Supplement to "Building the Bridge to Phase II: Efficacy Estimation in Dose-Expansion Cohorts" by Boonstra, et al.

- Section S1 describes the actual simulation process.
- Sections S2 and S3 present additional details on the dose-assignment and efficacy analyses, respectively.
- Section S4 presents auxiliary results for Scenarios 1–6 in the manuscript
- Section S5 presents additional results for an additional set of scenarios (Scenarios 7–10).

S1 Description of Simulation

The following steps outline how the simulation proceeds. The scenarios are visually presented in Figure 2 of the manuscript.

1. **Specify generating parameters** Select true toxicity curve and efficacy curves A and B, i.e. one of the scenarios in Figure 1 (manuscript) or Figure S2 (supplement). For all patients in the trial, toxicity and efficacy outcomes are simulated using the true, unknown probabilities from these curves corresponding to the assigned dose level.
2. **Select DEC configuration** Choices considered here are: $n = 15$ patients per DEC, $n = 30$ patients per DEC with no interim analysis, or $n = 30$ patients per DEC with an interim analysis
3. **Select dose assignment mechanism** Either Local, Global, or CRM (Section 2.1 [manuscript], with additional details in Section S2 [supplement])
4. **Select efficacy analysis** Either Empiric or Model-based (Section 2.2 [manuscript], with additional details in Section S3 [supplement])
5. **Conduct 2500 simulated trials** Repeat Steps 5a–5e 2500 times.
    - **5a. Simulate dose escalation** If the dose assignment mechanism is Local or Global, dose escalation is based upon the 3+3 algorithm. If the dose assignment mechanism is CRM, then this step enrolls a number of patients equal to the average number that the 3+3 algorithm enrolls, but making dose assignments according to the underlying statistical model used by the CRM.
    - **5b. Simulate dose expansion up to 75 patients (5 cohorts of up to 15 patients per cohort)** Simulate the first patient, assigned to the estimated MTD from Step 5a. The assigned dose level for subsequent patients will be made using the selected dose assignment mechanism applied to the simulated toxicity outcomes from the previous patients. DEC membership to one of the five DECs (labeled DEC 1, 2, 3, 4, and 5) is randomly simulated as each patient enrolls. For Local, the trajectory of dose assignments may differ between the five DECs; if dose reductions proceed below dose level 1 for any DEC, then no more patients are enrolled to that DEC. For Global or CRM, the dose assignment trajectory ignores DEC membership; if dose reductions proceed below dose level 1, then the entire trial is stopped. Any DEC that enrolls 15 patients proceeds to Step 5c.
    - **5c. Conduct initial efficacy analysis** If the DEC configuration is '$n = 30$ patients per DEC with no interim analysis', skip this step and proceed to Step 5d. Otherwise, for each of the DECs that completed Step 5b, simulate the 15 efficacy outcomes (response or no response). Without loss of generality, we generate efficacy outcomes for patients in DECs 1–3 according to efficacy curve A (the inefficacious DECs) and efficacy outcomes for DECs 4–5 according to efficacy curve B (the efficacious DECs). Then, separately for each DEC, conduct an efficacy analysis to determine
        - (in the case of '$n = 15$ patients per DEC') whether to recommend the current assigned dose level for this DEC or conclude that the current assigned dose level is not efficacious in this DEC (skipping Steps 5d and 5e in both cases)

- (in the case of '$n = 30$ patients per DEC with an interim analysis') whether to continue to Step 5d for this DEC or stop enrolling patients to this DEC

5d. Simulate dose expansion up to 75 more patients (up to 5 cohorts of up to 15 patients per cohort) Enroll additional patients for any DEC that successfully completed Steps 5b and 5c. The assigned dose level will continue to be potentially modified according to the selected dose assignment mechanism applied to all the simulated toxicity outcomes from the previous patients.

5e. Conduct second efficacy analysis For each of the DECs that completed Step 5d (meaning that a total of 30 patients were simulated), simulate the remaining efficacy outcomes and conduct a second efficacy analysis on all 30 patients to determine whether to recommend the current assigned dose level for this DEC or conclude that the current assigned dose level is not efficacious in this DEC.

6. Report Outcomes For each combination of DEC configuration, dose-assignment mechanism, and efficacy analysis, the trial outcome is recorded (whether a dose was recommended; which dose was recommended; and the reason for recommending–or not–a dose). The ideal outcome is that a dose is never recommended for DECs 1–3, because these are always the inefficacious DECs, and that a dose is always recommended for DECs 4–5, because these are the efficacious DECs. How frequently this occurs is captured by the metrics FPR and TPR. A crude estimate of the simulation variability of TPR and FPR can be calculated as follows: the binomial variance is maximized at 0.5, i.e. $0.5 \times (1 - 0.5)$. Thus, 2500 simulations with 2 efficacious DECs per simulation suggests an approximate upper bound on the variability due to simulation of $0.5 \times (1 - 0.5)/(2500 \times 2) = 5 \times 10^{-5}$, and so the reported values of TPR in Table 1 of the manuscript and Table S2 of the supplement are likely to be within 2 standard errors of the true values, i.e. within $\pm 2 \times \sqrt{5 \times 10^{-5}} \approx \pm 1.4\%$. The variability for FPR will be slightly smaller, because there are three inefficacious DECs per simulation: reported values of FPR in Table 1 are likely to be within $\pm 1.2\%$.

We note that the actual code for implementing the simulation differs slightly from the above description, for purposes of computational expediency. This code is available at `http://www.umich.edu/~philb`. We used the R statistical environment for the simulation and all analyses (1–4).

S2 Dose-Assignment Mechanisms, Additional Details

The Local and Global mechanisms start with the 3+3 algorithm for dose-escalation, which is implicitly targeting a rate of DLT between 1/6 and 2/6, as observed by Storer (5). The CRM mechanism extends the continual reassessment method (4, 6), which targets a specific rate of DLT. In order to compare all three mechanisms, in the CRM mechanism, we target a rate of DLT equal to 0.25, which exactly falls between the 1/6 and 2/6 rates from the 3+3 algorithm.

Local Dose escalation is according to the common 3+3 design (7). Following its completion, five DECs open, all starting at the same estimated MTD. An extension of the 3+3 design monitors for toxicity during the DEC: at any point after 3 patients have been enrolled to the current dose level, if the proportion of patients within that DEC experiencing a DLT ever exceeds 1/3, the dose level is reduced by one level for the next patient, if possible. For example,

- 2 or more DLTs among the first 3 patients in a DEC will result in de-escalation before enrolling the 4th patient;
- 2 or more DLTs among the first 4 patients in a DEC will result in de-escalation before enrolling the 5th patient;
- 2 or more DLTs among the first 5 patients in a DEC will result in de-escalation before enrolling the 6th patient;
- 3 or more DLTs among the first 6 patients in a DEC will result in de-escalation before enrolling the 7th patient;
- etc...

In contrast to the approach of the 3+3 algorithm, we deliberately and specifically defined the Local rule not to de-escalate if the observed DLT rate at a dose level is *exactly equal* to 1/3, e.g. 2 DLTs in 6 patients, as this could still plausibly occur when the true toxicity rate at that dose level is $p = 0.25$ (i.e. the targeted rate). Multiple dose de-escalations during a DEC are possible but escalation is not. If the Local threshold of exceeding an observed 1/3 toxicity rate is met when a dose assignment within a DEC is already at the lowest dose level, enrollment–at that DEC only–is stopped, i.e. toxicity monitoring is local to that DEC.

Global The Global mechanism differs from the Local by combining all DECs for purposes of toxicity monitoring. This is achieved using a Pocock-type toxicity boundary to trigger dose de-escalation (8). Specifically, the trigger monitors toxicity at the kth patient treated at a single dose level, with $k = 3, \ldots, 5n$ and 5 DECs $\times$ n patients/DEC. The observed number of DLTs is given by $X = x$, where X is binomial: $\text{Bin}(k, 0.25)$. Whenever $\Pr(X \leq x | k, p = 0.25) > \gamma$, with $\gamma = 0.99815$ when $n = 15$ and $\gamma = 0.9986$ when $n = 30$, the dose level is de-escalated for all DECs, if possible, or, if already at the lowest dose level, patient enrollment in all DECs stops. In words, assuming that the true DLT rate is 0.25, if the probability of having observed the current total number of DLTs is extraordinarily high, then the true DLT rate is most likely larger than 0.25, and de-escalation is warranted. As with the Local design, the rule may trigger further de-escalation if this lower dose also proves toxic. When the true toxicity rate is the targeted value of $p = 0.25$, the respective choices of γ yield a cumulative probability of a single (unwarranted) de-escalation over all $5n = 75$ or $5n = 150$ patients approximately equal to 0.05.

CRM This is a modified implementation of CRM, spanning all DECs, with total sample size equal to the average sample size that a 3+3 algorithm at that dose-toxicity curve would enroll, as in (9), plus $5n = 75$ or $5n = 150$. From a toxicity perspective, we considered a similar design in our early work (10). The dose assignment mechanism is kept open for all patients, assuming a common dose-toxicity model across all DECs. Thus, from a toxicity perspective, there is no distinction between dose escalation and expansion: assignments are made according to the same mechanism for the entire trial and across all DECs. The statistical model used is given by

$$\Pr(\text{DLT at dose } i) = s_i^{\exp\{\beta\}},$$

where $s_i = \{0.05, 0.15, 0.25, 0.35, 0.45\}$ is the "skeleton" (this is the toxicity curve corresponding to scenario 1). Each patient is assigned to the dose level that the above model estimates has a rate of DLT closest to 0.25 but not exceeding 0.25+0.05=0.30. The model is fit with a Bayesian analysis, a priori $\beta \sim N(0, 0.6^2)$. Thus, dose level 3 is a priori thought to be the true MTD. If the CRM-estimated rate of DLT at the lowest dose level ever exceeds 0.30 (i.e. the target rate of 0.25 plus a margin of 0.05), suggesting that all dose levels are highly toxic, enrollment in all DECs stops. All other settings are as described in the Supplement of Boonstra, et al. (10), including some modifications to the CRM to improve patient safety (11, 12).

S3 Efficacy Analyses, Additional Details

Empiric A two-stage phase II design requires a minimum number of responses at each analysis in order to proceed (13, 14), based on the probability of observing the data under a specified alternative, i.e. the efficacy target. However, it is possible that, due to dose de-escalation during the MTD, fewer than 15 patients are treated at the current estimated MTD, which may change over the course of the DEC, being defined as the dose level that to be assigned to the next patient. To accommodate this, we develop a decision rule based on confidence intervals (CIs) for the estimated response rate. After 15 patients are enrolled, the upper confidence limit (UCL) of a one-sided 80% score-based CI must exceed the flat efficacy rate (0.05 for scenarios 1, 2, 5 and 6 and 0.20 for scenarios 3 and 4). For $0/15$, $1/15$, and $2/15$ responses, this upper bound is 0.045, 0.143, and 0.224, respectively. When all 15 patients are treated at the same dose, this confidence limit reduces to a more intuitive decision rule requiring at least one response when the flat rate is 0.05 and two when it is 0.20. When only a few patients are treated at MTD, the UCL may exceed the threshold solely due to the small sample size. For example, the UCL is still 0.066 even for $0/10$ responses. Thus, we impose an overriding constraint that at least one response at the current dose level is required to continue study, regardless of the number of evaluable patients. After 30 patients, we require greater evidence to recommend a dose: the lower confidence limit (LCL) of a one-sided 80% score-based CI must exceed the flat response rate. When all 30 patients are treated at the MTD, this is equivalent to a minimum response rate of $3/30$ for a threshold of 0.05 and $8/30$ for a threshold of 0.20. Table S1 translates the CI-based decision rule for the Empiric analysis into a required minimum number of responders for all possible numbers of patients treated at the current MTD.

Table S1: Translation of the Empiric analysis' decision rule based on confidence interval for response rate into a minimum required number of responders, which depends on the number of patients enrolled at the current estimated MTD. An interim futility analysis is conducted after the 15 patients are enrolled in a DEC, and a second analysis is conducted after all 30 patients are enrolled in the DEC

# Patients in DEC at Current MTD	Null Efficacy Rate $= 0.05$	Null Efficacy Rate $= 0.20$
	Required minimum # of responders to proceed after first 15 patients enrolled in DEC	
1 –10	1	1
11–15	1	2
	Required minimum # of responders to proceed after 30 patients enrolled in DEC	
1–2	1	1
3–5	1	2
6–8	1	3
9	2	3
10–13	2	4
14–17	2	5
18–22	2	6
23–26	3	7
27–30	3	8

Model In the Empiric analysis, when the dose level is de-escalated due to toxicity after patient 14, the trial will continue or not on the solely on basis of the 15th patient's outcome, owing to the minimum required response rate. Although an extreme situation, this highlights that an approach like the Empiric method summarily ignores data from patients at other dose levels. In contrast, the Model-based analysis leverages that data in estimating the response rate at the current estimated MTD. We model the dose-efficacy curve as $p(D) \equiv \Pr(\text{Response at dose } D) = \text{logit}^{-1}(\alpha + \beta D)$, where D is an integer coding for each dose level and $\text{logit}^{-1}(x) \equiv e^x/(1+e^x)$. We fit the model in R with a Bayesian analysis (1, 2), assuming a priori that α is Cauchy with scale parameter 10 and β is Cauchy with scale parameter 2.5 (truncated to the set of non-negative real numbers), based on recommendations in Gelman, et al. (15). Then, similar to the Empiric analysis, we construct a one-sided 80% CI for $p(D_{15})$, where D_{15} is the dose assigned to the 15th

patient, and compare the UCL to the flat efficacy rate to determine whether to proceed. After 30 patients, the LCL from a one-sided 80% confidence interval for $p(D_{30})$, D_{30} analogously defined is compared to the flat efficacy rate. It is not possible to translate this into a specific required minimum number of responses in general, as in the Empiric analysis, because the confidence interval depends not only on the response rate at the current MTD but also on those above and below it.

S4 Scenarios 1–6: Auxiliary Results

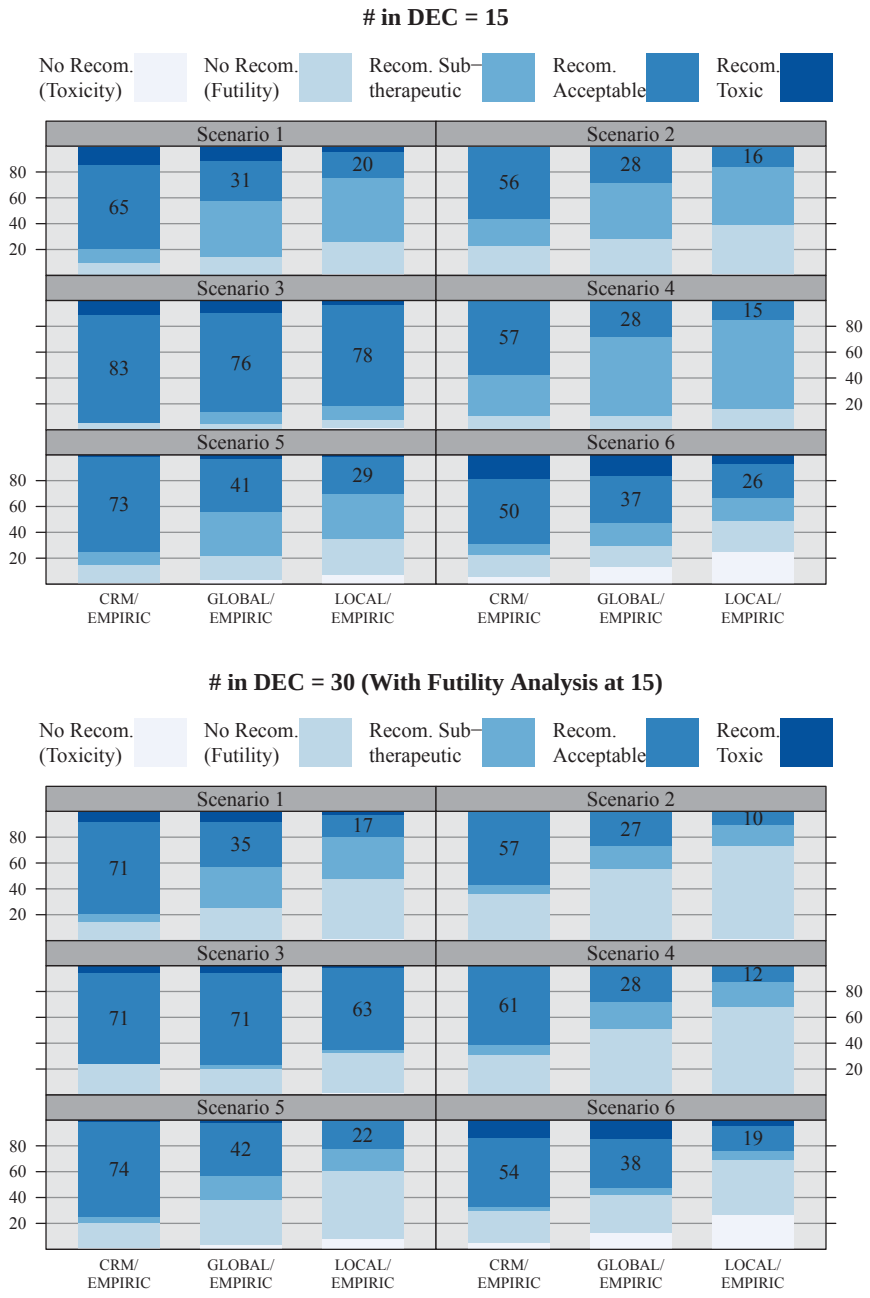

Figure S1: Breakdown of simulation-based frequencies of five possible outcomes in an efficacious DEC for the six scenarios of toxicity/efficacy curves in Figure 2 in the manuscript under each dose-assignment mechanism and for 15-patient DECs (top plot) and 30-patient DECs with a futility analysis (bottom). Only the Empiric efficacy analyses are given (the Model-based efficacy analyses are in Figure 3 of the manuscript). The ideal outcome is to recommend an acceptable, i.e. tolerable and efficacious, dose level; the proportion of simulated DECs recommending such a dose level is annotated.

S5 Scenarios 7–10: All Results

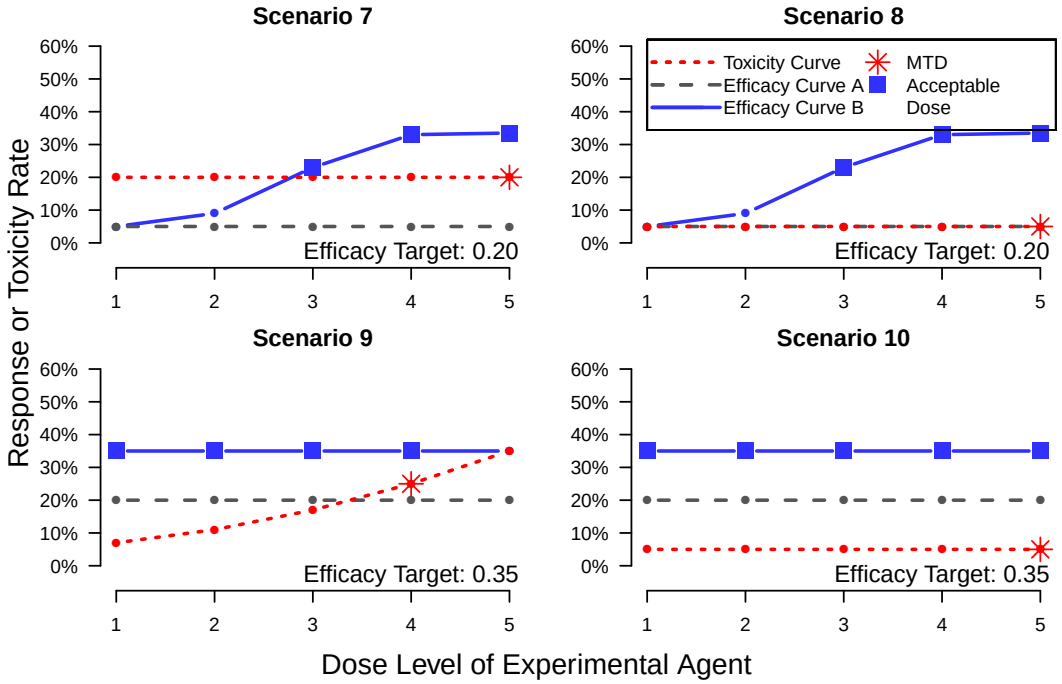

Figure S2: Four additional toxicity-efficacy scenarios. The desired targeted response, or efficacy, rate is in the lower-right of each panel. After dose-escalation, five DECs are simulated according to a common toxicity curve (short dashed). The MTD is indicated with a star. For three DECs, the targeted response rate is not achievable (long dashed; efficacy curve A); for two, it is achievable (solid; efficacy curve B). Acceptable dose levels in efficacy curve B, i.e. those with efficacy greater than the stated target and DLT rate no greater than 0.30, are boxed.

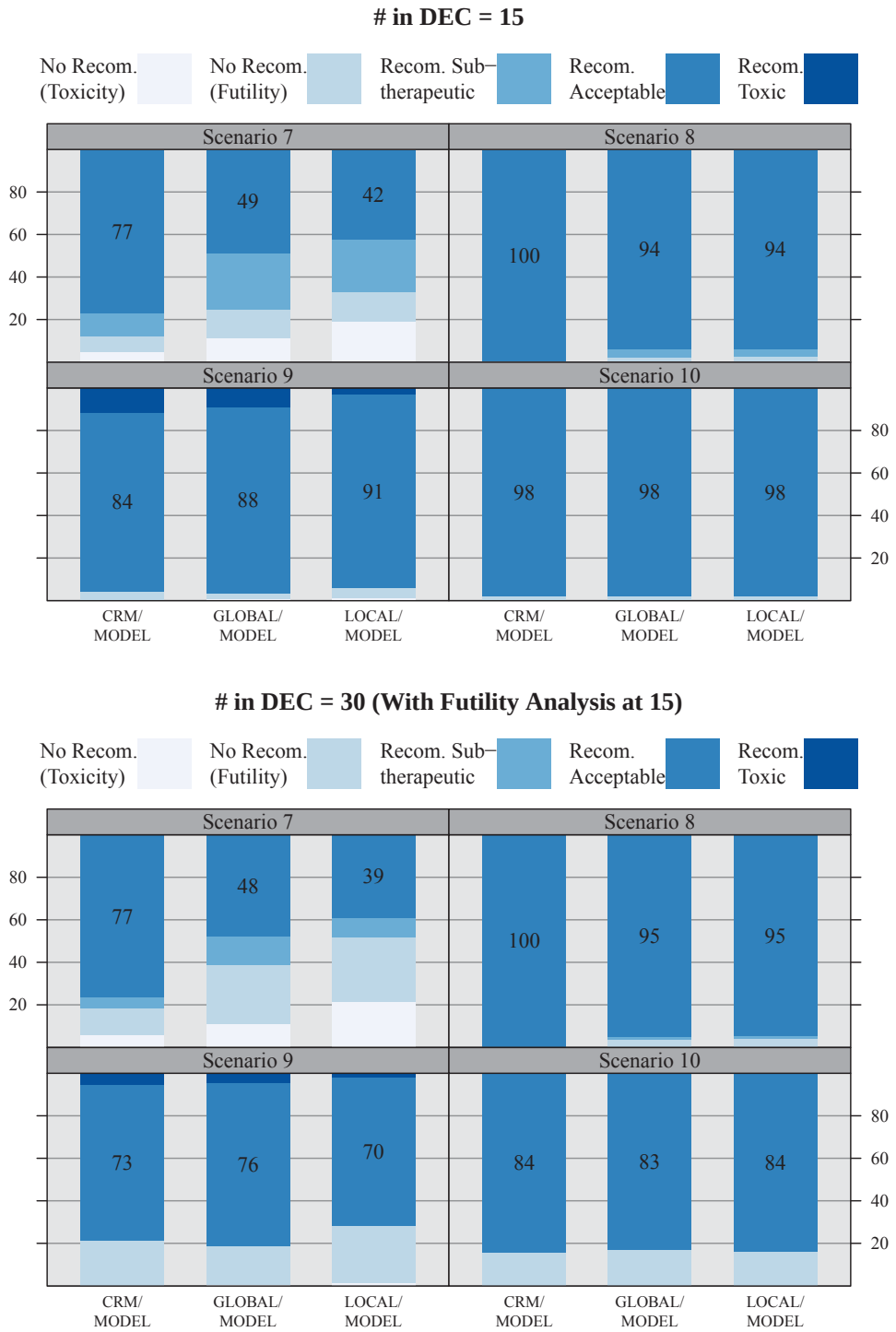

Figure S3: Breakdown of simulation-based frequencies of five possible outcomes in an efficacious DEC for the four additional scenarios of toxicity/efficacy curves in Figure S2 under each dose-assignment mechanism and for 15-patient DECs (top plot) and 30-patient DECs with a futility analysis (bottom). Only the Model-based efficacy analyses are given (the Empiric efficacy analyses are in Figure S4). The ideal outcome is to recommend an acceptable, i.e. tolerable and efficacious, dose level; the proportion of simulated DECs recommending such a dose level is annotated.

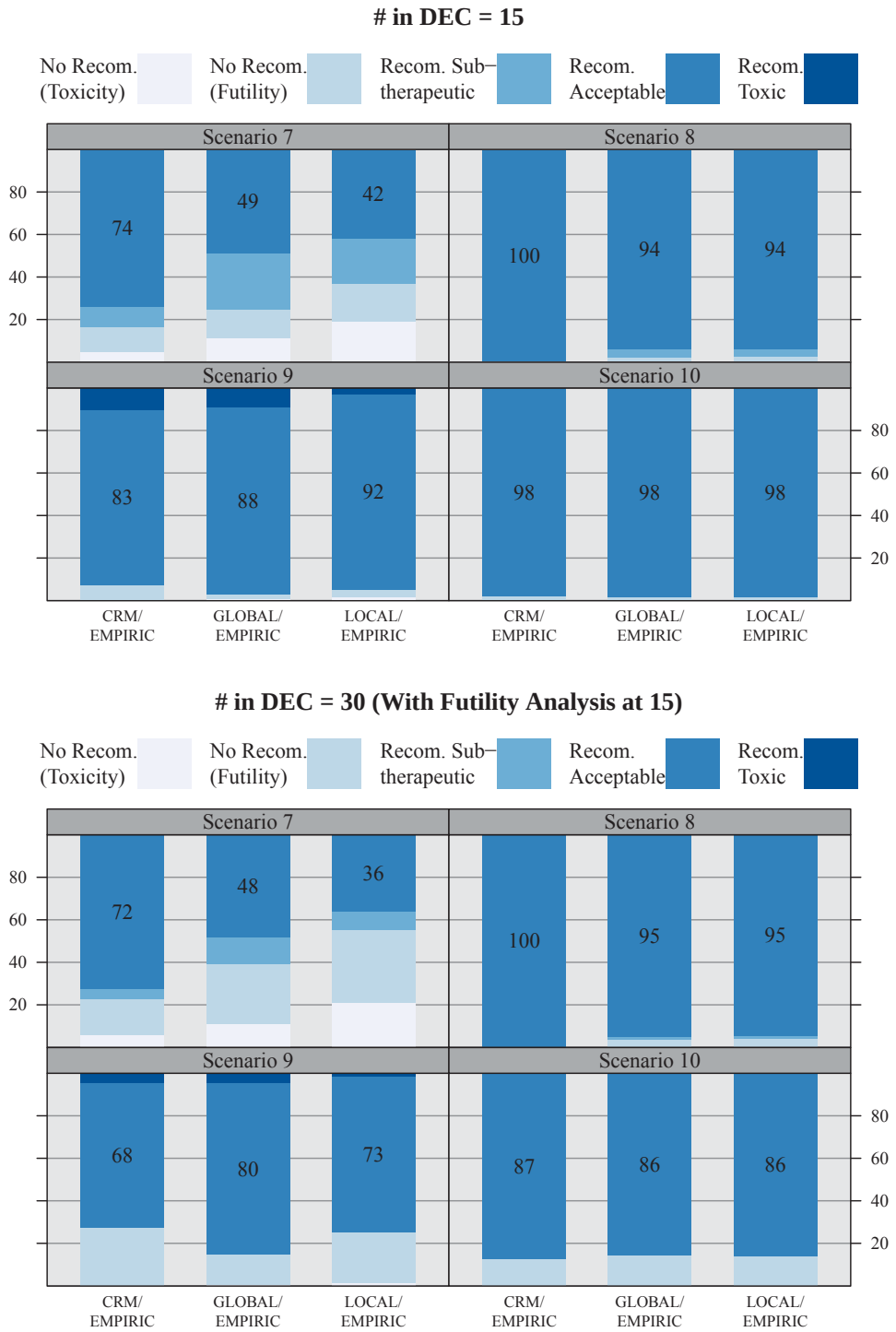

Figure S4: Breakdown of simulation-based frequencies of five possible outcomes in an efficacious DEC for the four additional scenarios of toxicity/efficacy curves in Figure S2 under each dose-assignment mechanism and for 15-patient DECs (top plot) and 30-patient DECs with a futility analysis (bottom). Only the Empiric efficacy analyses are given (the Model-based efficacy analyses are in Figure S3). The ideal outcome is to recommend an acceptable, i.e. tolerable and efficacious, dose level; the proportion of simulated DECs recommending such a dose level is annotated.

Table S2: True and false positive rates (TPR, FPR) for the four additional scenarios of toxicity/efficacy curves in Figure S2 under 18 combinations of dose-assignment mechanism, efficacy analysis, and 15-patient DECs, 30-patient DECs with no futility analysis, or 30-patient DECs with a futility analysis. TPR, in the left-hand columns, is the probability that a dose was recommended in an efficacious DEC, i.e. a DEC having at least one dose level with sufficiently large response rate. FPR, in the right-hand columns, is the probability that a dose was recommended in an inefficacious DEC, i.e. a DEC having no dose levels with sufficiently large response rate.

	TPR						FPR					
# in DEC = 15												
	CRM	Glob.	Loc.	CRM	Glob.	Loc.	CRM	Glob.	Loc.	CRM	Glob.	Loc.
	Scenario 7			Scenario 8			Scenario 7			Scenario 8		
Model	88	75	67	100	98	97	52	48	42	55	53	54
Empiric	84	75	63	100	98	97	44	48	39	54	53	53
	Scenario 9			Scenario 10			Scenario 9			Scenario 10		
Model	96	97	94	98	98	98	74	80	71	82	82	81
Empiric	93	97	95	98	98	98	75	82	78	83	84	84
# in DEC = 30 (No Futility Analysis at 15)												
	CRM	Glob.	Loc.	CRM	Glob.	Loc.	CRM	Glob.	Loc.	CRM	Glob.	Loc.
	Scenario 7			Scenario 8			Scenario 7			Scenario 8		
Model	84	63	53	100	96	96	19	18	12	18	19	18
Empiric	83	63	53	100	96	96	17	18	14	18	19	18
	Scenario 9			Scenario 10			Scenario 9			Scenario 10		
Model	80	81	77	84	85	83	20	20	16	20	20	20
Empiric	76	85	83	87	88	87	20	24	23	24	24	25
# in DEC = 30 (With Futility Analysis at 15)												
	CRM	Glob.	Loc.	CRM	Glob.	Loc.	CRM	Glob.	Loc.	CRM	Glob.	Loc.
	Scenario 7			Scenario 8			Scenario 7			Scenario 8		
Model	81	61	48	100	96	96	16	15	10	17	17	16
Empiric	77	61	45	100	96	96	13	15	11	17	18	16
	Scenario 9			Scenario 10			Scenario 9			Scenario 10		
Model	79	81	72	84	83	84	20	19	15	20	20	20
Empiric	73	85	75	87	86	86	18	23	20	23	23	24

Table S3: Average decrease in number of patients enrolled per 30-patient DEC as a result of the interim futility analysis for the four additional scenarios of toxicity/efficacy curves in Figure S2 under each dose-assignment and efficacy analysis, stratified by the efficacious DECs (left) and inefficacious DECs (right).

	Efficacious DECs						Inefficacious DECs					
	CRM	Glob.	Loc.	CRM	Glob.	Loc.	CRM	Glob.	Loc.	CRM	Glob.	Loc.
	Scenario 7			Scenario 8			Scenario 7			Scenario 8		
Model	1.2	2.6	2.3	0.0	0.3	0.3	7.2	7.1	6.4	7.0	6.9	6.9
Empiric	1.9	2.6	2.9	0.0	0.3	0.3	8.4	7.1	6.8	7.1	6.9	7.0
	Scenario 9			Scenario 10			Scenario 9			Scenario 10		
Model	0.5	0.3	0.7	0.3	0.3	0.2	3.7	2.9	4.4	2.8	2.8	2.8
Empiric	0.9	0.3	0.5	0.2	0.3	0.2	3.6	2.5	3.1	2.6	2.5	2.4

References

[1] R Core Team .*R: A Language and Environment for Statistical Computing*. R Foundation for Statistical ComputingVienna, Austria 2015.

[2] Stan Development Team .Stan: A C++ library for probability and sampling, version 2.8.0. 2015.

[3] Sarkar D.*Lattice: Multivariate Data Visualization with R*. New York: Springer 2008. ISBN 978-0-387-75968-5.

[4] Cheung K.*dfcrm: Dose-finding by the continual reassessment method* 2013. R package version 0.2-2.

[5] Hoering A, LeBlanc M, Crowley JJ. Choosing a phase I design. In: Crowley JJ, Hoering A., eds. *Handbook of Statistics in Clinical Oncology* 3rd ed. Boca Raton, FL: CRC Press; 2013:3–20.

[6] O'Quigley J, Pepe M, Fisher L. Continual reassessment method: A practical design for phase 1 clinical trials in cancer. *Biometrics.* 1990;46(1):33–48.

[7] Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM. A comparison of two phase I trial designs. *Stat Med.* 1994;13(18):1799–1806.

[8] Ivanova A, Qaqish BF, Schell MJ. Continuous toxicity monitoring in phase II trials in oncology. *Biometrics.* 2005;61(2):540–545.

[9] Ji Y, Wang SJ. Modified toxicity probability interval design: A safer and more reliable method than the 3+3 design for practical phase I trials. *J Clin Oncol.* 2013;31(14):1785–1791.

[10] Boonstra PS, Shen J, Taylor JMG, et al. A statistical evaluation of dose expansion cohorts in phase I clinical trials. *J Natl Cancer Inst.* 2015;107(3):dju429.

[11] Goodman SN, Zahurak ML, Piantadosi S. Some practical improvements in the continual reassessment method for phase I studies. *Stat Med.* 1995;14(11):1149–1161.

[12] Rosenberger WF, Haines LM. Competing designs for phase I clinical trials: A review. *Stat Med.* 2002;21(18):2757–2770.

[13] Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *J Chronic Dis.* 1961;13(4):346–353.

[14] Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clin Trials.* 1989;10(1):1–10.

[15] Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat.* 2008;2(4):1360–1383.