A Phase II Clinical Trial with Efficacy and Toxicity Outcomes and Baseline Covariates

Kristian Brock, Lucinda Billingham, Christina Yap and Gary Middleton

Abstract PePS2 is a phase II trial of the efficacy and safety of pembrolizumab in performance status 2 non-small-cell lung cancer patients. Previous studies have shown that efficacy is correlated with the extent to which PD-L1 is expressed in the tumour, and pretreatedness. There are few clinical trial designs that test co-primary efficacy and toxicity outcomes in phase II, and fewer still that incorporate baseline covariates. Thall, Nguyen and Estey present one such design but it has been scarcely used in trials. Their model incorporates terms to conduct a dose-finding study. This aspect is not required in PePS2 because a candidate dose has been widely tested. We introduce a novel simplification for phase II that focuses on testing efficacy and toxicity whilst adjusting for baseline covariates. The method shares information across cohorts. Simulations show it is far more efficient than analysing cohorts separately. Using the design in PePS2 with 60 patients to test the treatment in six cohorts, we can expect error rates typical of those used in phase II trials. However, we demonstrate that care must be used when specifying the models for efficacy and toxicity because more complex models require greater sample sizes for bias to be controlled.

Kristian Brock
University of Birmingham, UK, B15 2TT, e-mail: k.brock@bham.ac.uk

Lucinda Billingham
University of Birmingham, UK, B15 2TT, e-mail: l.j.billingham@bham.ac.uk

Christina Yap
University of Birmingham, UK, B15 2TT, e-mail: c.yap@bham.ac.uk

Gary Middleton
University of Birmingham, UK, B15 2TT, e-mail: g.middleton@bham.ac.uk

1 Introduction

There is a relative dearth of phase II clinical trial designs that incorporate patient covariates to assess efficacy and toxicity. Thall *et al.*[10] introduced a family of methods that perform dose-finding trials guided by binary efficacy and toxicity outcomes whilst accounting for baseline patient covariates. This enables dose recommendations tailored to individual patients. Our motivation is PePS2, a phase II trial of pembrolizumab in non-small-cell lung cancer patients of performance status 2. PePS2 is not a dose-finding trial. Instead, it seeks to estimate the probabilities of efficacy and toxicity at a dose of pembrolizumab previously demonstrated to be safe and effective in a closely-related group of patients[6]. In this piece, we implement a simplification of Thall *et al.*'s method. We remove the dose-finding components but retain aspects to study co-primary efficacy and toxicity outcomes that are associated with baseline covariates.

2 The PePS2 Trial

PePS2 is a phase II trial of pembrolizumab in non-small-cell lung cancer (NSCLC) patients with Eastern Cooperative Oncology Group (ECOG) performance status 2 (PS2). A patient with PS2 is ambulatory and capable of taking care of themselves but typically too ill to work. Critically, it is doubtful that a PS2 patient could tolerate the toxic side effects of chemotherapy.

The joint primary outcomes of the trial are (i) *toxicity*, defined as the occurrence of a treatment-related dose delay or treatment discontinuation due to adverse event related to pembrolizumab; and (ii) *efficacy*, defined as the occurrence of stable disease (SD), partial response (PR) or complete response (CR) without prior progressive disease (PD) at or after the second post-baseline disease assessment by RECIST v1.1[5], scheduled to occur at week 18. The primary objective of the trial is to learn if the treatment is associated with sufficient disease control with acceptably low toxicity to use in performance status 2 patients.

Pembrolizumab inhibits the programmed cell death 1 (PD-1) receptor via the programmed death-ligand 1 (PD-L1) protein. In a phase I study with 495 patients, Garon *et al.*[6] showed pembrolizumab to be active and tolerable in performance status 0 & 1 patients. Overall, 19.4% of patients had an objective response (PR or CR) and 9.5% experienced an adverse event of grade 3 or higher. The rate of toxicity compares favourably to those typically seen in advanced NSCLC patients using chemotherapy [9, 1]. We foresee no reason why these outcome rates should be materially dissimilar in PS2 patients.

Garon *et al.* introduce the PD-L1 proportion score biomarker, defined as the percentage of neoplastic cells with staining for membranous PD-L1. Efficacy outcomes for the 204 patients in their validation group, allocated to cohorts based on PD-L1 score, are shown in Table 1. Objective responses are observed in all cohorts and the rate of response increases with PD-L1 score.

Table 1 Objective response rates for the validation sample ($n = 204$) of Garon *et al.*[6].

PD-L1 Cohort	Criteria	Objective Response %, (95% CI)
Low	PD-L1 score $< 1\%$	10.7 (2.3, 28.2)
Medium	$1\% \geq$ PD-L1 score $< 50\%$	16.5 (9.9, 25.1)
High	PD-L1 score $\geq 50\%$	45.2 (33.5, 57.3)

Based on this information, we expect PD-L1 score to be predictive of response in our PS2 population. Additionally, in the Garon trial 24.8% of treatment-naive (TN) patients achieved a response, whereas only 18.0% did in the pre-treated (PT) patients. Pretreatment status represents a potentially small but important effect that should be considered when testing the treatment. We propose to investigate drug in the six cohorts formed by jointly stratifying by the three Garon PD-L1 classifications; and the PT or TN statuses. Each patient in PePS2 will belong to exactly one of these six cohorts, as demonstrated in Table 2.

Cohort	Treatment status	PD-L1 category	$x_i = (x_{1i}, x_{2i}, x_{3i})$
1	Treatment naive	Low	(0,1,0)
2	Treatment naive	Medium	(0,0,1)
3	Treatment naive	High	(0,0,0)
4	Pretreated	Low	(1,1,0)
5	Pretreated	Medium	(1,0,1)
6	Pretreated	High	(1,0,0)

Table 2 Cohorts used in the PePS2 trial. x_i shows the predictive variable vector.

In phase II, there is strong motivation to deliver findings quickly to inform the next study phase. Recruitment of approximately 60 PS2 patients within one year would be feasible but accrual materially higher would be unlikely. Given the relative dearth of treatment alternatives, we seek to offer the trial to all PS2 patients and not stratify accrual. Pembrolizumab has not been investigated in PS2 patients so the clinical scenario requires a trial design that tests efficacy and toxicity. Given the evidence that PD-L1 and pretreatedness are associated with response, it is highly desirable to use a trial design that incorporates this predictive information. We describe our search for a clinical trial design that achieves these objectives.

3 Review of Available Experimental Designs

We sought trial designs that use 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'.

Several manuscripts were identified detailing dose-finding designs that scrutinise both efficacy and toxicity[2, 11, 12]. Each of these presents a model that could be

adapted to our purpose. Very few phase II designs with co-primary outcomes were identified. Bryant & Day[3] take threshold values for required rates of efficacy and toxicity and return the threshold number of events to approve the treatment. For given levels of significance and power, the threshold counts define the optimal trial of competing efficacy and toxicity outcomes. The design implicitly assumes that the patient population is homogeneous and does not use covariates. Parallel Bryant & Day designs in our cohorts would require a prohibitively high sample size.

Finally, Thall, Nguyen & Estey (TNE)[10] introduce an extension of EffTox[11] that adds baseline patient covariates to analyse co-primary efficacy and toxicity at different doses. The objective of their Bayesian design is to recommend a personal dose of an experimental agent, after adjusting for baseline covariates. Their design has enjoyed limited use. On 05-Dec-2017, we identified 16 manuscripts listed on PubMed that cite Thall *et al.*[10], including 10 concerning further dose-finding methodology. None sought to adapt the design to the typical phase II task of investigating efficacy and toxicity at a single dose. Five papers were reviews. Konopleva, Thall *et al.*[8] use TNE in a dose-finding study of PR104 in relapsed or refractory AML and acute lymphoblastic leukaemia (ALL). This literature search suggests that TNE's method has only been used in blood cancer and only for the purposes of dose-finding. We found no suggestion that the method had been adapted for the non-dose-finding context. We introduce a simplification of TNE for use in phase II.

4 Assessing efficacy and toxicity and adjusting for covariates

In this section, we describe novel adaptations to Thall *et al.*[10] to derive a model that studies associated co-primary efficacy and toxicity outcomes, adjusted for baseline covariates. We call this design P2TNE, for *Phase II Thall, Nguyen & Estey*.

Where prior information is available on outcomes under historic treatments, Thall *et al.* present the general probability model

$$\operatorname{logit}\pi_k(\tau,Z,\theta) = \beta_k Z + \sum_{j=1}^{m}(\mu_{k,j}+\xi_{k,j}Z)\mathbb{I}(\tau=\tau_j) + \{g_k(x,\alpha_k)+\gamma_k Z\}\,\mathbb{I}(\tau=x) \tag{1}$$

for $k = E, T$ denoting efficacy and toxicity, respectively. Here, τ is the given dose; Z is a vector of covariates; θ is the vector of model parameters to be estimated; β_k is the vector of main covariate effects; $\mu_{k,j}$ is a vector of historic main treatment effects for m informative historic treatments; $\xi_{k,j}$ is vector of interactions between historic treatment j and Z; $\mathbb{I}(A)$ is the indicator function, taking value 1 if A is true; $g_k(x,\alpha_k)$ characterises the main dose effects; γ_k is a vector of dose-covariate interactions.

The authors introduce methods for associating the marginal efficacy and toxicity models, and in their example reuse the Gumbel model deployed in EffTox[11]:

$$\pi_{a,b}(\pi_E,\pi_T) = (\pi_E)^a(1-\pi_E)^{1-a}(\pi_T)^b(1-\pi_T)^{1-b} + (-1)^{a+b}(\pi_E)(1-\pi_E)(\pi_T)(1-\pi_T)\frac{e^{\psi}-1}{e^{\psi}+1} \quad (2)$$

where ψ is an association parameter and a,b take the value 1 when efficacy and toxicity occur respectively in a given patient, else 0.

We simplify this in P2TNE by removing dose-effect terms $(g(.) = \gamma = 0)$ and considering historic outcomes only for the treatment under investigation $(m = 1)$.

Let x_i denote the covariate data as specified in Table 2, and a_i, b_i the occurrence of efficacy and toxicity in patient i. For trial data

$$X = \{(x_1,a_1,b_1),...,(x_n,a_n,b_n)\} \quad (3)$$

the aggregate likelihood function is

$$\mathscr{L}(X,\theta) = \prod_{i=1}^{n} \pi_{a_i,b_i}(x_i,\theta) \quad (4)$$

Let θ have prior distribution function $f(\theta)$. For patients with covariate data x, the posterior expectation of the probability of efficacy under the treatment is

$$\mathbb{E}(\pi_E(x,\theta)|X) = \frac{\int \pi_E(x,\theta) f(\theta)\mathscr{L}(X,\theta)d\theta}{\int f(\theta)\mathscr{L}(X,\theta)d\theta} \quad (5)$$

and the posterior probability that the rate of efficacy exceeds some threshold π_E^* is

$$\Pr(\pi_E(x,\theta) > \pi_E^*|X) = \frac{\int \mathbb{I}(\pi_E(x,\theta) > \pi_E^*) f(\theta)\mathscr{L}(X,\theta)d\theta}{\int f(\theta)\mathscr{L}(X,\theta)d\theta} \quad (6)$$

The treatment is acceptable in patients with covariates x if

$$\begin{aligned} \Pr(\pi_E(x,\theta) > \pi_E^*|X) > p_E \\ \Pr(\pi_T(x,\theta) < \pi_T^*|X) > p_T \end{aligned} \quad (7)$$

where π_E^*, p_E, π_T^* and p_T are chosen by the trialists. Our lead clinician selected the values $\pi_E^* = 0.1$ and $\pi_T^* = 0.3$ in all cohorts. We identified that $p_E = 0.7$ and $p_T = 0.9$ gave acceptable performance in indicative scenarios in our simulation study described below. Our chosen models for marginal efficacy and toxicity are:

$$\begin{aligned} \text{logit}\,\pi_E(x_i,\theta) &= \alpha + \beta x_{1i} + \gamma x_{2i} + \zeta x_{3i} \\ \text{logit}\,\pi_T(x_i,\theta) &= \lambda \end{aligned} \quad (8)$$

, associated by (2). The four parameters in the efficacy model assume the log-odds for PT patients in each PD-L1 category are a common linear shift of those in TN patients, an assumption we call *piecewise parallelism*. Furthermore, the rate of toxicity is assumed uniform across groups, justified by data reported in Garon *et al.* and Herbst *et al.*[7]. We analyse models that relax these assumptions.

5 Simulation study

In a simulation study, we show that our model achieves error rates typical of phase II clinical trials in all cohorts in indicative scenarios where efficacy and toxicity rates are uniform across cohorts. Furthermore, we show that performance is very strong in heterogeneous scenarios inspired by outcomes from Garon[6] and Herbst[7], far surpassing that of beta-binomial conjugate analyses conducted in cohorts individually.

Choice of priors is contentious in clinical trials. We simulated performance under diffuse, sceptical, and informative priors. Bias, inferred by coverage of posterior credible intervals, was highest under the diffuse priors, suggesting benefit to using priors that provide some central tendency that truly reflect investigators' beliefs.

Our model choices (8) imply fairly strong assumptions. We analyse several model embellishments to infer the cost in required sample size of greater model freedom.

We relax the piecewise parallel assumption by adding interactions terms to the efficacy model (8). Using this model in the scenarios described above, the probability of approval decreases and bias increases. We find that 20 - 40 extra patients are required to restore performance to that previously observed and eradicate bias.

Furthermore, we relax the assumption that toxicity is uniform over groups by adding terms to the toxicity model. Again, the extra model freedoms without increases in sample size yield lower approval probabilities and increased bias. Bias is a particular problem in the toxicity model in scenarios when the expected event rate is low. For instance, a four-parameter toxicity model suffers from material bias estimating uniform 10% toxicity rates in all cohorts, particularly in the groups with smallest expected size. This model is unbiased when the assumed true toxicity rate is 30%. This is notable because the published data[6, 7] suggest low toxicity. This model successfully identifies differential toxicity associated with covariates.

Lastly, simulations reveal that model performance is seemingly not affected by efficacy and toxicity events being strongly associated. We investigated a model variant that assumes independent events by setting $\psi = 0$ in (2). The probability of treatment approval was practically unchanged. This is perhaps not surprising when we consider that ψ is not present in the marginal models for π_E or π_T. The association parameter would aid inference if outcomes were partly observed so that, for instance, the toxicity outcome is known but efficacy is unknown.

6 Further work and availability of materials

Statisticians are aware of the information loss that arises from dichotomising continuous variables. We use the PD-L1 categorisation published by Garon *et al.*[6]. In ongoing work, we use the underlying continuous score.

Models are implemented in Stan[4] and all materials are available on GitHub at: https://github.com/brockk/bebop

References

[1] H. Borghaei et al. “Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer.” In: *The New England journal of medicine* (2015), pp. 1–13. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1504627. URL: http://www.ncbi.nlm.nih.gov/pubmed/26028407.

[2] 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.

[3] 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.

[4] Bob Carpenter et al. “Stan: A Probabilistic Programming Language”. In: *Journal of Statistical Software* VV.Ii (2016).

[5] E. a. Eisenhauer et al. “New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)”. In: *European Journal of Cancer* 45.2 (2009), pp. 228–247. ISSN: 09598049. DOI: 10.1016/j.ejca.2008.10.026. URL: http://dx.doi.org/10.1016/j.ejca.2008.10.026.

[6] 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.

[7] Roy S. Herbst et al. “Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): A randomised controlled trial”. In: *The Lancet* 387.10027 (2016), pp. 1540–1550. ISSN: 1474547X. DOI: 10.1016/S0140-6736(15)01281-7.

[8] Marina Konopleva et al. “Phase I/II study of the hypoxia-activated prodrug PR104 in refractory/relapsed acute myeloid leukemia and acute lymphoblastic leukemia”. In: *Haematologica* 100.7 (2015), pp. 927–934. ISSN: 15928721. DOI: 10.3324/haematol.2014.118455.

[9] Joan H Schiller et al. “Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer.” In: *The New England journal of medicine* 346.2 (2002), pp. 92–8. ISSN: 1533-4406. DOI: 10.1056/NEJMoa011954. URL: http://www.ncbi.nlm.nih.gov/pubmed/11784875.

[10] Peter F. Thall, Hoang Q. Nguyen, and Elihu H. Estey. “Patient-specific dose finding based on bivariate outcomes and covariates”. In: *Biometrics* 64.4 (2008), pp. 1126–1136. ISSN: 0006341X. DOI: 10.1111/j.1541-0420.2008.01009.x.

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

[12] 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.