Comments on: Statistical Considerations for a Trial of Ebola Virus Disease Therapeutics by Michael A. Proschan, Lori E. Dodd and Dionne Price

John Whitehead
Department of Mathematics and Statistics, Lancaster University, UK.

Correspondence: John Whitehead, Department of Mathematics and Statistics, Fylde College, Lancaster University, LA1 4YF, UK. Email: j.whitehead@lancaster.ac.uk.

Proschan et al. [1] describe a "barely Bayesian design" (BBD) design currently being implemented to evaluate ZMapp as treatment for Ebola. All trial patients receive an optimised standard of care, and are randomised equally between an experimental group (E) that additionally receives ZMapp and a control group (C) that does not. The primary response is binary: survival to Day 28. Interim analyses occur when 12 responses are received; then after every two new responses up to 40; finally after every 40 new responses up to 200. These are total sample sizes.

Here, the probability of success (surviving until Day 28) on Treatment T is denoted by p_T, T = E, C. The BBD uses independent beta priors with parameters (1, 1) for p_E and p_C. At each interim analysis, the posterior probability that E is superior to C is calculated as $\Pi = P(p_E > p_C)$. If $\Pi \geq 0.999$, the trial is stopped to claim that E is superior to C. If $\Pi \leq 0.001$, it is stopped to due to inferiority of E. If the maximum sample size of 200 is reached and $\Pi \geq 0.975$, then superiority will be claimed for E.

An alternative design for comparing E with C in a randomised sequential trial can be based on a triangular test (TT) [2, 3]. Up to 20 interim analyses are conducted after every 25 patient responses are received. Suppose n_T patients have received Treatment T, and S_T have succeeded, T = E, C; $n_E + n_C = n$, $S_E + S_C = S$. At each interim analysis, $Z = (n_C S_E - n_E S_C)/n$ and $V = n_E n_C S(n - S)/n^3$ are calculated. The trial is stopped to claim E is better than C if $Z \geq 6.3990 + 0.2105V$, or to conclude that E is no better than C if $Z \leq -6.3990 + 0.6315V$. This particular TT forms the randomised component of a multi-stage approach [4, 5]: here it is considered in isolation.

Table 1 presents frequentist properties of both designs from million-fold simulations. For the TT the 25 new observations at each interim analysis alternate between one extra on E and one extra on C. The TT is constructed to have a one-sided type I error rate (α) of 0.025 and a power of 0.90 when the odds ratio is 2, and these properties are achieved quite accurately. By contrast for BBD, α = 0.032 and 0.027 and powers are 0.684 and 0.574 for p_C = 0.500 and 0.667 respectively. Maximum sample sizes for BBD and TT are 200 and 500 respectively, but average sample sizes are much closer. Also given for TT is the probability that the final sample size is less than 300. When Treatments E_1, E_2 and E_3 are tested in turn against C, with $p_C = p_1 = p_2$ while p_3 takes a larger value (p_i is the probability of success on E_i), the probability of correctly recommending only E_3 is much greater for TT than BBD. In the first case, average total sample sizes are similar. Also shown are results from two "Matching TTs", achieving the same α and power as the corresponding BBD. These show substantial reductions in expected sample size relative to the BBD. The TT is recommended for implementation, **but the Matching TTs are not recommended** due to their high α and low power.

The BBD fails to make use of opportunities afforded by the Bayesian approach, in particular using pre-existing data in constructing true representations of prior opinion. Its

frequentist properties include α levels exceeding those conventionally adopted in phase III and disturbingly low power. These deficiencies are compounded in a succession of comparisons. The BBD assigns excessive resource to determining whether a treatment is merely ineffective or actually harmful. The method does include a frequentist "advisory futility rule" but, because it is operated at the discretion of an Independent Data Monitoring Committee, it cannot be evaluated in simulations.

Any of a large number of existing group sequential approaches [2, 6] could have been adopted for the trial of ZMapp, or substituted for BBD in master protocols for evaluating a series of treatments. There is no good reason for the introduction of a new method for these purposes. The BBD does not, as it stands, allow for stratification or covariate adjustment in interim analyses, use of endpoints such as survival or ordered categories rather than binary, or for final valid frequentist analyses allowing for the interim analyses and for inclusion of data collected from patients still under treatment when the stopping criterion is met. All these extensions could be developed, but they have already been created and implemented for existing methods such as the TT. The TT as described has a first interim analysis after 25 patients, whereas the BBD starts after 12. The TT could easily be altered to allow such early looks, but they are not especially useful. To stop so early for safety according to BBD, not only have the results on E to be very bad, but those on C have to be very good. By the time data from 12 patients can be analysed, more patients will have started study treatment, especially if the endpoint is at 28 days and trial recruitment is rapid. Early safety signals are most likely to be detected by the Independent Data Monitoring Committee considering outcomes of patients on E, looking at their relationship to drug administration and without necessarily waiting the full 28 days.

Either the BBD or the TT can be used with or without preceding phase II trials to screen out poor treatments quickly: the advantage of doing so increases with the number of available treatments. Such phase II trials could be randomised, but with higher type I error rates. In exceptional circumstances, as encountered in the Ebola epidemic, they might be non-randomised [4, 5].

The BBD is recommended by its authors for research in future serious epidemics. Results presented here suggest that serious consideration should be given to alternative designs in such circumstances.

References

1. **Proschan MA, Dodd LE, Price D.** Statistical considerations for a trial of Ebola virus disease therapeutics. *Clinical Trials*, 2016, DOI: 10.1177/1740774515620145.
2. **Whitehead J**. *The Design and Analysis of Sequential Clinical Trials*, revised 2nd edition. Chichester: Wiley, 1997.
3. **Whitehead J.** Group sequential trials revisited: simple implementation using SAS. *Stat Methods Med Res* 2011; **20**: 635–656.
4. **Cooper BS, Boni MF, Pan-ngum W, Day NPJ, Horby PW, Olliaro P, Lang T, White NJ, White LJ, Whitehead J.** Evaluating clinical trial designs for investigational treatments of ebola virus disease. *PLoS Med* 2015; 12: e1001815. doi:10.1371/journal.pmed.1001815.
5. **Whitehead J, Olliaro P, Lang T, and Horby P.** Trial design for evaluating novel treatments during an outbreak of an infectious disease. *Clinical Trials*, 2015.
6. **Jennison C, Turnbull BW.** *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: CRC, 2000.

Table 1: Properties of the BBD and TT for various combinations of p_C and p_E

Results derived from million-fold simulations

When $p_C = 0.500$, the Matching TT has upper boundary $Z = 4.450 + 0.2764V$ and lower boundary $Z = -4.450 + 0.8292V$,
when $p_C = 0.667$, the Matching TT has upper boundary $Z = 4.144 + 0.3163V$ and lower boundary $Z = -4.144 + 0.9489V$,
for each Matching TT up to 20 interim analyses take place after every 14 new responses

1 experimental treatment			**BBD**		**TT**			**Matching TT**		
p_C	p_E	odds ratio	Probability of recommending experimental treatment	average total sample size	Probability of recommending experimental treatment	average total sample size	probability that the total sample size ≤ 300	Probability of recommending experimental treatment	average total sample size	probability that the total sample size ≤ 168
0.500	0.333	½	0.000	180	0.000	97	1.000	0.000	59	1.000
	0.500	1	0.032	198	0.025	184	0.923	0.031	98	0.947
	0.667	2	0.684	180	0.899	227	0.810	0.686	136	0.756
	0.800	4	0.995	112	1.000	121	0.999	0.996	87	0.979
0.667	0.500	½	0.000	180	0.000	97	1.000	0.000	52	1.000
	0.667	1	0.027	199	0.025	204	0.882	0.027	91	0.967
	0.800	2	0.574	187	0.875	278	0.646	0.573	145	0.698
	0.889	4	0.973	140	1.000	157	0.989	0.986	111	0.905
3 experimental treatments			**BBD**		**TT**			**Matching TT**		
p_C, p_1, p_2	p_3	odds ratio	Probability of recommending Treatment E_3 *only*	average total sample size	Probability of recommending Treatment E_3 *only*	average total sample size	-	Probability of recommending Treatment E_3 *only*	average total sample size	-
0.500	0.667	2	0.641	576	0.855	595	-	0.644	332	-
0.667	0.800	2	0.543	585	0.832	686	-	0.542	327	-