Published in final edited form as:

Clin Trials. 2013 February ; 10(1): 81–92. doi:10.1177/1740774512469313.

Discussion of “Small-Sample Behavior of Novel Phase I Cancer Trial Designs,” by Assaf P. Oron and Peter D. Hoff

Bradley P. Carlin, Wei Zhong, and Joseph S. Koopmeiners¹

¹Bradley P. Carlin is Professor and Head and Joseph S. Koopmeiners is Assistant Professor, Division of Biostatistics, University of Minnesota, 420 Delaware St. S.E., Minneapolis, MN 55455. Wei Zhong is Biostatistician, Genentech, Inc. 1 DNA Way, South San Francisco, CA 94080

Congratulations to the authors on a fine and comprehensive paper! The authors have uncovered an important drawback of the standard continual reassessment method (CRM) and related Bayesian Phase I (BP1) trials; namely, that they can be “sticky” in their dose allocation schemes, suffering from what the authors call “long memory.” This is the point made by the n^* distributions shown in Figures 5-7, which demonstrate that while the BP1 methods typically perform best on average (a common selling point for their use), they also tend to feature disconcertingly high trial-to-trial variability. Indeed, these figures are reminiscent of decision theory settings wherein a low-variance but biased estimator may actually be preferable to an unbiased one whose variance is much higher. It is unusual to find a commonly used Bayesian method (like BP1) on the wrong side of such an argument, so the paper's findings are worthy of praise and further discussion.

We do however wish to make two points on behalf of BP1 methods which may serve to “rehabilitate” them to some extent. First, like all Bayesian approaches, BP1 methods do rely on the “good model” assumption; that is, the notion that if the underlying statistical model is far from the truth, all bets for optimality of the procedure are off. Some of the authors' illustrations (most notably the disconcerting Neuenschwander et al. example in Section 3.3) suffer from exactly this problem: laboring under an overly simple (e.g., a one-parameter logistic) dose-response model and a weakly informative prior, there is little chance that early MTD estimates will be accurate. Safety constraints added under the “modified CRM” rubric of Goodman et al. (1995) and others, such as starting at the lowest possible dose and never skipping dose levels when escalating or deescalating, can make things even tougher for BP1 methods. Practical concerns can lead to even more ad hoc fixes, such as the decision in Neuenschwander et al. to escalate from d_4 to d_7 (rather than the recommended d_{12}), which can in turn have odd consequences down the line (here, the apparently “incoherent” decision to further escalate to d_9 after seeing 2 toxicities, when in fact this is a drop from the originally recommended d_{12}). To us, the repairs actually applied to this design (enriching the model to a two-parameter logistic and increasing the toxicity penalty) all seem quite reasonable, enabling the trial to end successfully and without an overly high DLT rate. Indeed, we would argue that the ability of Bayesian methods to flexibly adapt to this challenge should be viewed as an advantage.

To summarize our first point then, all Bayesian methods (including BP1) are model-based, and can thus be expected to pay dividends *when the model is at least approximately correct*.

They also offer the chance to learn about model flaws as we go, and thence to take corrective action. Indeed, most standard Bayesian textbooks (e.g. Gelman et al., 2004; Carlin and Louis, 2009) routinely recommend extensive model checking and comparison, and so the uncritical use of simple one-parameter CRM models is indeed to be eyed warily. In the context of BP1 design, Yin and Yuan (2009) use Bayesian model averaging across multiple parallel CRM models, each with a different toxicity probability “skeleton,” thus incorporating Bayesian model choice into a Phase I trial. But even when using only the standard CRM, the use of 3-patient (not 1-patient) cohorts, better priors (including that on which starting dose to select), and the modified CRM restrictions are all simple and sensible in practice. Discovering problems with our statistical model should not in our view make us want to abandon modeling entirely, as up-and-down and related 3+3 methods do. Instead, it should make us want to build a better model, where we remember that “model” includes the dose-response curve, the error distribution, the prior distribution, and so on. Up-and-down methods utilize some of the same external information (for example, imposing inequality constraints via a separate, post-hoc isotonic regression analysis), but do so in an ad hoc, one-step-at-a-time fashion, thus precluding accurate propagation of all sources of uncertainty throughout the analysis. The authors seem to view the fully Bayesian approach as guilty of making an “unrealistic promise” here, and seek to retreat to the safer ground of older methods useful for “dose selection only.” But surely point and interval estimation of the MTD is the primary goal of many Phase I investigators, so, armed as we are with loads of models, computational methods, expert information, published literature reviews, and yes a small sample of current binary observations, retreat seems premature.

Our reluctance to retreat from modeling brings us to our second point, which is the authors' contention that the “overarching feature” of BP1 methods is their “insistence upon treating every cohort with the estimated MTD at any given time.” While one might well get this impression from reading the existing BP1 literature (or the summary in Chapter 3 of the recent textbook by Berry et al., 2011), it is not our opinion that this feature is crucial to a principled BP1 modeling approach. Indeed, the main message we take from the authors' work is that, while it may be perfectly sensible to use the Likelihood Principle (LP) for *estimation* (as when using the posterior mean as our final MTD estimate), making it do double duty as a *design* principle (as BP1 methods do when using the current posterior to choose the next dose) may not be: the authors show this can lead to designs with long memory properties that biostatisticians and clinicians may well find undesirable.

To remedy this, we might simply *sample* the next dose from the MTD posterior distribution, instead of just using its mean. This would undoubtedly reduce “settling” in the MTD trajectories as well as lighten the tails of the trial-to-trial n^* distributions, but the corresponding dosing schedules would likely be too erratic to satisfy many clinicians. As a less drastic alternative, we might limit the memory of our BP1 procedure by utilizing only the data from the m most recent cohorts, Y_m^* , when determining the next dose. Reminiscent of temporally adaptive escalation with overdose control (EWOC) approaches, we might allow m to increase as the trial wears on, with m approaching the total number of cohorts n as our final MTD estimate becomes sufficiently precise. In our view this does not violate the

LP, since we only limit the Bayesian procedure's memory at the interim allocation stage, not the final estimation stage.

We offer a brief investigation of this approach in the context of a standard CRM using a one-parameter logistic dose-response curve,

$$p_j = \frac{\exp(\alpha + \beta X_j)}{1 + \exp(\alpha + \beta X_j)}, \quad (0.1)$$

where j indexes the ℓ possible dose levels $X_j, j = 1, \dots, \ell$. We fix α at -3 , and assume β follows an exponential distribution with mean 1. We also investigate an alternative, nonparametric model that avoids any smooth baseline link entirely, and instead assumes only that $0 < p_1 < p_2 < \dots < p_\ell < 1$. To implement this nonparametric isotonic regression in a fully Bayesian fashion, a simple solution is to add these constraints to an otherwise vague prior. For example, taking $\ell = 6$ we can set $p_1 \sim \text{Unif}(0, p_2)$, $p_2 \sim \text{Unif}(p_1, p_3)$, $p_3 \sim \text{Unif}(p_2, p_4)$, $p_4 \sim \text{Unif}(p_3, p_5)$, $p_5 \sim \text{Unif}(p_4, p_6)$, and $p_6 \sim \text{Unif}(p_5, 1)$, where $\text{Unif}(a, b)$ denotes the uniform distribution on the interval (a, b) . Such a nonparametric uniform prior (NUP) model is readily implemented in BUGS or JAGS, and should perform well when the number of dose levels is not too large.

After each look at the data, we update the memory-limited posterior means $E[p_j | Y_m^*]$, and then select as our next dose the one that minimizes the distance

$$\text{dist}_j = |E[p_j | Y_m^*] - p^*|, \quad (0.2)$$

where p^* is our physician-specified target rate of toxicity. In this simple illustration we do not permit early stopping, simply running all trials to the maximum sample size. Our simulation assumes the patient recruitment plan is to enroll a new cohort of size $k = 3$ every 6 weeks, with all toxicity outcome data from each cohort assumed available prior to the enrollment of the next cohort. For this illustration we take $p^* = 0.36$, $m = 5$, and set the maximum number of cohorts $n = 16$. Here we compare our logistic and NUP models as implemented using standard CRM rules (where the full interim dataset Y is used for dose allocation) to our “limited memory” version (where only Y_m^* is used for this purpose).

We present simulation results in the case where the optimal dose is Dose 4; entries for this dose are highlighted in boldface in the tables for easier identification. We used R2jags (cran.r-project.org/web/packages/R2jags) to call JAGS (mcmc-jags.sourceforge.net) from R version 2.12.2. Each of our simulation studies used 1000 simulated trials, analyzed by generating two MCMC chains run for 5000 iterations following a 5000-iteration burnin period. Random manual checks revealed no significant MCMC convergence issues using standard convergence diagnostics.

Table 1 shows the true parameter settings in our first scenario, for which the simple one-parameter logistic function can describe the true probability of toxicity reasonably well. Figure 1 provides histograms of trial-to-trial variability in n^* similar to those in the authors'

Figure 5. The first row shows the empirical distributions of n^* for the standard, long-memory (LM) CRM, while the second row shows the short memory (SM) results. The first column adopts the logistic model (which is the truth here), while the second utilizes the more flexible NUP model. The results confirm that our short memory approach does indeed reduce the undesirable heavy tail behavior in the n^* distribution. However, these gains do come at the price of a small reduction (approximately 1 cohort) in $\overline{n^*}$, marked as in the authors' figures by vertical lines. Table 2 investigates this tradeoff a bit further by showing the empirical selection probabilities and percents of patients treated at each dose using the competing parametric and nonparametric link functions in both the LM and SM cases. This table also confirms that our memory-limited CRM continues to perform nearly identically to the standard CRM with respect to these criteria.

Table 3 describes a second, unsmooth scenario, in which no logistic function can readily approximate the true probability of toxicity due to the large jump between Doses 3 and 4. Histograms showing trial-to-trial variability in n^* for this scenario appear in Figure 2, while the corresponding selection probabilities and proportions of patients treated at each dose appear in Table 4. The results indicate the NUP models performed better than the logistic models in this unsmooth scenario. Still, the results confirm that our simple memory-limiting device leads to dramatic improvements in n^* performance while leaving selection probabilities and the proportions of patients treated virtually unchanged. Particularly striking is the second column of Figure 2, where the SM BP1 method completely eliminates the bimodality of the LM version while reducing $\overline{n^*}$ by just 0.5 cohorts.

We note that the potential range of memory-limited BP1 modeling is much broader than investigated here. Potential enhancements include simultaneous models for toxicity and efficacy (or surrogate efficacy), range parameters to capture lower or upper bounds on response probabilities, differential weighting schemes to place a higher penalty on overdosing than underdosing, early termination rules to control overtotoxicity and to enable an early decision regarding the optimal dosage, and so on. See Zhong et al. (2012, 2013) for a discussion of some of these issues in the standard BP1 context.

In closing, we again congratulate the authors on their work, and look forward to future developments in this surprisingly resilient research area. In particular, we concur with the authors that hybrid methods which attempt to capture the best features of both up-and-down and BP1 approaches may well offer the more sensible path forward.

Acknowledgments

The work of the first two authors was supported in part by NCI grant R01-CA095955.

Additional References

- Berry, SM.; Carlin, BP.; Lee, JJ.; Müller, P. Bayesian Adaptive Methods for Clinical Trials. Boca Raton, FL: Chapman and Hall/CRC Press; 2011.
- Carlin, BP.; Louis, TA. Bayesian Methods for Data Analysis. 3rd. Boca Raton, FL: Chapman and Hall/CRC Press; 2009.
- Gelman, A.; Carlin, J.; Stern, H.; Rubin, DB. Bayesian Data Analysis. 2nd. Boca Raton, FL: Chapman and Hall/CRC Press; 2004.

- Yin G, Yuan Y. Bayesian model averaging continual reassessment method in phase I clinical trials. *Journal of the American Statistical Association*. 2009; 104:954–968.
- Zhong W, Carlin BP, Koopmeiners JS. Flexible link continual reassessment methods for trivariate binary outcome phase I/II trials. *Journal of Statistical Theory and Practice*. 2013 To appear.
- Zhong W, Koopmeiners JS, Carlin BP. A trivariate continual reassessment method for phase I/II trials of toxicity, efficacy, and surrogate efficacy. *Statistics in Medicine*. 2012 To appear.

Figure 1.

Between-run variability in n^* , the number of cohorts allocated to the MTD, for the long- and short-memory versions of the logistic and NUP models in the logistic scenario. The runs were 16 cohorts long with cohort size 3.

Figure 2.

Between-run variability in n^* , the number of cohorts allocated to the MTD, for the long- and short-memory versions of the logistic and NUP models in the unsmooth scenario. The runs were 16 cohorts long with cohort size 3.

Table 1

Simulation parameter settings when Dose 4 is optimal and a simple logistic function can readily fit the true probability of toxicity.

Dose	1	2	3	4*	5	6
True p	0.07	0.13	0.23	0.36	0.5	0.61
True distance	0.29	0.23	0.13	0	0.14	0.25

Table 2

Operating characteristics of the logistic and NUP models under long memory (LM) and short memory (SM), when Dose 4 is optimal and a simple logistic function can readily fit the true probability of toxicity.

Model (memory)	Operating characteristics	Dose					
		1	2	3	4*	5	6
Logistic (LM)	selection probability	0	0.005	0.206	0.627	0.152	0.01
	prop. of patients treated	0.066	0.095	0.245	0.387	0.162	0.045
Logistic (SM)	selection probability	0	0.002	0.170	0.632	0.189	0.007
	prop. of patients treated	0.067	0.097	0.249	0.332	0.187	0.068
NUP (LM)	selection probability	0	0.004	0.403	0.585	0.008	0
	prop. of patients treated	0.063	0.096	0.471	0.363	0.007	0
NUP (SM)	selection probability	0	0.002	0.366	0.618	0.014	0
	prop. of patients treated	0.063	0.106	0.530	0.297	0.004	0

Simulation parameter settings when Dose 4 is optimal and an unsmooth scenario where a parametric function cannot readily fit the true probability of toxicity.

Table 3

Dose	1	2	3	4*	5	6
True p_j	0.1	0.12	0.2	0.55	0.6	0.62
True distance	0.45	0.43	0.35	0	0.05	0.07

Operating characteristics of the logistic and NUP models under long memory (LM) and short memory (SM), when Dose 4 is optimal and a parametric function cannot readily fit the true probability of toxicity.

Table 4

Model (memory)	Operating characteristics	Dose					
		1	2	3	4*	5	6
Logistic (LM)	selection probability	0	0	0.010	0.465	0.330	0.195
	prop. of patients treated	0.063	0.064	0.097	0.364	0.245	0.167
Logistic (SM)	selection probability	0	0	0.020	0.457	0.354	0.169
	prop. of patients treated	0.063	0.063	0.135	0.335	0.240	0.164
NUP (LM)	selection probability	0	0	0.010	0.719	0.266	0.005
	prop. of patients treated	0.062	0.063	0.086	0.570	0.215	0.004
NUP (SM)	selection probability	0	0	0.005	0.723	0.266	0.006
	prop. of patients treated	0.062	0.063	0.135	0.540	0.196	0.004