Supplementary Information to "A Design for Phase II Clinical Trials in Stratified Medicine with Efficacy and Toxicity Outcomes and Predictive Variables"

Kristian Brock and UoB authors[1,2]

[1]Cancer Research UK Clinical Trials Unit, Institute of Cancer and Genomic Sciences, University of Birmingham

[2]Institute of Immunology and Immunotherapy, University of Birmingham

July 29, 2017

1 Search for Feasible Trial Designs for PePS2

We sought a clinical trial design that admits explanatory variables to study joint primary outcomes efficacy and toxicity. Using PubMed, we searched for publications under the MeSH major topic 'clinical trials' that are categorised with the MeSH Terms 'Drug-Related Side Effects and Adverse Reactions' and 'Models, Statistical'. Efficacy was not made explicit in our search because establishing efficacy is such a common motivation for trials. We expected the presence of a toxicity outcome to be a more effective discriminator. On 5-Aug-2015, this query returned 67 documents whose collective focus was primarily statistical clinical trial methodology in scenarios where toxicity is a key outcome.

Forty-eight of the papers were discarded because they focused on a univariate outcome: forty-four focused primarily on toxicity alone and a further four focused on efficacy

alone. Four papers were reviews or advisory in nature and did not contain specific model proposals. One paper was discarded because it was in Danish with no English translation.

This left fourteen papers for further consideration. Naturally, given the subject matter, these papers concerned a preponderance of dose-finding and early phase trials. With cytotoxic treatments, dose-finding has typically sought to find the maximum tolerable dose under the assumption that efficacy and toxicity increase in lock step as dose is increased. In so-called cytostatic treatments, disease may be controlled without reducing the overall tumour burden and the probability of efficacy may not be an increasing function of dose. As such, in cytostatic treatments, efficacy and toxicity can be jointly scrutinised to find the optimal dose rather than just the maximal dose. The growth of targeted therapies and immunotherapies has lead to a research focus on methods that jointly model efficacy and toxicity for dose-finding purposes.

Eight of the papers in our search describe dose-finding methods for cytostatic treatments. Although these works detail designs that address a different trial objective (i.e., finding a dose), they are pertinent to our problem because they potentially use probability models that could be redeployed for our purposes. We consider those briefly now.

Braun[1] introduced a bivariate extension of the Continual Reassessment Method (CRM) to two competing outcomes, toxicity and disease progression, where the two events are associated. CRM itself was originally published by O'Quigley *et al.*[2] with the purpose of conducting dose-finding trials under the cytotoxic assumption. Ivanova[3] presented a rule-based up-and-down design that seeks to maximise the number of subjects allocated in the neighbourhood of the optimal dose. Zhang, Mandrekar and Sargent[4, 5] introduced TriCRM, another extension of CRM that considers the ordinal trinary outcome: no response and no serious toxicity; efficacy without serious toxicity; and toxicity so serious that it precludes efficacy. Wang & Day[6] present a method where response and toxicity outcomes occur according to bivariate log-normally distributed patient thresholds. They allocate the next dose to maximise patient-oriented expected utility. Finally, Thall *et al.*[7, 8, 9] present EffTox, a Bayesian adaptive design for dose-finding that con-

tinually monitors binary efficacy and toxicity events. They use the posterior probabilities of efficacy and toxicity at each dose and an L^p norm to calculate dose attractiveness scores. Subject to rules governing dose transition and eligibility, the most attractive dose is iteratively selected for successive cohorts. Generally in dose-finding models, dose (or transformed dose) is used as the sole explanatory variable that determines the outcome probabilities. This provides opportunities to use other explanatory variables in a non-dose-finding setting.

Five papers present models for efficacy and toxicity in a non-dose-finding setting. Ghebretinsae *et al.*[10] present a method for modelling non-gaussian continuous outcomes for comet assay data. This is not applicable to our scenario. In the single arm setting, Cook & Farewell [11] present a sequential design to analyse correlated bivariate efficacy and toxicity events, accounting for multiple analyses over time. Jin[12] presents a two-stage method accounting for the trade-off between efficacy and toxicity. Brutti *et al.*[13] present a two-stage Bayesian method to compare the overall toxicity rate and the true efficacy-and-safety rate to pre-specified target thresholds. None of these methods explicitly include predictive variables, although that is not to say they could not be adapted to use them.

In the two-arm setting, Bouckaert & Mouchart[14] present a model to analyse a two arm randomised controlled trial from the view that trial outcomes can be attributed to therapeutic effects and toxic effects. They also do not explicitly consider predictive variables but their model uses binary variables to denote arm membership so it is sensible to conclude that this specification could be generalised to include arbitrary explanatory variables.

Not included in our PubMed search but frequently cited in similar work is Conaway & Petroni[15, 16]. They present sequential designs for phase II trials with bivariate, dependent activity and toxicity outcomes. In each case, their emphasis is on the development of stopping rules rather than the incorporation of predictive information.

Through knowledge of the field, we are also aware of Thall, Simon & Estey[17, 18]

and Thall & Sung's[19] work on monitoring multiple outcomes (commonly, efficacy and toxicity) using Dirichlet-multinomial models and stopping boundaries in single arm phase II trials. These methods do not use predictive information. Wathen, Thall *et al.*[20] published a method that uses predictive patient data to study efficacy in patient subgroups, but their method does not study toxicity.

To further supplement our search, we studied review articles of biomarker-guided clinical trial designs. Table 2 in Buyse *et al.*[21] lists the *targeted* (or selection) design (as used in the ToGA trial[22]) and *Bayesian adaptive* design (as used in the BATTLE trial[23], amongst others) as potential designs for validated, predictive biomarkers of an experimental treatment. These are multi-arm designs, randomly allocating patients to treatments, conditional on biomarker status. Neither of these designs analyses toxicity as a co-primary outcome, although naturally safety would be an important secondary outcome in trials that use either. Freidlin & Korn[24] review randomised designs that can be used to develop or validate biomarkers. Our setting is non-randomised and concerns studying the treatment modification effect of a biomarker that has already been validated in a closely related patient population. More recently, Antoniou *et al.*[25] described in detail the adaptive biomarker-guided clinical trial designs they encountered in a review that covered 171 papers and 14,436 candidate abstracts. None of the eight designs they describe explicitly incorporates a co-primary outcome.

Finally, our PubMed search returned Bryant & Day[26]. This is perhaps the best known and widely used phase II trial design for studying efficacy and toxicity. Theirs is a two-stage method that offers a chance to reject a treatment for being inactive or excessively toxic at an interim stage. The design takes threshold values for the probabilities of efficacy and toxicity that are acceptable and unacceptable and returns the minimum number of efficacy events and maximum number of toxicity events that should be observed to approve the treatment for further study. For given levels of statistical significance and power, the threshold event counts define the optimal trial of the competing outcomes of efficacy and toxicity. Their method considers different levels of association between

efficacy and toxicity events and chooses an optimal design. The design implicitly assumes that the patient population is homogeneous thus it does not use predictive variables.

2 Simulating cohort membership

In simulations, we will randomly sample cohort membership and this requires estimates of the cohort prevalences. In Table S9 in Garon *et al.*[27], 39% of the 824 patients screened with evaluable tumour sample had low PD-L1 expression, 38% medium and 23% high. Amongst TN patients, these percentages were 31%, 44% and 25%. Amongst PT patients, they were 41%, 36% and 23%. Testing for association between the two categories by chi-squared test yields $p = 0.049$, so there is reasonable evidence to believe that PD-L1 expression level is not identically distributed for TN and PT patients. Low PD-L1 scores appear to be less prevalent amongst TN patients.

The chief investigator of PePS2 expects approximately 50% of patients to have been previously treated, based on their experience with the patient population. Scaling the PD-L1 prevalences observed by Garon *et al.* in the TN and PT groups, we expect cohort membership probabilities

$$\tilde{\rho} = (0.157, 0.218, 0.124, 0.207, 0.180, 0.114)$$

For iteration j, we randomly sampled cohort membership probabilities, $\rho_j \sim Dirichlet(\hat{\rho})$, for $j = 1, ..., J$, where $\hat{\rho} = (15.7, 21.8, 12.4, 20.7, 18.0, 11.4)$ and J is the number of simulated trial iterations. In Section 5 of this supplementary appendix, we investigate the effect of alternative prevalences.

This yielded expected values that matched $\tilde{\rho}$ and 95% confidence intervals given in Table 1. With a randomly selected ρ_j, patient-level allocations to cohorts $1, ..., 6$ were randomly sampled from multinomial distributions with probability vector ρ_j. The mean cohort sizes and 95% confidence intervals are also shown in Table 1. These statistics are

	ρ_j	Num patients	
Cohort	95% CI	Mean	95% CI
1	(0.093, 0.234)	9.4	(3, 17)
2	(0.143, 0.303)	13.1	(6, 22)
3	(0.067, 0.195)	7.4	(2, 14)
4	(0.134, 0.292)	12.4	(6, 21)
5	(0.112, 0.261)	10.8	(4, 19)
6	(0.060, 0.183)	6.9	(2, 14)

Table 1: Simulated cohort prevalences and cohort sizes, based on 100,000 replicates.

based on 100,000 random samples. The distribution of these cohort sizes approximately concurred with our expectations.

The variance of Dirichlet random variables is determined by the size of the elements of the parameter vector, ρ. To consider alternatives and verify that we were using approximately the correct order of magnitude of randomness in our cohort allocations, we repeated the same exercise with Dirichlet parameter vectors $\hat{\rho}/10$ and $10\hat{\rho}$. The vector $\hat{\rho}/10$ yielded cohort sizes that were too wide, e.g. cohort sizes of zero were observed too frequently. The vector $10\hat{\rho}$ yielded cohort sizes that exhibited less variation, but looked plausible nonetheless. It is conservative to prepare for more variability rather than less, so we resolved to use $\hat{\rho}$.

3 Uniformative priors

As described in the manuscript we selected modestly informative priors in the BEBOP model for PePS2. To verify that inference was not being unduly influenced by our priors, we also investigated by simulation BEBOP performance under the completely uninformative priors shown in Table 2 in the same six scenarios used in the paper.

Table 2: Uninformative normal prior distributions for the elements of $\boldsymbol{\theta}$.

	μ	σ^2
α	0	10
β	0	10
γ	0	10
ζ	0	10
λ	0	10
ψ	0	10

Using the model structure described in the manuscript, these parameter priors yield prior beliefs on the efficacy and toxicity probabilities summarised in Table 3. The priors admit every possible value in the 95% credible intervals and are centred on 50% event rates, as we would expect.

Cohort	ProbEffL	ProbEff	ProbEffU	ProbToxL	ProbTox	ProbToxU
1	0.00	0.49	1.00	0.00	0.51	1.00
2	0.00	0.50	1.00	0.00	0.51	1.00
3	0.00	0.50	1.00	0.00	0.51	1.00
4	0.00	0.49	1.00	0.00	0.51	1.00
5	0.00	0.49	1.00	0.00	0.51	1.00
6	0.00	0.50	1.00	0.00	0.51	1.00

Table 3: Alternative uninformative priors to those in the main text.

These uninformative priors give the BEBOP operating characteristics in Table 4. These should be compared to those in Table 5 in the main manuscript. For comparison, we also give operating performance of beta-binomial models that use $\alpha = \beta = 0.001$ to give similarly uninformative priors.

We see that the modestly informative priors in the manuscript increase the probability of approving the treatment by approximately 1 or 2% in most cases. In scenarios 1 and

Scenario	Cohort	ProbEff	ProbTox	Odd	N	Eff	Tox	BEBOP	BetaBin
1	1	0.300	0.1	1.0	9.4	2.8	0.9	0.877	0.544
1	2	0.300	0.1	1.0	13.1	3.9	1.3	0.901	0.660
1	3	0.300	0.1	1.0	7.5	2.2	0.7	0.812	0.473
1	4	0.300	0.1	1.0	12.4	3.7	1.2	0.894	0.635
1	5	0.300	0.1	1.0	10.8	3.2	1.1	0.886	0.575
1	6	0.300	0.1	1.0	6.9	2.1	0.7	0.803	0.459
2	1	0.100	0.3	1.0	9.4	1.0	2.8	0.023	0.036
2	2	0.100	0.3	1.0	13.1	1.3	3.9	0.022	0.031
2	3	0.100	0.3	1.0	7.5	0.7	2.2	0.018	0.037
2	4	0.100	0.3	1.0	12.4	1.2	3.7	0.021	0.032
2	5	0.100	0.3	1.0	10.8	1.1	3.3	0.023	0.032
2	6	0.100	0.3	1.0	6.9	0.7	2.1	0.020	0.040
3	1	0.300	0.1	0.2	9.4	2.8	0.9	0.876	0.560
3	2	0.300	0.1	0.2	13.1	3.9	1.3	0.901	0.673
3	3	0.300	0.1	0.2	7.5	2.2	0.7	0.811	0.491
3	4	0.300	0.1	0.2	12.4	3.7	1.3	0.893	0.646
3	5	0.300	0.1	0.2	10.8	3.2	1.1	0.886	0.585
3	6	0.300	0.1	0.2	6.9	2.1	0.7	0.803	0.472
4	1	0.167	0.1	1.0	9.4	1.6	0.9	0.408	0.302
4	2	0.192	0.1	1.0	13.1	2.5	1.3	0.635	0.440
4	3	0.500	0.1	1.0	7.5	3.7	0.8	0.974	0.623
4	4	0.091	0.1	1.0	12.4	1.1	1.3	0.220	0.127
4	5	0.156	0.1	1.0	10.8	1.6	1.1	0.411	0.289
4	6	0.439	0.1	1.0	6.9	3.0	0.7	0.921	0.577
5	1	0.167	0.3	1.0	9.4	1.6	2.8	0.046	0.071
5	2	0.192	0.3	1.0	13.1	2.5	3.9	0.065	0.084
5	3	0.500	0.3	1.0	7.5	3.7	2.2	0.102	0.164
5	4	0.091	0.3	1.0	12.4	1.1	3.8	0.024	0.028
5	5	0.156	0.3	1.0	10.8	1.6	3.3	0.043	0.057
5	6	0.439	0.3	1.0	6.9	3.0	2.1	0.096	0.165
6	1	0.167	0.1	0.2	9.4	1.6	0.9	0.406	0.317
6	2	0.192	0.1	0.2	13.1	2.5	1.3	0.633	0.452
6	3	0.500	0.1	0.2	7.5	3.7	0.7	0.974	0.640
6	4	0.091	0.1	0.2	12.4	1.1	1.3	0.217	0.135
6	5	0.156	0.1	0.2	10.8	1.6	1.1	0.407	0.298
6	6	0.439	0.1	0.2	6.9	3.0	0.7	0.922	0.580

Table 4: BEBOP performance under uninformative priors. Compare to Table 5 in the main manuscript. ProbEff and ProbTox are the true probabilities of efficacy and toxicity. Odds denotes the ratio of odds of efficacy in patients that experience toxicity to those that do not. Odds=1 corresponds to no association; values less than one convey that efficacy is less likely when toxicity is observed; and vice-versa. N is the mean number of patients in a cohort; Eff and Tox the mean number of events. BEBOP is the probability that treatment is approved by the BEBOP model; BetaBin the probability it is approved by a beta-binomial model. 10,000 iterations were used in each scenario.

3, the probability of approving the treatment in cohorts 3 and 6 is increased by up to 9% by the paper priors compared to these uninformative priors. This is using flat efficacy profiles. Under the inclined efficacy profile in scenarios 4-6, there is no extra effect in cohorts 3 and 6, nor is the design more likely to approve in the adverse scenario 2.

4 Informative priors

We also consider here more informative priors for $\boldsymbol{\theta}$.

Table 5: Overall Response probabilities reported in Garon *et al.*[27].

	PD-L1		
	Low	Medium	High
Treatment naive	0.167	0.192	0.500
Pre-treated	0.091	0.156	0.439

The Overall Response probabilities observed in Garon *et al.*[27] are given in Table 5. To provide an interesting alternative to our priors in the main manuscript, we seek priors that anticipate efficacy probabilities broadly matching those in Table 5, and with slightly more precision (i.e. narrower credible intervals) than those in the manuscript.

Table 6: Informative normal prior distributions for the elements of $\boldsymbol{\theta}$.

	μ	σ^2
α	0.0	$1.3^2 = 1.69$
β	-1.0	1.69
γ	-2.75	1.69
ζ	-2.2	1.69
λ	-2.2	4
ψ	0	1

The parameter priors in Table 6 yield the event rate priors in Table 7.

Cohort	ProbEffL	ProbEff	ProbEffU	ProbToxL	ProbTox	ProbToxU
1	0.00	0.14	0.71	0.00	0.21	0.86
2	0.00	0.19	0.80	0.00	0.21	0.86
3	0.07	0.50	0.93	0.00	0.21	0.86
4	0.00	0.10	0.64	0.00	0.21	0.86
5	0.00	0.13	0.77	0.00	0.21	0.86
6	0.01	0.34	0.93	0.00	0.21	0.86

Table 7: Alternative informative priors to those in the main text. The event rates broadly match those in the Garon study, and the credible intervals are narrower than the priors used in the paper.

Operating characteristics for BEBOP using these priors in Table 8 show that the probability of approving in cohorts 3 and 6 is generally increased. This reflects what we would expect given the prior beliefs of efficacy in these cohorts. As undesirable effect is

Scenario	Cohort	ProbEff	ProbTox	Odd	N	Eff	Tox	BEBOP
1	1	0.300	0.1	1.0	9.4	2.8	0.9	0.871
1	2	0.300	0.1	1.0	13.1	3.9	1.3	0.914
1	3	0.300	0.1	1.0	7.5	2.2	0.7	0.999
1	4	0.300	0.1	1.0	12.4	3.7	1.2	0.807
1	5	0.300	0.1	1.0	10.8	3.2	1.1	0.842
1	6	0.300	0.1	1.0	6.9	2.1	0.7	0.995
2	1	0.100	0.3	1.0	9.4	1.0	2.8	0.016
2	2	0.100	0.3	1.0	13.1	1.3	3.9	0.022
2	3	0.100	0.3	1.0	7.5	0.7	2.2	0.130
2	4	0.100	0.3	1.0	12.4	1.2	3.7	0.008
2	5	0.100	0.3	1.0	10.8	1.1	3.3	0.009
2	6	0.100	0.3	1.0	6.9	0.7	2.1	0.082
3	1	0.300	0.1	0.2	9.4	2.8	0.9	0.869
3	2	0.300	0.1	0.2	13.1	3.9	1.3	0.914
3	3	0.300	0.1	0.2	7.5	2.2	0.7	0.998
3	4	0.300	0.1	0.2	12.4	3.7	1.3	0.806
3	5	0.300	0.1	0.2	10.8	3.2	1.1	0.841
3	6	0.300	0.1	0.2	6.9	2.1	0.7	0.994
4	1	0.167	0.1	1.0	9.4	1.6	0.9	0.379
4	2	0.192	0.1	1.0	13.1	2.5	1.3	0.678
4	3	0.500	0.1	1.0	7.5	3.7	0.8	0.999
4	4	0.091	0.1	1.0	12.4	1.1	1.3	0.140
4	5	0.156	0.1	1.0	10.8	1.6	1.1	0.347
4	6	0.439	0.1	1.0	6.9	3.0	0.7	0.983
5	1	0.167	0.3	1.0	9.4	1.6	2.8	0.054
5	2	0.192	0.3	1.0	13.1	2.5	3.9	0.090
5	3	0.500	0.3	1.0	7.5	3.7	2.2	0.132
5	4	0.091	0.3	1.0	12.4	1.1	3.8	0.020
5	5	0.156	0.3	1.0	10.8	1.6	3.3	0.044
5	6	0.439	0.3	1.0	6.9	3.0	2.1	0.130
6	1	0.167	0.1	0.2	9.4	1.6	0.9	0.379
6	2	0.192	0.1	0.2	13.1	2.5	1.3	0.678
6	3	0.500	0.1	0.2	7.5	3.7	0.7	0.999
6	4	0.091	0.1	0.2	12.4	1.1	1.3	0.140
6	5	0.156	0.1	0.2	10.8	1.6	1.1	0.346
6	6	0.439	0.1	0.2	6.9	3.0	0.7	0.984

Table 8: BEBOP performance under informative priors. Compare to Table 5 in the main manuscript. ProbEff and ProbTox are the true probabilities of efficacy and toxicity. Odds denotes the ratio of odds of efficacy in patients that experience toxicity to those that do not. Odds=1 corresponds to no association; values less than one convey that efficacy is less likely when toxicity is observed; and vice-versa. N is the mean number of patients in a cohort; Eff and Tox the mean number of events. BEBOP is the probability that treatment is approved by the BEBOP model; BetaBin the probability it is approved by a beta-binomial model. 10,000 iterations were used in each scenario.

observed in scenario 2, where there is an increased chance of inappropriately approving in cohorts 3 and 6, as the priors exert too much influence on the posterior. In comparison, this supports our choice of *modestly informative* priors in the paper. In general, the differences concur with what we would expect with standard Bayesian theory, i.e. the posterior is a blend of prior and data, and with stronger priors, the prior has greater weight in the posterior.

5 Alternative cohort prevalences

In all of the simulations presented in the main body, we use the parameter vector $\hat{\rho} = (15.7, 21.8, 12.4, 20.7, 18.0, 11.4)$ to sample cohort memberships, for the reasons described. For clarity in this section, we refer to that set of prevalences derived for the PePS2 trial as $\hat{\rho}_P$. We investigate the sensitivity of our BEBOP implementation to the prevalences used by comparing performance under the alternative vector $\hat{\rho}_A = (16.67, 16.67, 16.67, 16.67, 16.67, 16.67)$, labelled with subscript A to denote it as an alternative. Under $\hat{\rho}_A$, patients are uniformly distributed amongst the six cohorts and the expected size of each is 10 patients.

Table 9: Comparison of BEBOP performance in scenario 4 of Table 5 in the main text using the cohort prevalences derived in the main body and alternative, uniform prevalences. N is the expected cohort size. ProbApprove is the probability of the BEBOP design approving the treatment.

	Scenario 4		$\hat{\rho}_P$		$\hat{\rho}_A$	
Cohort	ProbEff	ProbTox	N	ProbApprove	N	ProbApprove
1	0.167	0.1	9.4	0.460	10.0	0.472
2	0.192	0.1	13.1	0.685	10.0	0.658
3	0.500	0.1	7.5	0.982	10.0	0.993
4	0.091	0.1	12.4	0.282	10.0	0.310
5	0.156	0.1	10.8	0.483	10.0	0.511
6	0.439	0.1	6.9	0.920	10.0	0.967

Table 9 compares the performance of BEBOP designs using cohort prevalences $\hat{\rho}_P$ and $\hat{\rho}_A$ in scenario 4 of our simulations in the main text. The first thing to note is that the probability of approving treatment has changed by no more than 5% in any cohort. The operating characteristics do not change in any great degree and the inferences do not change at all. The method performs as we might expect from a regression analysis.

6 Appendix

6.1 Practical Steps for Implementing BEBOP

Trialists should assess the operating performance of a BEBOP design in theoretical scenarios using computer simulation. At the very least, we conduct simulations to estimate the probability that a design will incorrectly approve a poor treatment (similar to the notion of significance in frequentist trial designs) and correctly approve a good treatment (essentially, statistical power). Simulated trials are conducted by randomly sampling outcomes for notional patients and invoking the acceptance decision determined by (9) in the main text at the final (and potentially also interim) stages. There are a number of choices to make when implementing BEBOP. Prior to simulating performance, trialists should:

1. Specify forms for the marginal efficacy and toxicity models (2) and (3) in the main text.

2. Specify a form for the joint model. (1) and (4) in the main text are two options.

3. Specify $f(\boldsymbol{\theta})$, the prior distribution for $\boldsymbol{\theta}$.

4. Specify efficacy and toxicity thresholds, π_E^*, π_T^* based on clinical rationale. These may vary by cohort or they may be common, as the clinical scenario dictates.

With this information, the trialists may simulate trial data sets, $\boldsymbol{X}_j$ for $j = 1, ..., J$ and infer the decision of the design on each. Values for p_E and p_T need not be specified before simulations are run. Instead, it is more flexible to record the value for $\Pr(\pi_E(x_i, \boldsymbol{\theta}) > \pi_E^* | \boldsymbol{X})$ and $\Pr(\pi_T(x_i, \boldsymbol{\theta}) < \pi_T^* | \boldsymbol{X})$ in simulated iterations for each distinct value of x_i. Then, the trialists may adjust the performance of the design by considering different values for p_E and p_T, inferring the operating characteristics of the pair by invoking (9) on the simulated output.

In summary, the values for π_E^*, π_T^* are based on clinical rationale and set at run-time.

In contrast, the values of p_E and p_T need not be, so it is easier to tweak model operating characteristics by varying p_E and p_T. We used this algorithm in the main text.

References

[1] Thomas M. Braun. "The bivariate continual reassessment method: Extending the CRM to phase I trials of two competing outcomes". In: *Controlled Clinical Trials* 23.3 (2002), pp. 240–256. ISSN: 01972456. DOI: 10.1016/S0197-2456(01)00205-7.

[2] J O'Quigley, M Pepe, and L Fisher. "Continual reassessment method: a practical design for phase 1 clinical trials in cancer." In: *Biometrics* 46.1 (1990), pp. 33–48. ISSN: 0006-341X. DOI: 10.2307/2531628.

[3] Anastasia Ivanova. "A New Dose-Finding Design for Bivariate Outcomes". In: *Biometrics* 59.4 (2003), pp. 1001–1007. ISSN: 0006341X. DOI: 10.1111/j.0006-341X.2003.00115.x.

[4] Wei Zhang, Daniel J. Sargent, and Sumithra Mandrekar. "An adaptive dose-finding design incorporating both toxicity and efficacy". In: *Statistics in Medicine* 25.14 (2006), pp. 2365–2383. ISSN: 02776715. DOI: 10.1002/sim.2325.

[5] Sumithra J. Mandrekar, Rui Qin, and Daniel J. Sargent. "Model-based phase I designs incorporating toxicity and efficacy for single and dual agent drug combinations: Methods and challenges". In: *Statistics in Medicine* 29.10 (2010), pp. 1077–1083. ISSN: 02776715. DOI: 10.1002/sim.3706.

[6] Meihua Wang and Roger Day. "Adaptive Bayesian design for phase I dose-finding trials using a joint model of response and toxicity." In: *Journal of biopharmaceutical statistics* 20.1 (2010), pp. 125–144. ISSN: 1054-3406. DOI: 10.1080/10543400903280613.

[7] PF Thall and JD Cook. "Dose-Finding Based on Efficacy-Toxicity Trade-Offs". In: *Biometrics* 60.3 (2004), pp. 684–693.

[8] John D Cook. *Efficacy-Toxicity trade-offs based on L-p norms: Technical Report UTMDABTR-003-06*. Tech. rep. 2006, pp. 1–9.

[9] PF Thall, JD Cook, and EH Estey. "Adaptive dose selection using efficacy-toxicity trade-offs: illustrations and practical considerations." In: *Journal of biopharmaceutical statistics* 16.5 (2006), pp. 623–638. ISSN: 1054-3406. DOI: 10.1080/10543400600860394.

[10] Aklilu Habteab Ghebretinsae et al. "Joint modeling of hierarchically clustered and overdispersed non-gaussian continuous outcomes for comet assay data". In: *Pharmaceutical Statistics* 11.6 (2012), pp. 449–455. ISSN: 15391604. DOI: 10.1002/pst.1533.

[11] R J Cook and V T Farewell. "Guidelines for monitoring efficacy and toxicity responses in clinical trials." In: *Biometrics* 50.4 (1994), pp. 1146–1152. ISSN: 0006-341X. DOI: 10.2307/2533451.

[12] Hua Jin. "Alternative designs of phase II trials considering response and toxicity". In: *Contemporary Clinical Trials* 28.4 (2007), pp. 525–531. ISSN: 15517144. DOI: 10.1016/j.cct.2007.03.003.

[13] P. Brutti, S. Gubbiotti, and V. Sambucini. "An extension of the single threshold design for monitoring efficacy and safety in phase II clinical trials". In: *Statistics in Medicine* 30.14 (2011), pp. 1648–1664. ISSN: 02776715. DOI: 10.1002/sim.4229.

[14] A. Bouckaert and M. Mouchart. "Sure outcomes of random events: A model for clinical trials". In: *Statistics in Medicine* 20.4 (2001), pp. 521–543. ISSN: 02776715. DOI: 10.1002/sim.659.

[15] M R Conaway and G R Petroni. "Bivariate sequential designs for phase II trials." In: *Biometrics* 51.2 (1995), pp. 656–664. ISSN: 0006341X. DOI: 10.2307/2532952.

[16] M R Conaway and G R Petroni. "Designs for phase II trials allowing for a trade-off between response and toxicity." In: *Biometrics* 52.4 (1996), pp. 1375–1386. ISSN: 0006-341X. DOI: 10.2307/2532851.

[17] Peter F Thall, Richard M Simon, and Elihu H Estey. "Single-Arm Clinical Trials With Multiple Outcomes". In: *Statistics in Medicine* 14.October 1993 (1995), pp. 357–379.

[18] P F Thall, R M Simon, and E H Estey. "New statistical strategy for monitoring safety and efficacy in single- arm clinical trials". In: *J.Clin.Oncol.* 14.0732-183X SB - M SB - X (1996), pp. 296–303. ISSN: 0732-183X.

[19] Peter F. Thall and Hsi Guang Sung. "Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials". In: *Statistics in Medicine* 17.14 (1998), pp. 1563–1580. ISSN: 02776715. DOI: 10.1002/(SICI)1097-0258(19980730)17:14<1563::AID-SIM873>3.0.CO;2-L.

[20] J. Kyle Wathen et al. "Accounting for patient heterogeneity in phase II clinical trials J." In: *Statistics in medicine* 27 (2008), pp. 2802–2815. ISSN: 02776715. DOI: 10.1002/sim.

[21] Marc Buyse et al. "Integrating biomarkers in clinical trials". In: *Expert Review of Molecular Diagnostics2* 11.2 (2011), pp. 171–182. ISSN: 1473-7159. DOI: 10.1586/ERM.10.120.

[22] Yung Jue Bang et al. "Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): A phase 3, open-label, randomised controlled trial". In: *The Lancet* 376.9742 (2010), pp. 687–697. ISSN: 01406736. DOI: 10.1016/S0140-6736(10)61121-X. URL: http://dx.doi.org/10.1016/S0140-6736(10)61121-X.

[23] Edward S. Kim et al. "The BATTLE trial: Personalizing Therapy for Lung Cancer". In: *Cancer Discovery* 1.1 (2011), pp. 44–53. ISSN: 21598274. DOI: 10.1158/2159-8274.CD-10-0010. arXiv: 1112.3563.

[24] Boris Freidlin, Lisa M. McShane, and Edward L. Korn. "Randomized clinical trials with biomarkers: Design issues". In: *Journal of the National Cancer Institute* 102.3 (2010), pp. 152–160. ISSN: 00278874. DOI: 10.1093/jnci/djp477.

[25] Miranta Antoniou, Andrea L Jorgensen, and Ruwanthi Kolamunnage-Dona. "Biomarker-Guided Adaptive Trial Designs in Phase II and Phase III: A Methodological Review". In: *Plos One* 11.2 (2016), e0149803. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0149803. URL: http://dx.plos.org/10.1371/journal.pone.0149803.

[26] J Bryant and R Day. "Incorporating toxicity considerations into the design of two-stage phase II clinical trials." In: *Biometrics* 51.4 (1995), pp. 1372–1383. ISSN: 0006-341X. DOI: 10.2307/2533268.

[27] Edward B Garon et al. "Pembrolizumab for the treatment of non-small-cell lung cancer." In: *The New England journal of medicine* 372.21 (2015), pp. 2018–28. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1501824. URL: http://www.ncbi.nlm.nih.gov/pubmed/25891174.

[28] R Herrick et al. *EffTox*. 2015. URL: https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software{_}Id=2.