing different treatment durations

Yongdong Ouyang\textsuperscript{a,b,*}, Hong Qian\textsuperscript{b}, Lakshmi N. Yatham\textsuperscript{c}, Hubert Wong\textsuperscript{a,b}

\textsuperscript{a} School of Population and Public Health, University of British Columbia, 2206 East Mall, Vancouver, BC, V6T 1Z3, Canada
\textsuperscript{b} Centre for Health Evaluation and Outcome Sciences, 588-1081 Burrard Street, St Paul’s Hospital, Vancouver, BC, V6Z 1Y6, Canada
\textsuperscript{c} Department of Psychiatry, The University of British Columbia, 2255 Wesbrook Mall, Vancouver, BC, V6T 2A1, Canada

\textbf{ARTICLE INFO}

\textbf{ABSTRACT}

When the optimal treatment duration is uncertain, a randomized trial may allocate patients to receive active treatment for different durations. We use an example where patients receive treatment for 0, 24, or 52 weeks. In this trial, patients in the 24-weeks and 52-weeks arms receive the same treatment during the first 24 weeks. This overlap allows for more powerful analyses than conventional pair-wise comparisons of arms. When the outcome is the time-to-event, the power for the 0-weeks versus 24-weeks comparison can be increased by including patients in the 52-weeks arm as patients in the 24-weeks arm for the first 24 weeks and censoring at 24 weeks. Furthermore, differences observed between the 24-weeks and 52-weeks arms during the first 24 weeks can only reflect noise. Hence, the comparison of these two arms should be restricted to only patients who remain on the study at 24 weeks and include only the events after 24 weeks. Through simulation, we show that modified analyses accounting for these considerations increase study power substantially. Moreover, if patients were allocated equally to the arms, then events or discontinuations during the first 24 weeks will reduce the number of patients available for the 24-weeks versus 52-weeks comparison, and hence the power of this analysis will be lower than that for the 0-weeks versus 24-weeks comparison. We present a sample size calculation procedure for equalizing the power of these two analyses. Typically, this allocation requires much larger sample sizes in the 24-weeks and 52-weeks arms than in the 0-week arm.

\textbf{1. Introduction}

When the optimal duration for treatment is uncertain, a randomized trial may assign patients at random to receive active treatment for different lengths of time before discontinuation. Trials comparing two treatment durations are common, but they can be used to compare three (or more) treatment durations as well. Including three treatment durations in a single trial can be more cost- and time-efficient than conducting a sequence of two-arm trials. In these trials, patients may be allocated using either a single randomization into the three arms or a two-stage randomization design (TSRD) wherein a patient initially is randomized to one of two treatment periods and then those who were randomized to the longer of these two periods and who remain on study at the end of this period are randomized again to discontinue active treatment or to continue for a further period (i.e., into the third arm). Freedman and de Stavola \cite{1} provide an extended discussion comparing the potential benefits and disadvantages, both practical and methodological, of these two types of randomization designs.

In a trial with three or more treatment durations, there will be one or more initial ‘overlap’ periods during which subjects in multiple arms are all receiving active treatment. Any differences observed between the arms during this period can only reflect noise, not treatment effects. The idea that power can be improved by ignoring events during an overlap period has been discussed in the context of two-arm trials \cite{2}.

As a motivating example of a three-arm trial, presented in detail later, Yatham et al. \cite{3} conducted a double-blind, randomized trial to determine the optimal duration of atypical antipsychotic maintenance therapy in bipolar patients. In this trial, hereafter referred to as the “AAM trial,” patients entering manic remission following treatment with an atypical antipsychotic and a mood stabilizer combination were randomized to taper and discontinue their antipsychotic within 2 weeks (0-weeks arm), continue for 24 weeks (24-weeks arm) or continue for the entire 52 weeks (52-weeks arm) of follow-up. The primary outcome

\textsuperscript{*} Corresponding author. School of Population and Public Health, University of British Columbia, 2206 East Mall, Vancouver, BC, V6T 1Z3, Canada.

E-mail address: douyang@cheos.ubc.ca (Y. Ouyang).
was the time to relapse of a mood (manic or depressive) episode. Results from the AAM trial were reported with a primary analysis that consisted of the three pairwise comparisons of arms using standard Cox proportional hazards (PH) models applied to all data collected throughout the full trial follow-up period [4]; this approach generally would be regarded as the conventional choice for a trial with three arms. These analyses are sub-optimal for several reasons. From a clinical practice perspective, the two primary questions faced by the clinician are: (i) Should active treatment be given for the first 24 weeks? And (ii) given that a patient on active treatment has reached 24 weeks without experiencing the event, should active treatment be continued up to 52 weeks? Conventional pairwise comparisons of arms do not address these questions appropriately because these comparisons do not consider the two distinct time periods of interest or the time pattern of treatment delivery and focus on comparisons over the full follow-up period. Re-casting the hypotheses to be tested to match these questions leads to ways to improve the analysis.

To address the problem, Freedman and de Stavola [1] proposed using a time-varying coefficient Cox proportional-hazards model. Their analysis consisted of fitting all of the data to a single model from which the answers to the two primary questions can be extracted using appropriate contrasts on the regression coefficients. However, they did not quantify the potential gain in statistical power using their approach, and they did not consider sample size calculations for the two primary hypothesis tests. In the work presented here, the two primary questions will be addressed using two separate models to enable a more straightforward development and presentation of our sample size calculation method; however, the estimates will be equivalent to those from Freedman and de Stavola’s model. Such analyses were conducted by the authors of the AAM trial, but because these analyses yielded the same general conclusions as the conventional analysis, the authors judged that the simplicity of the conventional approach and its familiarity to readers would allow readers to more easily interpret the results. In the general case, the conclusions from the two analyses may differ, and thus, it is important to recognize when the more complex analysis is warranted.

Planning the sample size in these trials is challenging because the conventional sample size formulae do not consider two distinct follow-up periods and assume proportional hazards between arms throughout follow-up. Barthel et al. [5] have shown that when the proportional hazards assumption does not hold, sample size calculations making this assumption (e.g., based on a log-rank test or the standard Cox PH model) could substantially over- or underestimate the true required sample size. Additionally, if patients are allocated equally across the three arms, the events that occur during the first 24 weeks would significantly reduce the number of patients available for the 24-week versus 52-week comparison, leading to lower power for testing the 24-week versus 52-week comparison than for the 0-week versus 24-week comparison. This consideration is important for trial design since an underpowered primary comparison both jeopardizes the value of the trial and may be deemed unethical [6]. Obtaining equal power for both comparisons will require unequal sample sizes across the arms, but discussion of the required calculations has not appeared in the literature.

This article has two aims. In Aim One, we describe an efficient analysis plan which uses separate analyses for the two primary questions to identify the optimal treatment duration in a three-arm trial. Via simulation, we quantify the statistical power that is achieved using this analysis plan when the true hazard in the 24-week arm is different before and after 24 weeks, and we show that under these conditions, conventional pairwise comparison of arms yields substantially lower power. We illustrate the efficient analysis plan by applying it to data from the AAM trial and comparing the results with those obtained using the conventional analysis. In Aim Two, we derive and evaluate via simulation a sample size calculation procedure that nearly equalizes the powers of the two key comparisons of interest. In the concluding section, we discuss the implications of our results, limitations, and suggest potential future work.

2. Method

2.1. Aim one: improving power using efficient analyses

For simplicity in the following presentation, we will continue to refer to the three treatment duration arms as “0-weeks”, “24-weeks”, and “52-weeks”. Application of the results to trials with a different set of treatment durations requires only straightforward modifications. In this study, we assume the following hazard functions which use a step hazard function [7] that allows for different treatment effects in the periods before and after 24 weeks:

\[
\begin{align*}
0 - \text{week arm: } h(t) &= h_0(t); \\
24 - \text{week arm: } h(t) &= h_0(t)\exp(\beta - \gamma (t - t_1)); \\
52 - \text{week arm: } h(t) &= h_0(t)\exp(\beta),
\end{align*}
\]

where \(h_0(t)\) reflects the baseline hazard for patients who discontinue active treatment immediately. \(H(x)\) is a function that takes the value one when \(x\) is positive, and the value zero otherwise. In this study, \(t_1 = 24\). In this formulation, \(\beta\) and \(\gamma\) are the parameters of primary interest. The parameter \(\beta\) reflects the effect of active treatment compared to control during the first 24 weeks. The parameter \(\gamma\) reflects the effect of “long duration” treatment (i.e., continuing active treatment after 24 weeks) compared to “short duration” treatment (i.e., discontinuing after 24 weeks of active treatment) in the period after 24 weeks. The difference \(\beta - \gamma\) can be used to assess the carry-over effect of short duration treatment versus control in the period after 24 weeks but it is of secondary interest since, if a benefit of 24 weeks of active treatment is found, the focus switches to assessing the benefit of continuing active treatment (i.e., to estimating \(\gamma\)). Allowing the hazard to change at 24 weeks in the 24-weeks arm better reflects what plausibly occurs in the real world and allows for a clear interpretation of each estimated hazard ratio. In a conventional pairwise comparison of arms, the estimated hazard ratios are difficult to interpret as they result from averaging over the entire follow-up period over which proportionality of hazards should not be expected to hold. These hazard functions are equivalent to the ones in Freedman and de Stavola’s model but using a different parameterization.

2.1.1. Description of efficient analyses

Although a single Cox proportional hazards model can be fit to the full dataset, we propose splitting the analysis into two separate analyses that target each primary question separately. Doing so also enables clearer presentation of our proposed sample size calculation procedure. Our proposed ‘efficient analyses’ are:

Comparing active treatment versus control during the first 24 weeks of follow-up.

Using data from all three arms but with the data on individuals in the 52-weeks arm censored after 24 weeks of follow-up, fit a time-dependent treatment effect Cox proportional hazards model comparing events in the 0-weeks arm (control) to events in the 24-weeks and 52-weeks arms combined (active treatment) with the hazard functions described above. Note that this estimate for \(\beta\) could be obtained more simply by fitting a standard (i.e., without time-dependent treatment effects) Cox model using only data up to 24 weeks. However, fitting the more complex model enables assessment of the carry-over effect of short duration treatment.

Comparing active treatment for 52 weeks versus active treatment for 24 weeks after 24 weeks of follow-up.

Re-setting time zero to the end of the 24th week after the randomization, fit a Cox proportional hazards model that includes only patients in the 24-weeks and 52-weeks arms who remain on the study at the end of 24 weeks of follow-up and include only events that occur after this time-point. Since the individuals in these two arms receive the same
treatment during the first 24 weeks, they have the same probability of experiencing events (and other outcomes), so the distributions of characteristics of the individuals who are included in this analysis are expected to remain the same across the two arms.

2.1.2. Simulation study

We compared the power of our efficient analyses to the conventional pairwise analyses via simulation using R [8]. For each set of input parameters investigated, 10,000 datasets were simulated, and treatment comparisons were made using a likelihood ratio test with a two-sided constant hazard less than 0.5%.

Survival outcomes were generated assuming a constant hazard \( h_0 \) (i.e., an exponential survival function) while the patient was receiving the control treatment and a constant hazard \( h_1 = HR \times h_0 \) while receiving active treatment, where \( HR \) is the hazard ratio. Specifically, 0-week arm patients experience a hazard \( h_0 \) for the full 52 weeks of follow-up, 24-weeks arm patients experience a hazard \( h_1 \) for the first 24 weeks and a hazard \( h_0 \) from 24 weeks to 52 weeks, and 52-week arm patients experience a hazard \( h_1 \) for the entire 52 weeks (See Fig. 1). The assumption of a hazard \( h_0 \) during the 24 to 52-week period in the 24-weeks arm (i.e., setting \( \beta = \gamma \)) is made for convenience of illustrating the sample size calculation method described below. Adjusting that method to accommodate a different hazard is straightforward. It was assumed that no patient dropped out. A set of values for \( h_0 \) were selected to correspond to 52-week survival probabilities in the range 10%–90% in 10% increments. Hazard ratios were set in the range of 0.2–0.8 in 0.1 increments. The simulation scenarios covered all of the combinations of values of \( h_0 \) with \( HR \). In each scenario investigated, equal numbers of patients were allocated to each arm, and the total sample size was set so that 80% power would be achieved if the conventional analyses were used. The sample sizes were obtained using the ART package [9,10] in Stata [11], which accommodates with complex features including non-proportional hazards. See S.1 in the Supporting Information for additional simulation details.

2.2. Aim two: equalizing power through sample size reallocation

As noted earlier, events (and dropouts) occurring during the first 24 weeks reduce the number of patients available after 24 weeks for the 24-weeks vs. 52-weeks comparison. Therefore, if patients are allocated equally to the three arms, the 24-weeks vs. 52-weeks comparison will have lower power than the 0-weeks vs. 24-weeks comparison. If both of these comparisons are equally important, equalizing the power of the two comparisons requires the allocation algorithm to assign fewer patients to the 0-weeks arm and more to the 24-weeks and 52-weeks arms. The procedure presented below can be used for planning both single randomization and two-stage randomization designs.

Let \( N_1, N_2, \) and \( N_3 \) denote the number of patients randomized to the 0-weeks, 24-weeks, and 52-weeks arms, respectively. Since the sample size needed to achieve a desired power for the 24-weeks versus 52-weeks comparison does not involve \( N_1 \), we seek to calculate \( N_2 \) and \( N_3 \) independently from \( N_1 \). The complication is that the relevant numbers of at-risk patients are only those who remain on the study at 24 weeks and these numbers are random variables, \( N_2' \) and \( N_3' \), say. To simplify, we propose setting aside this complication and first determining \( N_2' \) and \( N_3' \) as if they were fixed numbers. Then we back-calculate to obtain \( N_2 \) and \( N_3 \) by dividing these sample sizes by the expected survival at \( t_1 = 24 \) weeks, i.e., setting \( N_2 = N_2'/\exp(-h_1 t_1) \) and \( N_3 = N_3'/\exp(-h_1 t_1) \) where \( h_1 \) is the (common, by design) hazard during the first 24 weeks in the 24-weeks and 52-weeks arms. For simplicity, we set \( N_2' \) and \( N_3' \) to be equal and solve using standard software (See S.3 in the Supporting Information for more details) [12-14].

For the 0-weeks versus 24-weeks comparison, the 24-weeks and 52-weeks arms are combined, so the power calculations involve only two groups with sample sizes \( N_1 \) and \( N_2+N_3 \). Several relatively simple algorithms have been proposed for power calculations in trials with unbalanced arms [15-18]. However, in our empirical assessment through simulation, we found that these algorithms do not produce accurate results when the sample sizes are greatly unequal (see Supporting Information Table S3 for an example). Instead, we use the power formulae developed by Strawderman [19], that is based on the second-order approximation to the distribution function for the two-sample log-rank statistic. These power formulae do not provide a closed-form expression for calculating the sample sizes, but given a value for \( HR \), the equations can be solved numerically using standard software (see S.3 for more details) [20]. To evaluate the performance of the proposed sample size re-allocation algorithm, we applied the algorithm to calculate the sample sizes required to achieve a target power (80% or 90%) and then simulated and analyzed data (10,000 replicates) based on these sample sizes to assess the empirical power. For comparison, we also conducted simulations assuming that the total sample size had been allocated equally among the three arms. Data were analyzed using the efficient analyses described in Section 2.1.1.

![Fig. 1. Assumed Hazard function in each arm in the data generator for the simulation studies. Patients in the 0-weeks arm receive placebo for the full 52 weeks with a constant hazard \( h_0 \). Patients in the 24-weeks arm receive active treatment for the first 24 weeks with hazard \( h_1 \), then receive placebo for the rest of follow-up time with hazard \( h_0 \). Patients in the 52-weeks arm receive active treatment for the full 52 weeks with hazard \( h_1 \).](image-url)
3. Results

3.1. Improving power using efficient analyses

3.1.1. Results of simulation study

Table 1 displays the power achieved by the proposed efficient analyses for selected values of $h_0$ and HR (see S.2 in Supporting Information for a complete table). The efficient analyses provided substantially greater power than the 80% power obtained using conventional analysis. For the 0-weeks versus 24-weeks comparison, the efficient analyses achieved near 100% power across every scenario tested. For the 24-weeks versus 52-weeks comparison, the power achieved by the efficient analysis ranged from roughly 90% when the hazard ratio was 0.2 up to more than 95% when the hazard ratio was over 0.5. At first glance, it may seem paradoxical that the power is lower when the treatment effect is higher (i.e., HR = 0.2 vs. HR = 0.5). However, these calculations are from scenarios with different sample sizes, so the power values are not comparable. The relatively low power when HR = 0.2 arises because, with this large effect, few people survive to the end of the 24th week, leading to very small sample sizes for the 24-weeks versus 52-weeks comparison.

3.1.2. An illustrative application of the efficient analysis

The AAM trial was a multi-center, double-blind, placebo-controlled trial that enrolled bipolar patients who recently remitted from an acute manic episode following treatment with a mood stabilizer (lithium or valproate) + an atypical antipsychotic (olanzapine or risperidone). Patients were recruited from 17 academic centers in Canada and collaborating sites in Brazil. Patients remained on their mood stabilizer as open label treatment for the full 52 weeks duration of the trial but were randomized to taper and discontinue the antipsychotic within two weeks at the start of the trial (0-weeks arm), after 24 weeks (24-weeks arm), or continue the antipsychotic for the full 52 weeks of follow-up (52-weeks arm). The atypical antipsychotic was blindly substituted with a placebo pill in the 0 weeks and 24 weeks arms. Thus, at no time did the patient or the healthcare providers know whether the patient was receiving the active or the placebo pill, nor did they know which pill the patient would receive in the period after 24 weeks. Randomization was stratified by center and by mood stabilizer/antipsychotic combination. The trial enrolled a total of 159 patients: 52 into the 0-weeks arm, 54 into the 24-weeks arm, and 53 into the 52-weeks arm. The primary outcome was the time to relapse by the arm. The number of patients who discontinued by study arm were 8 (0-weeks), 14 (24-weeks), and 12 (52-weeks). In particular, the number of patients who discontinued within the first 24 weeks were 9 (24-weeks) and 8 (52-weeks). Thus, after excluding patients who experienced the event or discontinued within the first 24 weeks, only 26 (24-weeks) and 22 (52-weeks) patients remained for the 24-weeks versus 52-weeks comparison. The upper section of Table 3 displays the results from a pair-wise comparison of the three arms with the shorter treatment duration arm as the reference. Assuming a conventional 5% significance level is used, these results indicate that the 24-weeks arm has better outcomes than the 0-week arm, and no difference is detected between the 24-weeks and 52-weeks arms. However, note that the Kaplan-Meier plot, the curves for the 24-weeks and 52-weeks arms, which are expected to behave similarly during the first 24 weeks in outcomes between the 24-weeks and 52-weeks arms during the first 24 weeks reflects only noise as participants in these two arms are receiving identical treatment during this period. Including these events in the comparison of the 24-weeks and 52-weeks arms potentially obscures the true differences in hazards between these two arms. As a caveat, although balance in patient characteristics is expected at 24 weeks, heterogeneity in patient responses may lead the trial to evolve in a way that creates some degree of imbalance during the first 24 weeks. Thus, this analysis should adjust for observed imbalances at 24 weeks. (Alternatively, this potential imbalance could have been mitigated at the design stage by adopting a TSRD). Third, the conventional analysis does not account for a potential sharp change in the hazard at the 24th week in the 24-weeks arm (arising from the discontinuation of the active treatment), which could lead to a violation of the proportional hazard assumption in the Cox model. Thus, the hazard in the 24-weeks arm should be modeled using a time-varying coefficient. The results presented here are intended to illustrate the use of the efficient analyses. The Cox models are adjusted for drug combination but do not consider potential heterogeneity in treatment effects. Readers seeking clinical interpretations are referred to the original publication [3], which contains more complete analyses and nuanced interpretations.

Table 2 shows the number of events in each arm during the first 24 weeks and after 24 weeks. Fig. 2 reproduces the Kaplan-Meier plot for the time to relapse by the arm. The number of patients who discontinued by study arm were 8 (0-weeks), 14 (24-weeks), and 12 (52-weeks). In particular, the number of patients who discontinued within the first 24 weeks were 9 (24-weeks) and 8 (52-weeks). Thus, after excluding patients who experienced the event or discontinued within the first 24 weeks, only 26 (24-weeks) and 22 (52-weeks) patients remained for the 24-weeks versus 52-weeks comparison. The upper section of Table 3 displays the results from a pair-wise comparison of the three arms with the shorter treatment duration arm as the reference. Assuming a conventional 5% significance level is used, these results indicate that the 24-weeks arm has better outcomes than the 0-week arm, and no difference is detected between the 24-weeks and 52-weeks arms. However, note that in the Kaplan-Meier plot, the curves for the 24-weeks and 52-weeks arms, which are expected to behave similarly during the first 24 weeks

Table 1

| 52-week Survival in the 0-week arm, % | Hazard Ratio ($h_0/h_0$) | 0-weeks vs 24-weeks Comparison | 24-weeks vs 52-weeks Comparison |
|--------------------------------------|--------------------------|-------------------------------|-------------------------------|
|                                      |                          | Sample size per arm needed for 80% power based on conventional analysis | Sample size per arm needed for 80% power based on conventional analysis |
|                                      |                          | Simulation power under efficient analysis, % | Simulation power under efficient analysis, % |
| 30                                   | 0.2                      | 67                            | 99.9                          | 40                            | 90.9                          |
| 60                                   | 0.2                      | 171                           | 99.9                          | 79                            | 90.0                          |
| 90                                   | 0.2                      | 878                           | 100.0                         | 343                           | 89.1                          |
| 30                                   | 0.5                      | 197                           | 99.4                          | 321                           | 97.1                          |
| 60                                   | 0.5                      | 487                           | 99.9                          | 355                           | 95.8                          |
| 90                                   | 0.5                      | 2449                          | 99.0                          | 1381                          | 94.6                          |
| 30                                   | 0.8                      | 1400                          | 99.1                          | 4309                          | 97.3                          |
| 60                                   | 0.8                      | 3352                          | 99.7                          | 6188                          | 96.9                          |
| 90                                   | 0.8                      | 16574                         | 99.9                          | 11929                         | 96.3                          |
because all of these patients have received identical treatment during this period, unexpectedly diverge with higher event rates in the 52-weeks arm. This raises two concerns: (1) if the divergence reflects underestimation of event rates in the 24-weeks arm, and if so, the 0-weeks arm versus 24-weeks arm comparison may be overestimating the efficacy, and (2) in the 24-weeks arm versus 52-weeks arm comparison, perhaps a beneficial effect of continued treatment after 24 weeks is being washed out by the “noise” effect in the opposite direction in the period before 24 weeks. The magnitude of the divergence is not large, so likely has limited impact on the results for this trial, but in trials that show larger divergences, the impact may be much greater. The lower section of Table 3 displays the results obtained using the efficient analysis. For the period before 24 weeks, the effect obtained by combining the data before 24 weeks from the 24-weeks and 52-weeks arms was reduced slightly compared to the conventional analysis (HR = 0.53 versus HR = 0.57), but the range of the 95% confidence interval (CI) and the p-value remained similar, and thus yielded the same qualitative conclusion. When events occurring after 24 weeks in the 24-weeks versus 52-weeks arms were compared, the effect obtained was similar to that found using the conventional analysis (HR = 1.18 versus HR = 1.03), indicating that the effect after 24 weeks was not greatly affected by the divergence seen during the first 24 weeks. (The characteristics of the two groups remaining on study at 24 weeks were similar, so no additional adjustment variables were added to the model.) Note that both the conventional and efficient analyses yielded a wide confidence interval. Thus, although the two analyses did not disagree, the power in this design was insufficient for assess whether continuing treatment beyond 24 weeks is beneficial.

3.2. Equalizing power through sample size reallocation

Table 4 shows the sample sizes and power achieved by the proposed

Fig. 2. Kaplan-Meier plot for the time to relapse in each of the three arms. Event rates during the first 24 weeks appeared to be higher in the 52-weeks arm than in the 24-weeks arm, when these rates were expected to be the same.

Table 3

| Time to any mood episode comparisons for all patients and by antipsychotic subgroup based on Cox analysis with adjustment for mood stabilizer antipsychotic. |
|---------------------------------|-----------------|-----------------|-----------------|
|                                 | HR              | 95% CI          | P               |
|                                 |                 |                 |                 |
| **Pair-wise Analysis**          |                 |                 |                 |
| 24-weeks vs. 0-weeks arms       | 0.53            | 0.33, 0.86      | 0.01            |
| 52-weeks vs. 0-weeks arms       | 0.63            | 0.39, 1.02      | 0.06            |
| 52-weeks vs. 24-weeks arms      | 1.18            | 0.71, 1.99      | 0.52            |
| **Efficient Analysis**          |                 |                 |                 |
| Early events:                   | 0.57            | 0.34, 0.93      | 0.02            |
| Late events:                    | 1.03            | 0.36, 2.93      | 0.95            |

HR = Hazard ratio, 95% CI = 95% confidence interval, P = p-value.

a Compares the 0-weeks arm versus the combined 24-weeks and 52-weeks arms during first 24 weeks of follow-up.

b Compares the 24-weeks arm versus the 52-weeks arm between 24 and 52 weeks of follow-up among only patients still on study after 24 weeks.
reallocation (Columns 4 through 8) and those obtained based on equal allocations (Columns 9 through 11). The proposed re-allocation algorithm achieved empirical powers that were nearly equal for the two comparisons and close to the target power. In the scenarios with low (30%) baseline survival rates, the samples size in each of the 24-weeks and 52-weeks groups needed to be from two to two and a half times larger than in the 0-weeks group. In the scenarios with high (90%) baseline survival rates, the 24-weeks and 52-weeks groups needed about one and a half times more observations. As expected, balanced allocation led to over-powering of the 0-weeks versus 24-week comparison and to under-powering of the 24-weeks versus 52-weeks comparison.

4. Discussion

We have argued that pair-wise comparison of arms, commonly used in multi-arm trials, is not optimal in trials comparing three or more treatment durations. In our example trial, we recast the hypotheses to reflect the ones most likely to be relevant to treating clinicians, namely (1) efficacy of 0 weeks vs 24 weeks of treatment during the first 24 weeks, and (2) efficacy of 24 weeks vs 52 weeks treatment during the 24–52 weeks period in patients still on treatment at 24 weeks. We have proposed fitting a pair of Cox models that address the two follow-up time periods separately. These models utilize the information contributed by the 52-weeks arm during the first 24 weeks when assessing efficacy during the period before 24 weeks and removes the noise introduced during the first 24 weeks when assessing efficacy in the period after 24 weeks. The models are equivalent to the single model proposed by Freedman and de Stavola [1] and thus share the advantages of having more interpretable hazard ratios and increased power than the conventional analysis. This work extends Freedman and de Stavola [1] by quantifying the magnitude of the gains in power that can be expected. For a trial designed to obtain 80% power based on conventional analysis, the efficient analysis achieves greater than 95% power for the 0-weeks versus 24-weeks comparison and greater than 90% power for the 24-weeks versus 52-weeks comparison, across a wide range of conditions that are encountered in real trials.

We have argued that using a balanced allocation of patients to the three treatment arms leads to a much higher power for the 0-weeks versus 24-weeks comparison than for the 24-weeks vs. 52-weeks comparison. The second main contribution of this work is the proposed sample size allocation algorithm that will approximately equalize the powers for testing the two primary hypotheses. As seen in the simulation results, equalizing power requires a highly unbalanced design, from roughly 40% more up to more than twice as many patients in each of the 24-weeks and 52-weeks arms compared to the 0-weeks arm, depending on the specific scenario. The optimal treatment duration question also has arisen in trials evaluating the efficacy of a new agent for induction and maintenance therapy. Freidlin et al. [21] review trial designs for this problem. These designs include both the single-randomization three-arm trial in which participants are allocated to receive the new agent during both the induction and maintenance phases, during the induction phase only or during neither phase as well as the 2 × 2 “phase x new agent” factorial design (typically utilizing a TSRD). In the 2 × 2 design, removal of the arm receiving the new agent only during the maintenance phase results in a three-arm trial design with three different durations for the new agent. This design has been criticized as being potentially inefficient for assessing the maintenance effect for the relatively low proportion of patients who ultimately receive the new agent during the maintenance phase. Our proposed sample size calculation procedure provides one approach to remedying this shortcoming.

From a mathematical perspective, our approach is the same irrespective of whether the underlying objective is to compare different treatment durations when efficacy has already been established, or to test also for efficacy (by setting the active treatment duration to zero in one arm). For the latter objective, investigators may wish to design the trial to provide greater power for the test of efficacy. This is easily accommodated by simply changing the input power values to the desired values in our proposed sample size calculation algorithm. If single randomization is used and treatment allocations are not blinded, selection bias may be introduced into the 24-weeks versus 52-weeks comparison if outcomes during the first 24 weeks are influenced by knowledge of future treatment allocation. Another potential concern is that the number of patients that enter into the 24-weeks versus 52-weeks comparison is the result of a random process, and there is a risk that the attained power (conditional on the achieved sample sizes) may be lower than the power calculated assuming the expected sample sizes [22]. Although we have presented an example trial that compared only three treatment durations, the extension to trials involving more treatment durations is straightforward. For data analysis, the same step hazard function for the time-varying treatment effect Cox PH model can be used, and data from different arms can be pooled in the way described in this paper. Sample size calculations to equalize the powers for the comparisons of interest would begin with the calculation for comparing one treatment duration only or during neither phase as well as the 2 × 2 “phase x new agent” factorial design (typically utilizing a TSRD). In the 2 × 2 design, removal of the arm receiving the new agent only during the maintenance phase results in a three-arm trial design with three different durations for the new agent. This design has been criticized as being potentially inefficient for assessing the maintenance effect because of the relatively low proportion of patients who ultimately receive the new agent during the maintenance phase. Our proposed sample size calculation procedure provides one approach to remedying this shortcoming.

We have argued that using a balanced allocation of patients to the three treatment arms leads to a much higher power for the 0-weeks versus 24-weeks comparison than for the 24-weeks vs. 52-weeks comparison. The second main contribution of this work is the proposed sample size allocation algorithm that will approximately equalize the powers for testing the two primary hypotheses. As seen in the simulation results, equalizing power requires a highly unbalanced design, from roughly 40% more up to more than twice as many patients in each of the 24-weeks and 52-weeks arms compared to the 0-weeks arm, depending on the specific scenario. The optimal treatment duration question also has arisen in trials evaluating the efficacy of a new agent for induction and maintenance therapy. Freidlin et al. [21] review trial designs for this problem. These designs include both the single-randomization three-arm trial in which participants are allocated to receive the new agent during both the induction and maintenance phases, during the induction phase only or during neither phase as well as the 2 × 2 “phase x new agent” factorial layout (typically utilizing a TSRD). In the 2 × 2 design, removal of the arm receiving the new agent only during the maintenance phase results in a three-arm trial design with three different durations for the new agent. This design has been criticized as being potentially inefficient for assessing the maintenance effect because of the relatively low proportion of patients who ultimately receive the new agent during the maintenance phase. Our proposed sample size calculation procedure provides one approach to remedying this shortcoming.

From a mathematical perspective, our approach is the same irrespective of whether the underlying objective is to compare different treatment durations when efficacy has already been established, or to test also for efficacy (by setting the active treatment duration to zero in one arm). For the latter objective, investigators may wish to design the trial to provide greater power for the test of efficacy. This is easily accommodated by simply changing the input power values to the desired values in our proposed sample size calculation algorithm. If single randomization is used and treatment allocations are not blinded, selection bias may be introduced into the 24-weeks versus 52-weeks comparison if outcomes during the first 24 weeks are influenced by knowledge of future treatment allocation. Another potential concern is that the number of patients that enter into the 24-weeks versus 52-weeks comparison is the result of a random process, and there is a risk that the attained power (conditional on the achieved sample sizes) may be lower than the power calculated assuming the expected sample sizes [22]. Although we have presented an example trial that compared only three treatment durations, the extension to trials involving more treatment durations is straightforward. For data analysis, the same step hazard function for the time-varying treatment effect Cox PH model can be used, and data from different arms can be pooled in the way described in this paper. Sample size calculations to equalize the powers for the comparisons of interest would begin with the calculation for comparing the two longest treatment durations and then proceed by working backward iteratively for each subsequent comparison under the assumption the number of at-risk patients is fixed at the expected survival counts. All of these calculations can be performed in a manner analogous to what has been illustrated here. A practical concern is that trials comparing larger numbers of treatment durations may not be feasible as the required sample sizes typically will grow quickly.

### Table 4

Unbalanced sample size allocations are needed to equalize the power of the 0-weeks versus 24-weeks and 24-weeks versus 52-weeks comparisons. \( N_1 \) & \( N_2 \) & \( N_3 \) were calculated using ‘powerSurvEpi’. \( N_1 \) was calculated using Strawderman’s formula. Equal allocation over-powers the 0-weeks versus 24-weeks and under-powers the 24-weeks versus 52-weeks comparisons.

| Baseline Survival Rate (%) | Hazard Ratio | Target Power | Unequal allocation | Power, % | Equal allocation |
|----------------------------|--------------|--------------|--------------------|---------|-----------------|
|                            |              |              | \( N_1 \) | \( N_2 \) | \( N_3 \) | 0 vs 24 | 24 vs 52 | \( N \) per arm | Power, % |
| 30                         | 0.5          | 80%          | 60 125 125        | 80.4 82.2| 104 90.8 74.2 |
| 30                         | 90%          | 79 167 167   | 89.8 91.7        | 138 97.0| 85.6           |
| 60                         | 80%          | 127 217 217  | 79.2 81.4        | 187 87.3| 75.3           |
| 60                         | 90%          | 167 290 290  | 88.0 90.9        | 249 94.0| 86.6           |
| 90                         | 80%          | 577 871 871  | 77.8 80.1        | 773 85.3| 75.6           |
| 90                         | 90%          | 757 1167 1167| 87.4 90.6        | 1031 92.7| 86.7           |
|                            | 0.8          | 80%          | 486 1127 1127    | 80.1 80.3| 913 93.2 71.1 |
| 30                         | 90%          | 646 1570 1570| 89.6 90.1        | 1262 98.0| 84.2           |
| 60                         | 80%          | 1051 1754 1754| 79.6 80.8        | 1520 86.7| 72.8           |
| 60                         | 90%          | 1396 2348 2348| 90.0 90.2        | 2031 94.7| 84.6           |
| 90                         | 80%          | 4862 6642 6642| 78.7 79.9        | 6049 83.9| 76.7           |
| 90                         | 90%          | 6459 8892 8892| 89.0 90.1        | 8081 92.3| 87.5           |
In this study, we assumed a simple data generator wherein the hazard was a constant value $h_0$ while a patient was off active treatment and a different constant value $h_1$ while on active treatment and analyzed data using the Cox proportional hazards model. As discussed in Barthel et al. [5], violation of assumptions could lead to under- or over-estimated assessments of power. We expect that our proposed approach will also improve the efficiency in studies with more complex data generators and models (time-varying treatment effects, non-proportionality, etc.), but additional investigation is needed to determine the magnitude of benefit. In addition, extending our sample size calculation approach to allow for dropouts would be useful.

In summary, this study has shown that in trials seeking to compare three or more treatment durations, balanced designs, and analysis based on pairwise comparison of treatment arms using the conventional Cox proportional hazards model are not efficient and may yield hazard ratios that are difficult to interpret. The efficiency of the trial can be improved greatly by using an unbalanced design and analysis with appropriately parameterized time-varying treatment effect Cox proportional hazards models.

Availability of data

Data from AAM trial will not be shared. At the time the trial was initiated (nearly 20 years ago), open access to trial data was not common nor expected; no agreements were made that allow for sharing of the data. Simulation codes are available from the authors on request.

Authors contributions

YO conducted computational work, contributed to the development of the paper’s ideas, drafted the initial manuscript, and contributed to the critical revision of the manuscript. HQ conducted the original analysis of the example trial and contributed to the critical revision of the manuscript. LNY contributed to the development of the paper’s ideas and contributed to the critical revision of the manuscript. HW conceived the ideas of this paper and contributed to the critical revision of the manuscript.

Funding sources

The research is funded by Canadian Institutes of Health Research (Grant ID: CIHR MCT-94835).

Ethics and patient consent

Not applicable.

Declaration of competing interest

L.N.Y. has received research support from or served as a consultant or speaker for Alkermes, Allergan, CANMAT, CIHR, Dainippon Sumitomo, Forest, Janssen, Lundbeck, Otsuka, Sanofi, Sunovion, Teva, and Valeant. Other authors declare that there is no conflict of interest.

Acknowledgments

We thank Dr. Ehsan Karim for constructive comments during the drafting of this manuscript and Dr. Robert Strawderman for sharing his power calculation code.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2020.100588.

References

[1] L.S. Freedman, B.L. de Stavola, Comparing the effects of different durations of the same therapy, Stat. Med. 7 (1988) 1013–1021, https://doi.org/10.1002/ sim.4780071003.

[2] J.A. Kopec, M. Abrahamowicz, J.M. Esdaile, Randomized discontinuation trials: utility and efficiency, J. Clin. Epidemiol. 46 (1993) 959–971, https://doi.org/10.1016/0895-4356(93)90163-u.

[3] L.N. Yatham, S. Beaulieu, A. Schaffer, M. Kaser-Sant’Anna, F. Kapczinski, B. Lafer, V. Sharma, S.V. Parikh, A. Daigneault, H. Qian, D.J. Bond, P.H. Silverstone, N. Wolji, R. Milve, P. Baruch, A. da Gumba, J. Quevedo, R. Dias, M. Kunz, L.T. Young, R.W. Lam, H. Wong, Optimal duration of risperidone or olanzapine adjunctive therapy to mood stabilizer following remission of a manic episode: a CANMAT randomized double-blind trial, Mol. Psychiat. 21 (2016) 1050–1056, https://doi.org/10.1038/mp.2015.156.

[4] D.R. Cox, Regression models and life-tables, J. Roy. Stat. Soc. B 34 (1972) 187–220.

[5] F.M.-S. Barthel, A. Babiker, P. Royston, M.K.R. Parmar, Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over, Stat. Med. 25 (2006) 2521–2542, https://doi.org/10.1002/sim.2517.

[6] S.D. Halpern, J.F.T. Karlswich, J.A. Berlin, The continuing unethical conduct of underpowered clinical trials, J. Am. Med. Assoc. 288 (2002) 358–362, https://doi.org/10.1001/jama.288.3.358.

[7] T. Therneau, C. Cromwell, E. Atkinson, Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model, (accessed 15 Jun 2017).

[8] R. R Development Core Team, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. URL, http://www.R-project.org. (Accessed 20 June 2018).

[9] P. Royston, A. Babiker, A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome, STATA J. 2 (2002) 151–163, https://doi.org/10.1177/1536867X0200200204.

[10] F.M.-S. Barthel, P. Royston, A. Babiker, A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome: update, STATA J. 5 (2005) 123–129, https://doi.org/10.1177/1536867X0500500114.

[11] StataCorp. Stata Statistical Software: Release 15, StataCorp LLC, College Station, TX, 2017, https://www.stata.com/. (Accessed 30 January 2018).

[12] L.S. Freedman, Tables of the number of patients required in clinical trials using the logrank test, Stat. Med. 1 (1982) 121–129, https://doi.org/10.1002/ sim.4780010204.

[13] B. Rosner, Fundamentals of Biostatistics, 6th edition, Thomson Brooks/Cole, 2006.

[14] W. Qiu, J. Chavarro, R. Lazarus, B. Rosner, J. Ma, powerSurvEpi: power and sample size calculation for survival analysis of epidemiological studies. https://CRAN.R-project.org/package=powerSurvEpi. 2018.

[15] D. Schoenfeld, The asymptotic properties of nonparametric tests for comparing survival distributions, Biometrika 68 (1981) 316–319, https://doi.org/10.2307/2335833.

[16] F.M.-S. Barthel, A.B. Babiker, A.R. Parmar, Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over, Stat. Med. 25 (2006) 2521–2542, https://doi.org/10.1002/sim.2517.

[17] J.J. Shuster, CRC Handbook of Sample Size Guidelines for Clinical Trials, CRC Press, 2019, https://doi.org/10.1201/9781351508644.

[18] F.M.-S. Barthel, Comparing survival analysis methods using the logrank test with unequal allocation, Stat. Med. 11 (1992) 1091–1098.

[19] R.L. Strawderman, An asymptotic analysis of the logrank test, Lifetime Data Anal. 3 (1997) 225, https://doi.org/10.1023/A:1009648914586.

[20] R.P. Brent, Algorithms for Minimization without Derivatives, Courier Corporation, 2013.

[21] B. Freidlin, R.F. Little, E.L. Korn, Design issues in randomized clinical trials of maintenance therapies, J. Natl. Cancer Inst. 107 (2015), https://doi.org/10.1093/jnci/djv225.

[22] H. Wong, Y. Ouyang, M.E. Karim, The randomization-induced risk of a trial failing to attain its target power: assessment and mitigation, Trials 20 (2019) 360, https://doi.org/10.1186/s13063-019-3471-8.