International Journal of Statistics and Probability; Vol. 6, No. 1; January 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Statistical Monitoring of Clinical Trials Using Brownian Bridges

Qiang Zhang[1,2] & Michael R. Kosorok[3]

[1] American College of Radiology, 1818 Market St, Suite 1720, Philadelphia, Pennsylvania, 19103, U.S.A

[2] Department of Biostatistics & Epidemiology, University of Pennsylvania, Philadelphia Pennsylvania, U.S.A

[3] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A

Correspondence: Qiang Zhang, American College of Radiology, 1818 Market St, Suite 1720, Philadelphia, Pennsylvania, 19103, U.S.A. E-mail: Ed.Zhang.Jr@gmail.com

Received: October 26, 2016 Accepted: December 2, 2016 Online Published: December 27, 2016

doi:10.5539/ijsp.v6n1p142 URL: https://doi.org/10.5539/ijsp.v6n1p142

Abstract

The Brownian bridge is not yet used widely in the statistical monitoring of clinical trials. In this paper, we investigate properties of the Brownian bridge and formally derive monitoring rules from these results. We will present four related main methods: (1). derivation of group sequential boundaries; (2). calculation of conditional power; (3). a new alpha spending function and (4). repeated confidence intervals, all under a Brownian bridge framework. Simulation results show that the type I error rate is well controlled and power is satisfactory for the group sequential design. We apply the proposed methods to monitor the interim results from the Beta Blocker Heart Attack Trial (BHAT) and a Head and Neck cancer trial with comparisons to the commonly used monitoring tools. Overall, the proposed methods when used together as one framework are more powerful and sensitive to interim positive and negative trends that are clinically meaningful and lead to timely early stopping with potentially more savings on sample sizes, time and costs. These tools are valuable additions to the existing group sequential methods which can be utilized in trial design, routine monitoring, and to answer important questions from data monitoring committees.

Keywords: Brownian motion, Brownian bridge, group sequential design, log-rank test

1. Introduction

There are multiple methods developed for the interim monitoring of clinical trials, these include, for example, the well-known Pocock (1977) and O'Brien-Fleming (OBF, 1979) designs, alpha spending function (Lan & Demets, 1983), stochastic curtailment (conditional power, CP, Davis & Hardy, 1990; Zhang, Lai, & Davis, 2015) and repeated confidence intervals (RCI, Jennison & Turnbull, 2000, Zhang, 2011). Most of these methods are built upon Brownian motion (BM). A well-known property regarding the original OBF boundary is that it has a conditional power of only approximately 50% according to the BM. For example, at 20% of the information time, the boundary value is 3.92 for a five-stage OBF design with 5% error rate, however, if we base this on rejecting the null hypothesis at 80% conditional power, the interim Z value would have to be greater than 5.35 under the BM. This may be an over-conservative stopping rule. For the interim monitoring of survival data, Tsiatis (1981) proved that under the proportional hazards assumption, the sequential log-rank statistics follow a Brownian motion process with the independent increment property. The Brownian bridge (BB), also known as the conditional Brownian motion, given past observed and future expected observations is fundamental and key to studying the distribution of empirical and stochastic processes including, for example, inference and interim monitoring on the quantiles for the survival distributions (Kosorok, 1999, 2008). Doob (1949) demonstrated the transformation between a Brownian motion and a Brownian bridge for proving the Kolmogorov-Smirnov theorems. A recent review of this concept can be found in Chow (2009), in which it was pointed out although Brownian motion and Brownian bridges differ by a single constraint the unique properties of Brownian bridges enable them to serve as stable statistical models of real world experiences. In fact, Brownian bridges are the natural framework for studying the limiting distributions of empirical processes (see, for example, Section 2.1 of Kosorok, 2008). As mentioned earlier, there is an inconsistency between the commonly used stopping boundaries with a 50% conditional power and the typical expected 80% or higher conditional power requirement. An ideal solution would be finding stopping rules with satisfactory conditional power and a reasonable critical value that represents an achievable clinical difference. In addition, we need statistical monitoring methods that are more sensitive to early evident positive or negative trends as we will see in the BHAT and oropharyngeal cancer trial examples. In this paper, we aim to first demonstrate that Brownian bridges can be used for the interim monitoring of clinical trials through group sequential methods. We will then show that with a simple transformation and the resulting smaller variance at each interim analysis, Brownian bridges are a more powerful tool for this purpose when using the conditional power procedure. This means we would be able to obtain designs with monitoring rules that possess the above mentioned two properties. Here, for the first time, we derive four trial monitoring methods for efficacy or futility monitoring using properties of Brownian bridges to help design a clinical study and answer important questions for data monitoring committees. We will consider two-sample comparison of survival data using the log-rank test as an example.

2. Methods

Assume we have two-sample right censored data from independent pairs of failure and censoring times, $\{T_{ik}, C_{ik}\}, i = 1, \dots, n_k$ for samples $k = 1, 2$ and $n = n_1 + n_2$, where we observe $\{X_{ik}, \delta_{ik}\}$, where $X_{ik} = T_{ik} \wedge C_{ik}$ and $\delta_{ik} = 1$ if $X_{ik} = T_{ik}$ or $\delta_{ik} = 0$, otherwise. Under the null hypothesis, $H_0: d\Lambda_1(t) = d\Lambda_2(t)$, in which t is the calendar time and $\Lambda_1(t)$, $\Lambda_2(t)$ are the cumulative hazard functions, the test statistic is defined as:

$$Z_n(t) = n^{-\frac{1}{2}} \int_0^t \widehat{W_n}(u) \frac{\overline{Y_1}(u)\overline{Y_2}(u)}{\overline{Y_1}(u) + \overline{Y_2}(u)} (\frac{d\overline{N_1}(u)}{\overline{Y_1}(u)} - \frac{d\overline{N_2}(u)}{\overline{Y_2}(u)})$$

as expressed in counting process integral form (Fleming & Harrington, 1991). $\overline{N_k}$ and $\overline{Y_k}$ are event counting and at risk processes for groups k = 1, 2, respectively. $\widehat{W_n}$ is the estimated weight function and n is the sample size. When $\widehat{W_n} = 1$, we have the usual log-rank test, which we will focus on in the current paper. The variance of $Z_n(t)$ is consistently estimated by

$$\sigma_n^2(t) = n^{-\frac{1}{2}} \int_0^t [\widehat{W_n}(s)]^2 \frac{\overline{Y_1}(s)\overline{Y_2}(s)}{\overline{Y_1}(s) + \overline{Y_2}(s)} (\frac{d\overline{N_1}(s) + d\overline{N_2}(s)}{\overline{Y_1}(s) + \overline{Y_2}(s)})$$

The asymptotic distribution of the standardized process $T_n(s) = Z_n(s)/\sigma_n(t)$, where t is the upper limit of integration, has been shown to follow Brownian motion (formula 5 in Eng & Kosorok, 2005). Basic properties for BM that are important for clinical trial monitoring can be found in Lan and Wittes (1988).

Let Z be the interim test statistic at information time t_{BM}, then $W(t_{BM}) = Z\sqrt{t_{BM}}$ is Brownian motion for $0 \leq t_{BM} \leq 1$ (e.g., observed events at the interim analysis divided by the total expected number of events for survival data), then $B(t) = (1-t)W(\frac{t}{1-t})$ follows a stationary Markovian process and is a Brownian bridge. Here t is the transformed time from t_{BM}. Important properties of the BB process can be found in Doob (1949) and Chow (2009):

(i) $E(B(t)) = 0, E[B(t)B(s)] = t \wedge s - ts$ for all $t, s \in [0,1]$.

(ii) $for\ 0 \leq t_1 \leq t_2 \leq t_3 \leq t_4 \leq 1, E\{[B(t_4) - B(t_3)] - [B(t_2) - B(t_1)]\} = (t_1 - t_2)(t_4 - t_3)$, which are dependent increments.

(iii) The construction of BB can be defined for $t \in [0, T]$ with covariance $t \wedge s - \frac{ts}{T}$.

(iv) The BB with a drift μt is defined as $B^{\mu}(t) = (1-t)W(\frac{t}{1-t}) + \mu t$.

(v) The BB has continuous sample paths almost surely.

First, group sequential designs such as those proposed by Pocock (1977) and O'Brien-Fleming (1979), especially the latter are among the most commonly used methods for the design and monitoring of efficacy in confirmative clinical trials. We aim to first show that the Brownian bridge can be utilized to obtain group sequential monitoring boundaries based on the above properties. This is an inevitable first step before we derive the other three methods including the more powerful conditional power approach which are often used together to make interim decisions regarding trial outcomes. Note that for a standard BM transformed to BB, we need to shift the time domain from [0,1] to $[0, \frac{1}{2}]$ for application to group sequential trials. Here the required number of events for the current interim analysis before and after the transformation is the same. The covariance in (i) is smaller than that of BM: $E[W(t_{BM})W(s_{BM})] = t_{BM} \wedge s_{BM}$. For $t \in [0, \frac{1}{2}]$, we can derive stopping boundaries based on the joint distributions of $B(t_1), B(t_2), \dots, B(t_J)$ at each of the planned interim analysis times. For a trial with three interim analyses, we have $t = t_1, t_2, t_3$ then at each of the three analyses, the joint distributions under the null are:

$$B(t_1) \sim N\{0,\ (1-t_1)t_1\},$$

$$B(t_1), B(t_2) \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{Bmatrix}(1-t_1)t_1 & (1-t_2)t_1\\(1-t_2)t_1 & (1-t_2)t_2\end{Bmatrix}\right],$$

$$B(t_1), B(t_2), B(t_3) \sim N\left[\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{Bmatrix}(1-t_1)t_1 & (1-t_2)t_1 & (1-t_3)t_1\\(1-t_2)t_1 & (1-t_2)t_2 & (1-t_3)t_2\\(1-t_3)t_1 & (1-t_3)t_2 & (1-t_3)t_3\end{Bmatrix}\right],$$

Given an alpha spending function, such as $(t) = \alpha t^{\lambda}$ and the above joint distributions, group sequential stopping boundaries for each analysis can be found as follows for a one-sided hypothesis $d\Lambda_1(t) > d\Lambda_2(t)$:

For *j*=1, $1 - \int_{-\infty}^{c_1} f(B_1) db_1 = \alpha(t_1)$,

For *j*=2, $\int_{-\infty}^{c_1} \int_{c_2}^{\infty} f(B_1, B_2) db_1 db_2 = \alpha(t_2) - \alpha(t_1)$,

For *j*=3, $\int_{-\infty}^{c_1} \int_{-\infty}^{c_2} \int_{c_3}^{\infty} f(B_1, B_2, B_3) db_1 db_2 db_3 = \alpha(t_3) - \alpha(t_2)$.

Hence the overall alpha is protected at the design level, and the stopping boundaries for more than 3 stages can be found by following similar algorithms. Three types of spending functions are used here to represent relatively aggressive ($\lambda = 1$, Pocock design type), intermediate ($\lambda = 2$) and conservative ($\lambda = 3$, O'Brien-Fleming design type) interim monitoring plans. We will apply these stopping boundaries to the simulated two-sample survival data under the null and alternative hypotheses and discuss the type I error rate and power in the results section. We will also illustrate how to monitor interim data from BHAT using these methods.

Second, since conditional power is one of the most important tools for both efficacy and futility monitoring of clinical trials (Lan & Wittes, 1988) which provides additional insights regarding interim data, we derive conditional power under BB here and demonstrate in the two clinical trial examples later that it is a more powerful and sensitive tool due to the smaller variances as we pointed out earlier. As in Lan and Wittes (1988), write $B(1/2) = B(t) + B(1/2) - B(t)$, the increment $B(1/2) - B(t)$ has mean and variance:

(vi) $E\left\{B\left(\frac{1}{2}\right) - B(t)\right\} = \frac{\mu}{2} - \mu t$

(vii) $var\left\{B\left(\frac{1}{2}\right) - B(t)\right\} = var\left\{B\left(\frac{1}{2}\right)\right\} + var\{B(t)\} - 2cov\left\{B\left(\frac{1}{2}\right), B(t)\right\} = \frac{1}{4} + t(1-t) - 2\left(t - \frac{t}{2}\right) = \frac{1}{4} - t^2$

Therefore, the conditional distribution of $B\left(\frac{1}{2}\right) | B(t) \sim N(b + \mu/2 - \mu t, 1/4 - t^2)$, and the conditional power is $CP_{BB}(t) = 1 - \Phi\left\{\frac{Z_{\frac{\alpha}{2}} - b - \left(\frac{\mu}{2} - \mu t\right)}{\sqrt{\frac{1}{4} - t^2}}\right\}$, whereas the variance for the increment under BM is greater, $1 - t_{BM}$. In the real trial examples, we will show how to calculate the non-sequential conditional power under BB for efficacy and futility monitoring and compare the results to the conditional power under BM. For illustration purpose, we will calculate conditional power for one interim analysis as shown above for the rest of the paper. We can follow similar algorithms as in Zhang et al. (2015) when there are multiple interim analyses in practice.

Third, the alpha spending function is another key method that is frequently used to derive interim monitoring plans that are flexible in terms of the exact timings of the analyses. Lan and DeMets (1983) defined an alpha spending function based on the supremum of a BM. The supremum of Brownian bridge can also be used to obtain a new alpha spending function with different design characteristics. According to Borodin and Salminen (p67, 2002), the distribution of the supremum of $B(t)$ is as follows.

(viii) $P[\sup_{0\le s\le t} B(t) \le u] = 1 - \exp(-2\frac{u^2}{t})$,

By following the method in Lan and DeMets (1983), we can find critical value $c = \sqrt{-\frac{\log(\alpha)}{2}} = 1.36$ if $\alpha = 0.025$ when we set t_{BM} = 1, such that all alpha is spent at the end of the study. Here let $\alpha(t_{BM} = 0) = 0$. We will compare this new spending function $\alpha_{BB}(t_{BM}) = e^{-\frac{1.36^2}{t_{BM}}}$ to the power family spending function in the BHAT example and discuss how it can be used in different clinical trial scenarios.

Finally, Jennison and Turnbull (2000) described the method of repeated confidence intervals to monitor clinical trials. It is a well-known approach for both efficacy and futility analyses, especially the latter in oncology trials. For this, we need to extend our algorithm above for method one to a two-sided test setting as in Zhang (2011). Repeated confidence intervals will be constructed based on the derived boundaries. Briefly, we determine the boundary value b_j by $P\left(\left|B(t_j)\right| < b_j, j = 1, \dots, i-1, B(t_i) > b_i\right) = \frac{\alpha(t_i)-\alpha(t_{i-1})}{2}$, where $t_1, t_2, \dots, t_i$ are the interim analysis times. According to Jennison & Turnbull (2000), The RCIs for a parameter δ are derived using the above boundary values as a sequence of intervals Int_j, $j = 1, \dots, K$: $P_\delta\left(\delta \in Int_j\ for\ all\ j = 1, \dots, K\right) = 1 - \alpha$. So, the RCI at stage k is $\left\{\frac{Z_k - b_k}{\sqrt{I_k}}, \frac{Z_k + b_k}{\sqrt{I_k}}\right\}$, in which Z_k, b_k and I_k are the test statistic, the boundary value derived from two-sided group sequential test, and information at stage k, respectively; and, for an unadjusted interval, b_k is replaced by $\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$. If we define boundary values under BM and BB models at stage k as b_k^{BM} and b_k^{BB}, then the ratio of the width of the two confidence intervals is expressed as $\frac{b_k^{BB}}{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}$ when comparing a BB and a fixed design. While for comparison between a BB and a BM, the ratio is b_k^{BB}/b_k^{BM}. In the BHAT example, we will show how to monitor interim data using RCIs under BB and BM.

3. Simulations

For method one, we derive stopping rules with type I error rates of 0.05 and 0.01 using the algorithm in the method section and assess the statistical power using sample sizes 249 and 170 as derived in Eng and Kosorok (2005) for each error rate. We also evaluate the statistical power under a flexible class of alternatives. Specifically, we simulate survival data according to scenarios I and II, and conduct interim analyses using boundaries derived as above. In scenario I, the control group and the treatment group have a constant hazard of 1.6 and 0.8. For scenarios II, the hazard function is kept the same for the control group, while the treatment group has a hazard function of 0.4 over the time interval [0, 1/3], 2.2 over the interval (1/3, 1], and 1.6 afterwards. We invoke uniform censoring between [0, 1] and patients are randomized equally into each group. In scenario II, the hazard functions cross but the survival functions are ordered and the treatment is beneficial. We plan to first verify type I error rates for group sequential designs under BM and BB. Each set of simulations has 500 trials and is repeated 100 times, Monte Carlo standard errors are calculated based on these results. Powers for the regular log-rank test are investigated to compare the design performance for both scenarios. In scenario II, we intend to examine the design performance when there is a transitory treatment effect, a more realistic case in many settings. We will apply trial monitoring methods 1-4 under BB to the clinical trial examples. All simulations and computations are done using R 3.2.0.

4. Results

For this research, we conduct two interim analyses at information times 0.33 and 0.67 for the simulated two-sample survival data. Under the null hypothesis, the error rates (Table 1) are controlled at the design levels with Monte Carlo standard errors ranging from 0.3% to 0.8%. For example, with a one-sided type I error rate of 0.025 for the OBF design type, the overall false positive rate is 0.024 (0.007) under BM and 0.025 (0.007) under BB. In general, the type I error rates based on the BB are well controlled and are similar to those from boundaries under BM. Under the alternative hypothesis for scenario I (Table 2, first and third rows), the statistical powers for each model are mostly the same. For example, for a design with alpha of 0.025 and 80% power according to Eng and Kosorok (2005), the statistical powers are 0.85 (0.018) under BM and 0.85 (0.018) under BB for the OBF design type. In scenario II (Table 2, second and fourth rows), as expected most of the powers are higher than those in scenario I, except for the OBF design, for which these are slightly lower or similar. For the same design above the powers are 0.81 (0.017) under BM and 0.81 (0.018) under BB. The two powers are almost identical but lower than those under scenario I. This could be due to the early time-dependent treatment effects and the fact that it may be easier to cross the other less conservative boundaries. We will comment on this again in the discussion section. However, in general, both monitoring boundaries are equally powerful in each row for scenario II.

Table 1. Error rates for Pocock, power and O'Brien-Fleming design types

α (1-sided)	BM boundary	BB boundary	BM boundary	BB boundary	BM boundary	BB boundary
	$(\lambda = 1)$	$(\lambda = 1)$	$(\lambda = 2)$	$(\lambda = 2)$	$(\lambda = 3)$	$(\lambda = 3)$
0.025	0.026 (0.008)	0.026 (0.008)	0.025 (0.007)	0.025 (0.007)	0.024 (0.007)	0.025 (0.007)
0.005	0.005 (0.004)	0.005 (0.003)	0.005 (0.003)	0.005 (0.003)	0.005 (0.004)	0.005 (0.004)

Table 2. Power for Pocock, power and O'Brien-Fleming design types

Error rates		$\alpha = 0.025$, one-sided						$\alpha = 0.005$, one-sided					
Power		80%			90%			80%			90%		
Design		$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Sample size		170			227			249			317		
BM	Scenario I	0.82	0.84	0.85	0.91	0.93	0.93	0.82	0.84	0.85	0.92	0.93	0.94
		0.018	0.016	0.018	0.013	0.012	0.013	0.016	0.018	0.015	0.012	0.012	0.012
BM	Scenario II	0.91	0.86	0.81	0.97	0.95	0.92	0.93	0.90	0.85	0.98	0.96	0.94
		0.013	0.015	0.017	0.008	0.009	0.012	0.012	0.013	0.017	0.006	0.009	0.011
BB	Scenario I	0.82	0.84	0.85	0.91	0.93	0.94	0.82	0.84	0.85	0.92	0.93	0.94
		0.018	0.016	0.018	0.013	0.012	0.012	0.016	0.018	0.015	0.013	0.012	0.012
BB	Scenario II	0.91	0.86	0.81	0.97	0.95	0.92	0.93	0.90	0.85	0.98	0.96	0.94
		0.013	0.015	0.018	0.008	0.009	0.013	0.012	0.013	0.016	0.006	0.009	0.011

5. Examples

We now apply each of the four methods to the BHAT example. The BHAT was a multi-center clinical trial, sponsored by the National Heart, Lung, and Blood Institute, and designed to test whether or not long term use of propranolol by patients who have recently suffered heart attacks reduces mortality (DeMets, Hardy, Friedman & Lan, 1984). The original study design is based on a two-sided test with type I error rate of 0.05 and 90% power for mortality reduction from 0.1746 to 0.1375 for the propranolol arm. The calculated sample size for a fixed design is 2010 patients each arm. For illustration purpose, we calculated boundary values at information times 0.206, 0.392, 0.616, 0.791 and 1. The interim test statistics are 2, 2.0494, 1.9494 and 2.8983 for the first four analyses. Figure 1 shows the monitoring boundaries under BM and BB, respectively. The two grey lines of the same type on the left panel are boundary values under BB but plotted on the Z value scale (under BM), while the two grey lines on the right panel are the boundary values under BM but plotted on the BB scale. Both stopping rules control the error rate at the design level, but since the joint distribution of the interim statistics based on the BB has different covariance matrices, the boundary shapes are different from those of the typical group sequential designs (left panel in Figure 1). This is easy to understand since the Brownian bridge is conditioned on 0 at information times 0 and 1 under the null. Please note that on the Z value scale, all the boundaries either under a BM or a BB are non-increasing, a typical required condition for these designs. The OBF design type is a more conservative rule in that it has greater boundary values during early analyses but the final critical value is less than that of Pocock design type under each model. We can see that at the fourth interim analysis, the Z value crosses both boundaries under BM and the BB. As mentioned in the methods section, after the transformation from BM to BB, we are able to arrive at stopping boundaries with similar properties as shown in the simulations and the BHAT example here. This builds a strong basis for our investigations regarding conditional power and other methods.

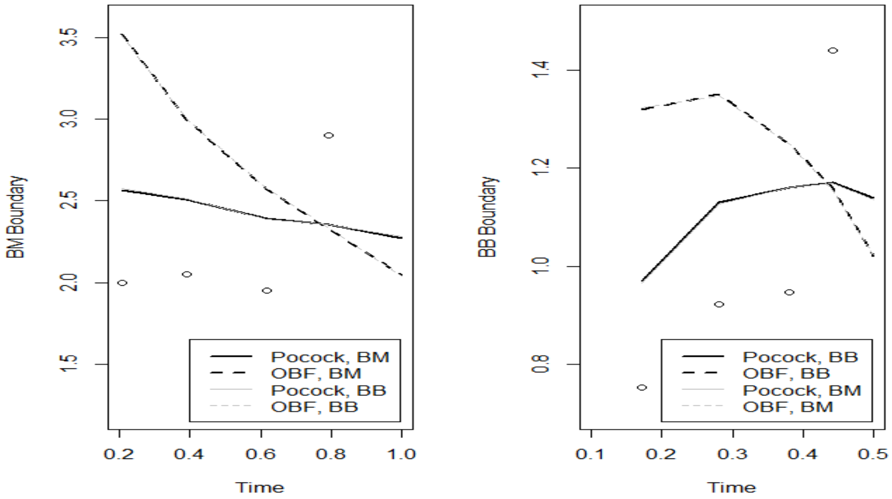

Figure 1. Left panel: Stopping boundaries under Brownian motion (dark lines) and the BHAT interim results (Grey lines are based on the Brownian bridge from figure 2 but plotted on Z value scale for comparison purposes).

Right panel: Stopping boundaries under Brownian bridge (dark lines) and the BHAT interim results (Grey lines are based on the Brownian motion from figure 1 but plotted on B value scale for comparison purposes).

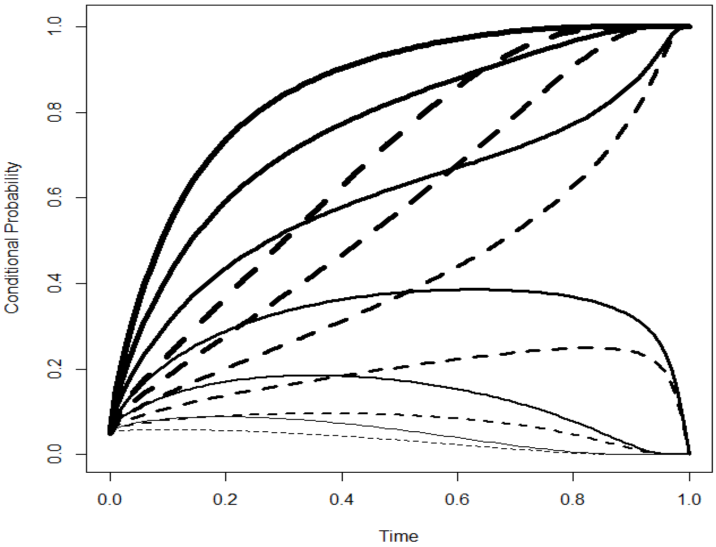

Figure 2. Conditional power based on the Brownian bridge (solid lines in increasing thickness are for Z values 0.5, 1.0, 1.5, 2.0, 2.5, 3.0) and the Brownian motion (dashed lines in increasing thickness are for Z values 0.5, 1.0, 1.5, 2.0, 2.5, 3.0) under the null hypothesis.

Now we come back to our point in the introduction. The conditional powers under the BM for the OBF design types in Figure 1 using the spending function are 0.34, 0.46, 0.54 and 0.59 for the first four analysis times. But these are 0.77, 0.81, 0.80 and 0.77 under the BB. This implies the OBF design types derived under both BM and BB meet the conditional power expectations according to BB. Also, the early treatment difference needed to reject the null hypothesis is more realistic than those using conditional power under the BM (e.g., Z values are 3.52 vs. 5.35 at $t_{BM} = 0.20$). We will discuss further regarding group sequential designs based on conditional power (Zhang et al., 2015) under the BB in the discussion. For BHAT, we can calculate the conditional probabilities of rejecting the null hypothesis given the similar positive trend of around $Z \cong 2$ for the first three analyses in addition to the fourth one. According to the methods section, under the null the conditional probabilities of a final Z value greater than 1.96 are 0.31, 0.44, 0.46, and 0.97 for each of the analyses under the BB and these are higher than those under BM, 0.12, 0.19, 0.24 and 0.91. Again, this is because of the smaller variances under BB at all four information times, especially at the fourth interim analysis (0.055 vs. 0.209). At the second and third analyses, the interim Z values are almost the same (both> 1.96), but the conditional power under BM is not close to 50% at all. However, under BB these are much higher and closer to 50%. This is a more sensitive reflection of the positive trend in the data being close to the OBF boundary.

To look at the impact of this method on futility monitoring, we revisit an example presented by Jennison and Turnbull (2000, chapter 13). This is a randomized trial comparing chemo-radiation to radiation alone for the oropharyngeal cancer. At information times 0.2, 0.41, 0.64, 0.90 and 1.0 the Z values are -1.04, -1.00, -1.21, -0.73 and -0.87. Under the hypothesis that $\frac{d\Lambda_2(t)}{d\Lambda_1(t)} \leq 0.61$, the conditional powers under BB are 0.23 and 0.024 for the first two interim analyses, but these are 0.50 and 0.14 under BM. As we can see if we consider the usual 10% conditional power rule often used in oncology protocols, we would not stop the trial early yet according to BM, however there is a strong indication of futility under the BB.

To investigate the difference of conditional power between each approach, we plotted the conditional probabilities under the null and alternative hypotheses (5% error, 80% power, $\mu_{BM} = 2.49$) in Figures 2 and 3. When no treatment effect is assumed (Figure 2), the conditional power under the BB is higher than those under the BM for each of the interim Z values (0.5, 1.0, 1.5, 2.0, 2.5 and 3.0). For example, the conditional powers under the BB are 0.06, 0.17, 0.38, 0.63, 0.83 and 0.94 for these Z values at t_{BM} = 0.5, however, these are only 0.03, 0.09, 0.20, 0.37, 0.57 and 0.75 under the BM. The conditional power reaches 50% (approximately same as the conditional power of an OBF boundary) for a Z value of 2 at as early as $t_{BM} = 0.29$ under the BB, however, it is around $t_{BM} = 0.69$ under the BM. A Z value of 2.5 or 3 reaches 80% conditional power at as early as $t_{BM} = 0.45, or\ 0.25$ under the BB, however, it is $t_{BM} = 0.71\ or\ 0.55$ under the BM, a 26-30% difference in information. This is important since many studies finish accrual before 50% information time, so it would lead to savings not only on time but also on the number of patients if we observe a positive result early in a study. Additionally, if during this analysis, a Z value of 2.5 crosses an OBF boundary at only 50% conditional power under the BM, then it is more consistent and convincing that the actual conditional power is 80% under the BB. Also, with a Z value of $\leq$ 1.5, these results imply that it is unlikely we would stop the study early for efficacy ($CP < 50\%$, therefore less than the boundary of an OBF design) for both models. If the future trend is under H_1 (Figure 3), the conditional power is only higher under the BB early in the analysis ($t_{BM} < 0.3$) then quickly becomes lower if we continue to observe the same treatment effects ($Z \leq 1.5$). For example, when $t_{BM} = 0.5$, the conditional powers are 0.47, 0.67 and 0.82 under the BM, however, these are 0.32, 0.57 and 0.79 under the BB. This is an important property especially when the new treatment is unlikely to be effective. We will elaborate more on the influence of the conditional power difference on futility monitoring in the discussion section for the setting where a null or negative trend is observed. When the treatment effect is more evident ($Z = 2.5, 3.0$), the conditional power is higher under the BB than those under the BM. For example, the conditional powers under the BB are 0.79, 0.92, 0.98, and 0.997 for these Z values > 1.5 at t_{BM} = 0.5, however, these are 0.82, 0.92, 0.97 and 0.992 under the BM. This difference is greater when $t_{BM} < 0.5$. Please note that although the conditional power under a Brownian bridge is derived on $t \in \left[0, \frac{1}{2}\right]$, it is plotted on the same time scale for comparison purpose.

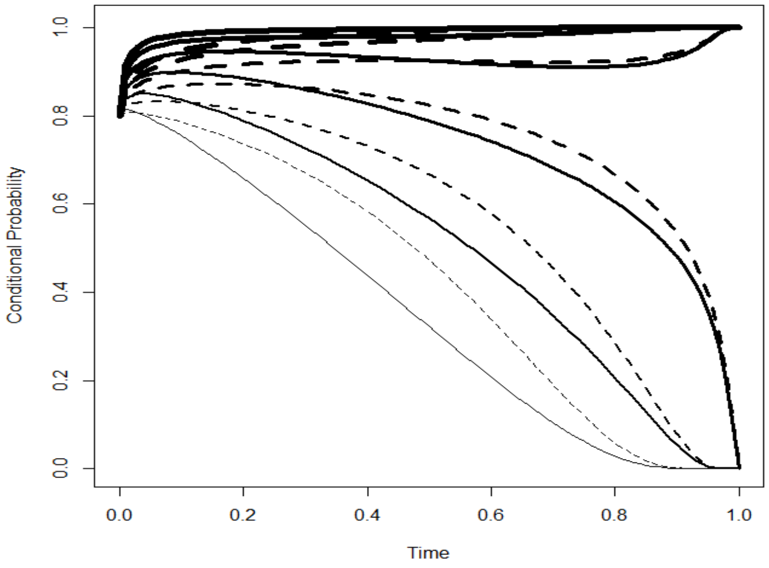

Figure 3. Conditional power based on the Brownian bridge (solid lines in increasing thickness are for Z values 0.5, 1.0, 1.5, 2.0, 2.5, 3.0) and the Brownian motion (dashed lines in increasing thickness are for Z values 0.5, 1.0, 1.5, 2.0, 2.5, 3.0) under the alternative hypothesis.

The newly introduced alpha spending function using the supremum of Brownian bridges gives boundary values of 5.52, 3.77, 2.81, 2.37 and 2.00 at each analysis time under BM. So it spends less alpha for the first two interim analyses comparing to the other three spending functions we use above. If in a study we do not expect to stop the trial at the early stage ($t_{BM} < 0.4$), then this is a good choice, otherwise, we can truncate the first two boundary values such that the stopping rule is less conservative. The last three boundary values are similar to those of the OBF and Pocock approaches. The BHAT results would suggest rejecting the null hypothesis at the fourth analysis according to this spending function as well. We will also reject the null hypothesis at the fourth interim analysis if the design is under the BB using this new alpha spending function.

According to the repeated confidence interval method, none of the first three RCIs has a lower limit greater than 1 on the hazard ratio scale favoring the experimental arm. However, at the fourth interim analysis, we would reject the null hypothesis, and the RCIs are (1.06, 1.81) and (1.07, 1.80) for the Pocock and OBF design types under BB and BM. Note that the lower limits are slightly closer to 1 for the RCIs under the Pocock desgin, and this agrees with the boundary shapes in Figure 1 at the fourth interim analysis.

6. Discussion

The Brownian bridge is one of the cornerstones in building the theory of empirical and stochastic processes. These processes are important for the robust estimation and comparison of treatment effects and for forming the theoretical framework for interim monitoring. In this paper, for the first time, we show that BB can be used for the statistical monitoring of clinical trials through the development of four kindred methods. It is very interesting that the underlying statistical models for each step converge to the same stochastic process, the Brownian bridge. We believe this unified view will bring further insights to the research of both fields. During data monitoring committee meetings, the group sequential designs, conditional power and RCI methods are usually considered together to aid the discussions. The former is necessary and critical at the design stage for setting up stopping guidelines for mostly efficacy monitoring. The other two approaches are used more often for futility monitoring. Our results show that the proposed boundaries have desired properties under each of the hypotheses. Investigators may choose group sequential designs under BB or BM if the performance is similar for the corresponding monitoring rules. However, the simple transformation from BM to BB gives rise to more powerful and sensitive or new monitoring tools, for example, the improved conditional power and the new alpha spending function.

The intriguing results of the non-sequential conditional power analyses from the two trial examples and the numerical studies indicate its potential advantages of gaining further insights from the same interim results in addition to the group sequential method under the BB and/or BM. As shown in the numerical studies, for obvious treatment difference even under the null trend, we may be able to reach high conditional power earlier (as much as 30% information for the examples we considered) than when using the Brownian motion approach. For non-significant interim results

$(e.g., Z(t) = 0, -0.5, -1, -1.5, -2, -2.5, -3)$, we may be able to stop the study for futility earlier than if we considered the conditional power based on Brownian motion. For example:

$$t_{BM} = 0.55, 0.41, 0.30, 0.22, 0.17, 0.13 \; and \; 0.11 \; when \; CP \leq 0.1$$

under H_1 for BB, but these are $0.65, 0.55, 0.46, 0.38, 0.32, 0.27 \; and \; 0.23$ for BM. In large clinical trials, especially if the patients have good outcomes, it may take a long time to observe 10%-16% more events. Thus, for either efficacy or futility monitoring, the more timely stoppings of effective or ineffective treatments according to the BB will potentially lead to less patients being exposed to the inferior therapies. Of course, stopping a trial early due to efficacy or futility will also depend on other considerations, including whether the actual timing is too early or not, secondary objectives, planned subgroup analyses, safety, quality of life, translational research, etc. Additionally, the stopping boundary for the group sequential design using stochastic curtailment and Brownian bridge (for more than one interim analysis which will be explored in future research) would be less than those under the Brownian motion, again according to which the monitoring rule is very conservative especially during early analysis times (Zhang et al., 2015). This is due to the smaller variances of Brownian bridges which lead to the higher predicted probabilities of rejecting the null hypothesis. Of course, we could also apply this approach to other analogous monitoring methods, such as Bayesian predictive power.

From the distribution of the supremum of a Brownian bridge we introduced a new spending function, this is different from existing ones in the statistical literature, and the operating characteristics of the corresponding group sequential designs and repeated confidence intervals should be explored further and its performance can be compared to the existing alpha spending functions.

We focus on survival data as an example for this research, but all tools developed here can be used for the interim monitoring of different endpoints, such as binary and continuous variables, since their asymptotic joint distributions can be transformed to Brownian bridges. We conducted a simulation study that is sufficient to confirm that the total error rates and power are satisfactory for the log-rank test under both models. More advanced theoretical or simulated results that show the sequential log-rank test follows Brownian motion have been published in the literature. Our goal is to transform this proven process and investigate the design properties instead of reproducing previously established results. Our simulation results show similar good statistical power for each model under both scenarios. For time-varying treatment effects (early difference) in scenario II we saw that the OBF design has slightly lower power than in scenario I for certain cases: one reason could be the maximum treatment difference occurred before the interim analysis, so that the regular log-rank test is not able to catch this trend which could cross the more conservative OBF boundary. The supremum weighted log-rank test is proposed for this case and has been proved to be more powerful in such settings (Eng & Kosorok, 2005, Zhang, Liu, & Kosorok, 2016). The corresponding group sequential design for time-varying treatment effects under a Brownian bridge will be explored in future research.

Overall, the proposed methods when used together as one framework are more powerful and sensitive to interim positive and negative trends that are clinically meaningful and lead to timely early stopping with potentially more savings on sample sizes, time and costs. These tools should be used widely for the interim monitoring of confirmatory trials in various disease areas.

References

Borodin, A. N., & Salminen, P. (2002). *Handbook of Brownian Motion – Facts and Formulas*. 2nd edition. Birkhäuser Verlag, Boston. https://doi.org/10.1007/978-3-0348-8163-0

Chow, W. C. (2009). Brownian bridge. *WIREs Comp Stat, 1*, 325–332. https://doi.org/10.1002/wics.38

Davis, B. R., & Hardy, R. J. (1990). Upper bounds for type I and type II error rates in conditional power calculations. *Communications in Statistics, Theory and Methods, 19*(10), 3571-3584. https://doi.org/10.1080/03610929008830398

DeMets, D. L., Hardy, R., Friedman, L. M., & Lan K.K. (1984). Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. *Contr. Clin.Trials, 5*, 362–372. https://doi.org/10.1016/S0197-2456(84)80015-X

Eng, K. H., & Kosorok, M. R. (2005). A sample size formula for the supremum log-rank statistic. *Biometrics,* 61, 86-91. https://doi.org/10.1111/j.0006-341X.2005.031206.x

Fleming, T. R., & Harrington, D. P. (1991). *Counting Processes and Survival Analysis,* New York: Wiley.

Doob, J. L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems, *Annals of Math. Stat*, *20*, 393-403. https://doi.org/10.1214/aoms/1177729991

Jennison, C., & Turnbull, B.W. (2000). *Group sequential methods with applications to clinical trials.* Chapman and Hall/CRC, Boca Raton, FL.

Kosorok, M. R. (1999). Two-sample quantile tests under general conditions. *Biometrika*, *86*, 909–921. https://doi.org/10.1093/biomet/86.4.909

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer: New York. https://doi.org/10.1007/978-0-387-74978-5

Lan, K. K. G., & DeMets, D. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, *70*, 659–63. https://doi.org/10.2307/2336502

Lan, K. K. G., & Wittes, J. (1988). The B-value: A tool for monitoring data. *Biometrics*, *44*, 579–85.

https://doi.org/10.2307/2531870

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, *35*(3), 549-556. https://doi.org/10.2307/2530245

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*, 191–199. https://doi.org/10.1093/biomet/64.2.191

Tsiatis, A. A. (1981). The asymptotic joint distribution of the efficient scores test forthe proportional hazards model calculated over time. *Biometrika*, *68*, 311–315. https://doi.org/10.1093/biomet/68.1.311

Zhang, Q. (2011). Repeated confidence intervals under fractional Brownian motion in long term clinical trials. *Communications in Statistics, Simulations and Computations*, *40*(8), 1130-1145. https://doi.org/10.1080/03610918.2011.563008

Zhang, Q., Lai, D. J., & Davis, B. R. (2015). Stochastically curtailed tests under fractional Brownian motion. *Communications in Statistics – Theory and Methods*, *44*(5), 1053-1064. https://doi.org/10.1080/03610926.2012.754469

Zhang, Q., Liu, J. F., & Kosorok, M. R. (2016). Group sequential designs for supremum weighted log-rank test for survival data with time-varying treatment effects. *To be submitted.*

Copyrights