Statistical Analysis Plan

A phase II trial of pembrolizumab in patients with non-small cell lung cancer and a performance status of 2

Version: 5.0, 31-Jan-2019

Sponsor	University of Birmingham
Sponsor number	RG_14-172
CRCTU number	LU2008
EudraCT number	2015-002241-55

Author(s):

Name: Kristian Brock **Signature:**

Trial role: Trial statistician

Date:

Reviewed and approved by:

Name: Lucinda Billingham **Signature:**

Trial role: Lead Biostatistician

Date:

Name: Gary Middleton **Signature:**

Trial role: Chief Investigator

Date:

Cancer Research UK Clinical Trials Unit @CRCTU

Key personnel involved in the Statistical Analysis Plan:

Name	Trial role
Kristian Brock	Trial statistician
Prof Lucinda Billingham	Lead Biostatistician
Prof Gary Middleton	Chief Investigator
Rhys Mant	Trial coordinator

Document Control Sheet

Statistical Analysis Plan version & date:	Reason for update:	To match protocol version & date:
0.1, 12-Sep-2016	Initial creation	2.0a, 07-Jun-2016
0.2, 19-Oct-2016	Revision after LB comments	2.0a, 07-Jun-2016
1.0, 19-Oct-2016	Initial release	2.0a, 07-Jun-2016
2.0, 18-May-2017	Created "Rules of Inference" section. Refined outcome definition in protocol and SAP.	4.0
3.0, 10-Apr-2018	Clarified intention to publish primary outcomes before the longer term secondary outcomes. Clarified that the co-primary DCB outcome is treated as success if: CR or PR or SD is measured at or after the second scheduled CT scan; and the scan date is at least 16 weeks from registration. The second scan is scheduled at 18 weeks but frequently these are conducted slightly early for reasons outside the control of the trials unit. The third scheduled scan is not until 27 weeks. Clarified that PD-L1 scores expressed as a range are interpreted as the midpoint, e.g. 10-20% is interpreted as PD-L1 = 15%. Added hierarchical model as a proposed supplementary method of analysing primary outcomes. The described BEBOP method requires that baseline covariates be present. This assumption became pertinent when we allowed PD-L1-assay failures to register. The hierarchical	5.0, 14-Aug-2017

	method does not mandate covariates be present. Added spider plot to specified graphical methods.	
4.0, 18-Jun-2018	Clarified the analysis method for DOR and DSD, and minor typographical changes.	6.0
5.0, 31-Jan-2019	Added prose describing the prior predictive event distributions of DCB and toxicity. Removed the proposal of a hierarchical analysis. This was proposed to include patients with missing covariate data, because "missing" could be included as another variable level whilst providing some shrinkage to avoid overfitting the relatively small group. However, the inherent exchangeability assumption is not appropriate. We added instead a method that will use the original proposed BEBOP model and multiply-impute missing covariates. Added plans to impute sum of target lesions when it is not reported but a RECIST response with respect to target lesions is reported, and vice versa. Clarified some minor points. Updated storage locations of snapshots, analyses, etc.	6.0, 10-Jul-2018

1. INTRODUCTION

Please refer to the protocol for an introduction to non-small-cell lung cancer (NSCLC) and pembrolizumab.

1.1 Purpose of the Statistical Analysis Plan

This Statistical Analysis Plan (SAP) provides guidelines for the analysis and presentation of results for the PePS2 trial. This plan, along with all other documents relating to the analysis of this trial, will be stored in the 'Statistical Documentation' section of the Trial Master File. The statistical analysis will be carried out by the Trial Statistician.

1.2 Trial Synopsis

Title

PePS2: A phase II trial of pembrolizumab in patients with non-small cell lung cancer and a performance status of 2

Trial Design

Multi-centre, single-arm phase II trial, testing pembrolizumab in a population of patients with non-small cell lung cancer (NSCLC) and an Eastern Cooperative Oncology Group (ECOG) performance status of 2.

Trial Objectives

This is a phase II trial of programmed cell death protein 1 (PD-1) blockade in patients with programmed death ligand 1 (PD-L1) defined NSCLC and an ECOG performance status of 2 with the primary purposes:

- To determine that pembrolizumab is safe and tolerable at the selected dose
- To detect the durable clinical benefit in this population of patients treated with pembrolizumab that would justify further investigation.

Secondary objective:

- To measure patient health related quality of life (HRQoL)

Exploratory objective:

- To discover possible biomarkers to predict for a response to pembrolizumab.

Outcome Measures

Primary outcome measures

- Toxicity, defined as the occurrence of a treatment-related dose delay or treatment discontinuation due to an adverse event
- Durable clinical benefit (DCB), defined as the occurrence of a complete response (CR), partial response (PR) or stable disease (SD) without prior progressive disease (PD) at or after the second scheduled CT scan (scheduled in the protocol to occur at 18 weeks).

Secondary outcome measures:

- Objective response (OR)
- Progression-free survival time (PFS)
- Time to progression (TTP)
- Overall survival time (OS)
- Health related quality of life (HRQoL)
- Duration of objective response (DOR) and duration of stable disease (DSD).

Sample Size

60 patients

Patient Population

Patients with non-small cell lung cancer and an ECOG performance status of 2. Analysis of response will be stratified by PD-L1 proportion score and pre-treatment status in order to assess the rate of response in each cohort. Please refer to protocol for full inclusion and exclusion criteria.

2. TIMING AND REPORTING OF INTERIM AND FINAL ANALYSES

Interim analyses of primary outcomes and safety data will be presented at annual trial steering committee (TSC) meetings. There are no formal stopping rules defined.

The primary outcome will be submitted for publication within 6 months of the resolution of both co-primary outcomes for all patients. Available data on secondary outcomes may be included. A further publication may be required to report the full information on all non-primary outcomes. This will be submitted 12 months of the last patient completing treatment.

3. RECRUITMENT AND RANDOMISATION

3.1 Recruitment

We will report:

- The number of patients registered, summarised by site;
- The date the snapshot was taken;

3.2 Ineligible Patients

Ineligible patients are defined as those registered patients who are subsequently found to not meet the protocol eligibility criteria.

The statistician will report:

- The number of ineligible patients, and reasons for their ineligibility.

The primary analysis will include those patients found to be ineligible but a sensitivity analysis may be conducted and reported if the number of ineligible patients is substantial. Protocol deviations relating to treatment will be reported as part of treatment compliance (Section 7).

4. DATA QUALITY

The length of patient follow-up will be estimated using a reverse Kaplan-Meier curve.

The statistician will report:

- Median (95% CI) follow-up;
- The number of patients with disease response assessments at each assessment time.

5. ANALYSIS POPULATIONS

The *registered population* will be the set of all patients registered to the trial.

The general *analysis population* will be the set of all patients that received pembrolizumab.

5.1 Baseline Patient Characteristics

The statistician will report summary values (with appropriate measures of distribution) for patient characteristics including:

- Sex
- Age
- Number and type of previous therapies

- PD-L1 proportion score

The registered population will be used.

6. TREATMENT RECEIVED

The statistician will report:

- The number of patients starting treatment;
- The number of registered patients not starting treatment, plus reasons;
- The median (and range) number of cycles administered;
- The median (and range) time from registration to first treatment;
- The number of patients (%) in which treatment delays occurred, plus reasons;
- The number of patients (%) in which dose reduction occurred, plus reasons;

Treatment will be given every three weeks. A reported administered dose greater than zero will be taken to be a cycle of treatment. The registered population will be used.

7. DEVIATIONS

The statistician will report:

- The number of patients (%) who deviate from protocol and the different types of deviation

8. TOXICITY AND SAFETY ANALYSIS

Adverse events will be recorded according to CTCAE v.4.0.

In addition to the analysis of the toxicity primary outcome, the statistician will report:

- The number (and %) of patients experiencing at least one any grade adverse event and adverse reaction (i.e. considered at least possibly related to pembrolizumab);
- The number (and %) of patients experiencing at least one grade 3-4 adverse event and adverse reaction (i.e. considered at least possibly related to pembrolizumab);
- The incidence of each adverse event and adverse reaction (all grades and grade 3-4) as a per-patient rate;
- The number of deaths considered to be associated with pembrolizumab;
- The number (and %) of patients experiencing at least one serious adverse event (SAE).

Additionally, we may present summaries of adverse events for those that are and are not immune-related.

The analysis population will be used.

Note: Adverse Events experienced within a Serious Adverse Event will be included in the AE and AR reporting.

9. ANALYSIS

9.1 Definition and Calculation of Outcome Measures

9.1.1 Primary Outcomes

The trial design is based on two co-primary outcome measures, toxicity and durable clinical benefit (DCB).

Toxicity

The toxicity co-primary outcome measure for the trial is defined as the occurrence of a treatment-related dose delay or treatment discontinuation due to an adverse event.

DCB

Patients are scheduled to have CT scans every 9 weeks from baseline until disease progression. On each occasion, overall tumour burden will be assessed using RECIST 1.1, according to the study protocol. DCB is defined as the occurrence of CR, PR or SD without prior PD at or after the second scheduled CT scan. The second scan is scheduled to occur at the end of treatment cycle 6 at 18 weeks. For many reasons, scans often do not take place on the desired date. The exact timing of scans is largely out of the control of the trials office. The third scheduled scan is at 27 weeks. In NSCLC, disease status could easily change during the intervening 9 weeks. When calculating DCB, we include scans that occurred at least 16 weeks after registration.

For example, a disease response of (CR / PR / SD) at the first scheduled scan at 9 weeks, followed by a disease response of (CR / PR / SD) at the second scheduled scan at 16+ weeks would constitute DCB. Likewise, (CR / PR / SD) at 9 weeks, missing data at 18 weeks, and (CR / PR / SD) at the third scheduled scan at 27 weeks would also constitute DCB.

Patients for whom the DCB outcome could not be determined will be taken to be non-responders.

9.1.2 Secondary Outcomes

9.1.2.1. Objective Response (OR)

Best overall response is the best response recorded over the whole period of assessment and could be complete response (CR), partial response (PR), stable disease (SD), progressive disease (PD) or inevaluable for response (NA, for which reasons such as early death due to disease or early death due to toxicity will be specified). Objective Response (OR) is the occurrence of CR or PR as best overall response. Objective Response will be based on responses confirmed using the subsequent 9-weekly scan but objective response based on unconfirmed responses will also be reported.

Furthermore, the number (and %) of patients in each best response category will be reported.

9.1.2.2. Health Related Quality of Life (HRQoL)

The purpose of HR QoL measurement is to quantify the degree to which the medical condition or its treatment impacts the individual's life in a valid and reproducible way. Health-related quality-of-life will be measured using the FACT-L and EQ-5D questionnaires, and a patient-generated subjective global assessment questionnaire, as identified in the protocol (see protocol appendix for questionnaires). The FACT-L questionnaire generates 5 measures for analysis: physical well-being, social/family well-being, emotional well-being, functional well-being and the lung cancer subscale. The EQ5D questionnaire generates 2 measures for analysis: an EQ5D utility measure and an EQ5D Visual Analogue Scale. Questionnaires will be administered on day 1 of every cycle prior to receiving treatment and also at the end of treatment visit.

9.1.2.3. Time to Progression (TTP)

This is defined as the time from commencement of trial treatment to the date of CT scan when progressive disease is first recorded. Patients with no recorded progression at the time of analysis or who die without recorded progression will be censored at the date of the CT scan when they were last recorded with an evaluable measure that was not progression.

9.1.2.4. Progression-Free Survival Time (PFS)

This is defined as the time from commencement of trial treatment to the date of CT scan when progressive disease first recorded or date of death without previously recorded progression. Patients who are alive with no recorded progression at the time of analysis will be censored at the date of the CT scan when they were last recorded with an evaluable measure that was not progression.

9.1.2.5. Overall Survival Time (OS)

This is defined as the time from commencement of trial treatment to the date of death. Patients who are alive at the time of analysis will be censored at the date last confirmed alive.

9.1.2.6. Duration of Objective Response (DOR) and Duration of Stable Disease (DSD)

Duration of objective response and duration of stable disease are defined as the time from commencement of trial treatment to the date of the subsequent CT scan when progressive disease is first confirmed or date of death without previously recorded progression. This outcome is calculated and reported separately for patients who achieve an OR or SD. Patients experiencing OR or SD who are alive with no recorded progression at the time of analysis will be censored at the date of the CT scan when they were last recorded with an evaluable measure that was not progression.

These outcomes are effectively a subgroup analysis of PFS using categories that reflect those patients whose best response is SD (DSD) and those patients who achieve an OR (DOR). Patients may belong to one analysis but not both. Patients who record no RECIST data or whose best response is PD will not be used in this analysis.

9.2 Methods of Analysis

9.2.1 Stratification variables

Each patient will have a PD-L1 proportion score. Where a range of PD-L1 proportion scores is specified for a tumour sample (e.g. 10-20%), the PD-L1 proportion score for that patient will be taken to be the mid-point (e.g. 15%).

9.2.2 Primary Outcomes

The co-primary outcomes will be summarised as toxicity rate and disease control (frequently referred to simply as *efficacy* in this section) rate and analysed simultaneously using the BEBOP (Bayesian Evaluation of Binary Outcomes with Predictive variables) method, developed by Brock, *et al*. (publication in draft).

Each patient will have a PD-L1 proportion score and this will determine membership to one of three PD-L1 categories, shown in Table 1. These categories were validated to be predictive of response in Garon, et al [1]. Additionally, each patient will be either previously treated or not previously treated. These two variables yield the six PePS2 cohorts shown in Table 2.

Table 1 – A patient's PD-L1 category is inferred from their PD-L1 proportion score

Criteria on PD-L1 proportion score z	PD-L1 category
$z < 1\%$	Low
$1\% \leq z < 50\%$	Medium
$z \geq 50\%$	High

Table 2 – PePS2 cohorts and corresponding vectors of predictive values, x_i = (x_{1i}, x_{2i}, x_{3i})

Cohort	Treatment status	PD-L1 category	x_{1i}	x_{2i}	x_{3i}
1	Not pre-treated	Low	0	1	0
2	Not pre-treated	Medium	0	0	1
3	Not pre-treated	High	0	0	0
4	Pre-treated	Low	1	1	0
5	Pre-treated	Medium	1	0	1
6	Pre-treated	High	1	0	0

Let $\theta = (\alpha, \beta, \gamma, \zeta, \lambda, \psi)$ be a vector of parameters. We model the marginal probability of efficacy in patients with the predictive values $\boldsymbol{x}$ = (x_1, x_2, x_3) using the logit-model

$$\pi_E(\boldsymbol{x}, \theta) = \alpha + \beta x_1 + \gamma x_2 + \zeta x_3$$

and the marginal probability of toxicity using the logit-model

$$\pi_T(\boldsymbol{x}, \theta) = \lambda$$

Using these formulae, the probability of efficacy is different for each cohort. In contrast, the probability of toxicity is uniform across the cohorts. These assumptions are supported by data published on similar studies in performance status 0 & 1 patients [1, 2].

Let *a*=1 for a given patient if they experience efficacy, and *b*=1 if they experience toxicity, else 0. The joint probability density of efficacy and toxicity events for this patient is modelled using the formula

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

The fraction on the right-hand side, which ranges in value over (-1, 1) for $\psi \in (\infty, -\infty)$, models the association between efficacy and toxicity events.

Let $\boldsymbol{X} = \{(\boldsymbol{x}_1, a_1, b_1), \ldots, (\boldsymbol{x}_n, a_n, b_n)\}$ be the trial outcomes for n patients. The aggregate likelihood function is

$$L(\boldsymbol{X}, \theta) = \prod_{i=1}^{n} \pi_{a_i, b_i}(\boldsymbol{x}_i, \theta)$$

With prior f(θ), the posterior distribution, up to proportionality, is

$$f(\theta | X) \propto f(\theta)\ L(\boldsymbol{X}, \theta)$$

We use normal prior distributions for the elements of θ with means and variances given in Table 3. These give prior events rates of approximately 20% for efficacy and toxicity in each cohort. This can be verified by setting $\boldsymbol{X} = \{\}$ and estimating the 'posterior' probabilities of efficacy and toxicity. The prior event rates we use represent conservative extrapolations of those published in performance

status 0 & 1 patients [1, 2], where, if we were to anticipate a difference compared to our performance status 2 population, the stronger PS0/PS1 patients would be more likely to achieve response and less likely to experience toxicity.

Table 3 – Parameters for normal prior distributions for the elements of θ

Parameter	Mean	Variance
α	-2.2	4
β	-0.5	4
γ	-0.5	4
ζ	-0.5	4
λ	-2.2	4
ψ	0	1

These priors generate prior predictive efficacy and toxicity outcomes shown in Figure 1. We see that modest event rates for each outcome in each cohort are anticipated, but that high event rates are possible in each case. In this regard, our priors can be regarded as having a regularising effect. The prior expected event rates with credible intervals are shown in Table 4. The expected event rates are approximately 20%, although materially lower and greater rates are facilitated in each cohort.

Table 4 - Credible intervals for events rates drawn from the prior predictive distribution of the regularising priors in Table 3. Eff and Tox show the probability of efficacy and toxicity, respectively. Lowercase l and u show the central 50% credible interval and upper-case L and U show the central 90% credible interval.

Previous	PD-L1	EffL	Effl	Eff	Effu	EffU
TN	Low	0.00	0.01	0.21	0.31	0..87
TN	Med	0.00	0.01	0.21	0.31	0.87
TN	High	0.00	0.03	0.20	0.30	0.75
PT	Low	0.00	0.00	0.21	0.30	0.92
PT	Med	0.00	0.00	0.21	0.30	0.92
PT	High	0.00	0.01	0.21	0.32	0.87
		ToxL	**Toxl**	**Tox**	**Toxu**	**ToxU**
TN / PT	Low-High	0.00	0.03	0.20	0.30	0.75

In a manuscript presently under review, we compared the operating performance under these priors to diffuse priors (i.e. greater prior variances) and informative priors (i.e. anticipating increasing efficacy in PD-L1). We demonstrated that informative priors coerced undesirable bias, and diffuse priors generate implausible prior predictive outcome rates with the majority of probability mass close to 0 or 1, offer poor posterior coverage and yield large empiric standard errors. We concluded that our regularising priors offer an attractive balance.

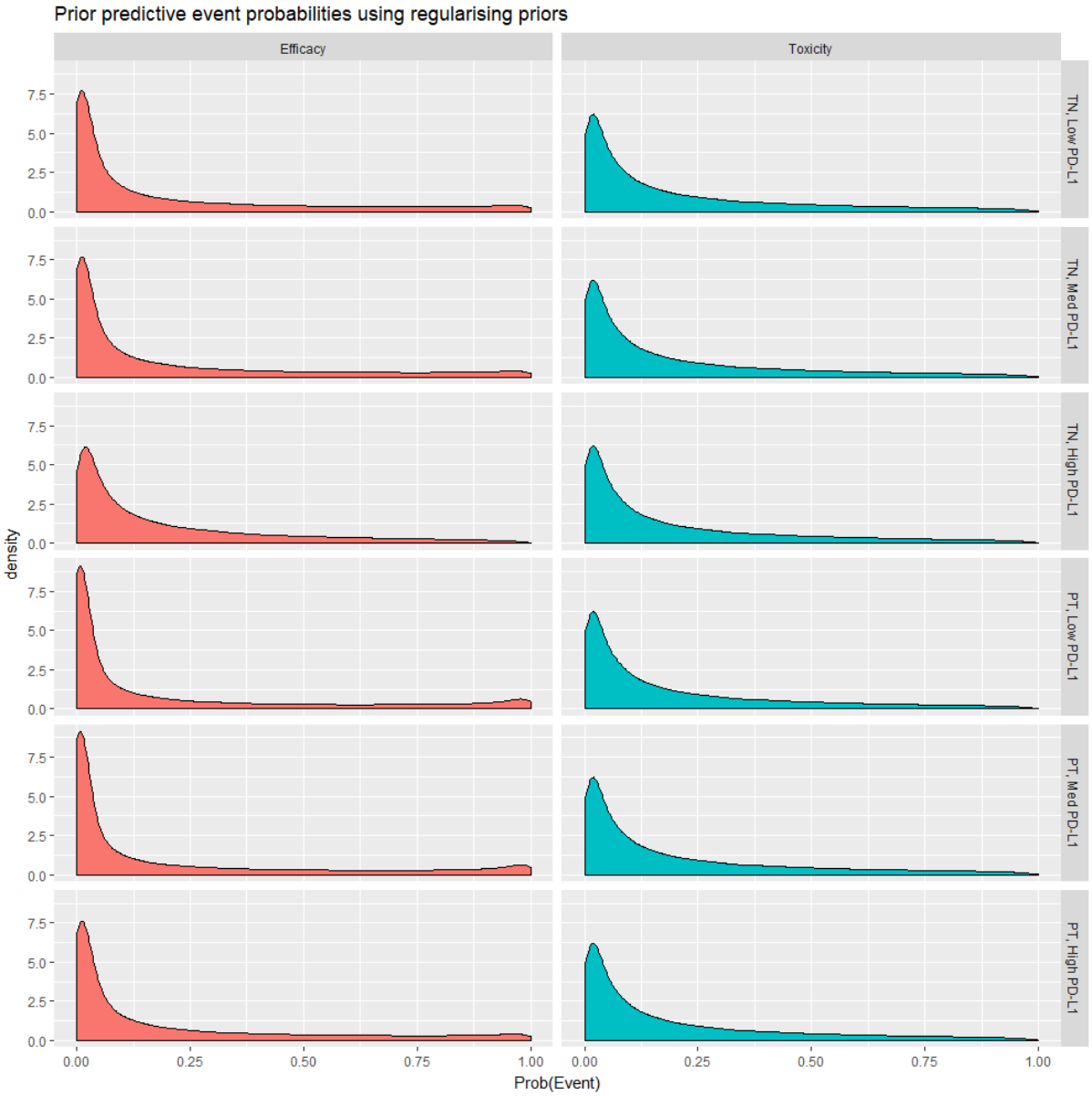

Figure 1 - Prior predictive distributions of the probabilities of efficacy and toxicity in all cohorts.

Given trial data $\boldsymbol{X}$, inferences on θ and functions of θ, such as $\pi_E(\boldsymbol{x},\theta)$ and $\pi_T(\boldsymbol{x},\theta)$, are made using the posterior predictive distributions.

If a patient does not report any adverse event that satisfies the criteria to be a toxicity event, that patient will be treated as not having had toxicity. If a patient does not report CT scans that satisfy the criteria to be a DCB event, that patient will be treated as not having had efficacy.

9.2.2.1. Missing data

The method described above requires covariate data be present. If a modest amount of covariate data is missing, we propose to:

1) Impute each possible combination of the missing covariates; this will yield *1,…, M* distinct complete datasets, each containing a small amount of imputed information;
2) Fit the model described above to each complete dataset to yield *M* sets of posterior samples;

3) Re-sample from the M sets of posterior samples with probabilities $p_1, \ldots, p_M$ determined by the likelihood of the imputed covariate vector, implied by the frequency of covariates in the fully-specified cases.

An example will be expository. Let us assume that two patients have missing PD-L1 status. We know that the PD-L1 category for each may be Low, Medium, or High, yielding $3^2 = 9$ possible covariate imputations. We supplement the observed covariate data with the imputed data to generate 9 complete sets. The efficacy and toxicity outcomes are known in each instance. Fitting the model to each complete dataset, we obtain 9 sets of posterior samples of the estimated probabilities of efficacy and toxicity in the six cohorts. Let us also assume that the incidence of Low / Medium / High PD-L1 in the set of patients that report PD-L1 categories is 10% / 30% / 60%. We attach scenario likelihood 0.1 * 0.1 = 0.01 to the imputation Low / Low; and likelihood 0.3 * 0.6 = 0.18 to the imputation Medium / High; etc. To obtain pooled posterior samples from the 9 distinct sets of posterior samples, we select from the datasets with probabilities implied by the scenario likelihoods, and then select with replacement and uniform probability from the posterior samples for that scenario.

We stress that this method has been proposed to impute required yet missing covariate data only. It will not be used to impute outcomes.

This method will allow us to use the intended analysis method to study the outcomes of all evaluable patients, irrespective the presence of covariate information.

9.2.2.2. Supporting graphical analyses

To support the analyses of the primary outcomes, the following graphical methods may be presented additionally:

1. A “waterfall plot”, showing the best percentage change from baseline in the sum of longest diameters of the target lesions, presented as a bar chart. The bars may be coloured to show cohort membership. This analysis may use the subset of patients in the analysis population with baseline tumour size measurement and at least one post-baseline tumour size measurement.
2. A “spider plot”, showing the sum of longest diameters of the target lesions through time, presented as a line chart. Mechanisms will be used to show cohort membership. This analysis will use the subset of patients in the analysis population with baseline tumour size measurement and at least one post-baseline tumour size measurement.
3. A “swimmer plot”, showing time on treatment and the chronological occurrence of responses and progressions, and the durability of responses. This analysis will use the analysis population.

9.2.3 Secondary Outcomes

9.2.3.1. Time-to-event outcomes

Time-to-event outcomes (TTP, PFS, OS, DOR, DSD) will be analysed using the Kaplan-Meier method. Median time-to-event (and 90% CIs) will be presented for each outcome. Milestone events rates will be presented (with 90% CIs) for each outcome. Milestone rates will be presented at appropriate times, selecting from 3, 6, 12, 18 and 24 months. Note that 90% confidence intervals are appropriate in this single arm phase II trial. These analyses will use the analysis population; the DoR analysis will use the subset that achieve partial or complete RECIST response.

9.2.3.2. HRQoL outcome

Descriptive longitudinal analyses of each quality of life score will be presented. If appropriate, the EQ5D utility score may be combined with the survival data to estimate quality-adjusted life years using the integrated quality-survival product (Billingham and Abrams 2002). This analysis will use the subset of the analysis population that provide any HRQoL data.

9.3 Sample Size Calculations and Operating Characteristics

PePS2 will recruit 60 patients.

9.3.1 Bryant & Day

Being a phase II trial with co-primary efficacy and toxicity outcomes, the original sample size was calculated using a Bryant and Day design. For the toxicity outcome, a cut-off of 10% was selected as the rate below which toxicity would be acceptable and a cut-off of 30% was selected as the rate above which toxicity would be considered unacceptable. A significance level of 5% was selected as the acceptable chance of approving an unacceptably toxic drug as worthy of further investigation.

It was always our intention to treat all PS2 NSCLC patients in PePS2, irrespective of PD-L1 status. However, we knew from [1] that PD-L1 score would affect the efficacy rate. For the purposes of sample size estimation, we estimated that 70% of patients would be PD-L1+ and that, in these patients, an efficacy rate of 30% would be unattractive and a 50% rate would be attractive. In the remaining 30% of PD-L1- patients, we considered that an efficacy rate of 10% would be unattractive and a 20% rate would be attractive. Weighting these efficacy thresholds together yields population thresholds of 24% and 41%. Using a significance level of 10% for the efficacy outcome as the chance of accepting an inactive drug as worthy of further investigation, and overall power of 90%, a Bryant and Day design requires 60 patients and would approve the treatment in the population if no more than 12 toxicities and at least 19 efficacies are observed.

The major flaw in using Bryant & Day in this way is that power is compromised when the prevalence of PD-L1+ is not 70%. In an all-comers trial, a design that uses the predictive information in PD-L1 score without assuming a prevalence would be preferable and to these ends, we developed BEBOP.

9.3.2 BEBOP[1]

The sample size of 60 patients was provisionally approved. We estimated the statistical operating performance in a broad simulation study using 12 scenarios and 60 patients.

In simulated trial iteration *j*, cohort membership probabilities ρ_j were randomly sampled from a Dirichlet distribution with parameter vector $\hat{\rho} = (15.7, 21.8, 12.4, 20.7, 18.0, 11.4)$. $\hat{\rho}$ was calculated by splicing together the prevalence of (low, medium, high) PD-L1 scores observed in [1] in previously-untreated patients (0.315, 0.436, 0.249) and previously-treated patients (0.414, 0.359, 0.227), scaling each by 50% to reflect that we expect equal numbers of previously-untreated and –treated patients. Given a draw from this Dirichlet distribution, cohort memberships for iteration *j* were randomly sampled using a multinomial distribution with probabilities ρ_j. Summary statistics for the simulated cohort prevalences and cohort sizes are given in Table 5.

Table 5 – Simulated cohort prevalences and sizes, based on 100,000 random draws using the described method.

	ρ_j	Number of patients
Cohort	Mean (95% CI)	Mean (95% CI)
1	15.7% (9.3%, 23.4%)	9.4 (3, 17)

[1] BEBOP is the name originally given to this method by the authors at CRCTU. A referee pointed out that the design was a specialisation of the dose-finding design by Thall, Nguyen & Estey (TNE) (2008). Externally, the name P2TNE was adopted to reflect the heritage of the design, and the focus on phase-2. For simplicity, the name BEBOP is retained in internal documents,

2	21.8% (14.3%, 30.4%)	13.1 (6, 17)
3	12.4% (6.7%, 19.5%)	7.4 (2, 14)
4	20.7% (13.4%, 29.1%)	12.4 (5, 21)
5	18.0% (11.1%, 26.1%)	10.8 (4, 19)
6	11.4% (6.0%, 18.2%)	6.8 (2, 14)

To give measure to the benefit of information sharing in BEBOP, we also consider beta-binomial Bayesian conjugate models that assess the treatment in cohorts individually. With prior $\pi \sim Beta(\alpha, \beta)$, the posterior beliefs are $\pi \mid q, m \sim Beta(\alpha + q, \beta + m - q)$ where m is the number of patients in a cohort and q is the number of events observed. Inferences are made on the posterior distribution. A Beta(0.4, 1.6) prior on the rates of efficacy and toxicity gives a prior mean event probability of 20% in each case, with a 95% credible interval of (0.0%, 80.5%). This prior is modestly informative, having an effective sample size of 2 patients.

For the purposes of measuring operating performance, using BEBOP and the beta-binomial models, we accepted the treatment at the end of each simulated trial if

$$\Pr(\pi_E(\boldsymbol{x}_i, \theta) > 0.1 \mid \boldsymbol{X}) > 0.7$$

and

$$\Pr(\pi_T(\boldsymbol{x}_i, \theta) < 0.3 \mid \boldsymbol{X}) > 0.9$$

Cohorts were evaluated individually. Table 6 shows operating characteristics of the described BEBOP design in 12 scenarios. The thresholds above were selected so that our BEBOP model under our regularising priors would approve with at least 80% probability in all cohorts in Scenario 1, and approve in Scenario 3 with no more than 5% probability.

In scenarios 1 to 4, the rates of efficacy and toxicity are uniform across the cohorts. Scenario 1 shows that if the true probability of efficacy is 30% and toxicity is 10%, we can expect BEBOP to approve the treatment with at least 85% probability in all cohorts. The cohorts have different approval probabilities because the average cohort sizes are different. A key benefit of the BEBOP design is the apportioning and sharing of information across cohorts via the Bayesian regression model. For instance, BEBOP will quite reliably approve the treatment in scenario 1 in cohorts 3 and 6, even though they each only receive approximately 7 patients who experience 2 efficacies. The high efficacy rate observed in other cohorts informs BEBOP. A statistical model that does not share information would not perform so well. The beta-binomial model would approve the treatment in cohort 3 in scenario 1 with probability only 50.5%. Using the described decision criteria with a cohort size of n=7, the beta-binomial model must observe at least q_E=4 efficacy events to conclude that the treatment is efficacious. In contrast, BEBOP reliably approves the treatment in cohort 3 despite only observing 2.2 efficacy events in 7.4 patients, on average. BEBOP manages this because it observes 15.6 efficacy events in the residual 52.6 patients in the other five cohorts, on average. The high efficacy rate seen in the other cohorts helps BEBOP to approve the treatment in cohort 3 when it is truly effective, despite observing relatively few responses.

Table 6 - Operating characteristics of the BEBOP model used in PePS2 over 12 simulated scenarios, reproduced from Brock et al [4]. Pr(Eff) and Pr(Tox) 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; Beta-Bin the probability it is approved by a beta-binomial model. 10,000 iterations were used in each scenario.

Scenario	Cohort	Pr(Eff)	Pr(Tox)	Odds	N	Eff	Tox	BEBOP	Beta-Bin
1	1	0.3	0.1	1.0	9.4	2.8	0.9	0.882	0.569
	2	0.3	0.1	1.0	13.1	3.9	1.3	0.903	0.679
	3	0.3	0.1	1.0	7.4	2.2	0.7	0.883	0.505
	4	0.3	0.1	1.0	12.4	3.7	1.3	0.893	0.658
	5	0.3	0.1	1.0	10.8	3.2	1.1	0.890	0.611
	6	0.3	0.1	1.0	6.8	2.0	0.7	0.872	0.489
2	1	0.3	0.3	1.0	9.5	2.8	2.8	0.178	0.123
	2	0.3	0.3	1.0	13.1	4.0	3.9	0.182	0.132
	3	0.3	0.3	1.0	7.5	2.2	2.2	0.178	0.132
	4	0.3	0.3	1.0	12.4	3.7	3.7	0.180	0.130
	5	0.3	0.3	1.0	10.8	3.2	3.2	0.180	0.130
	6	0.3	0.3	1.0	6.8	2.1	2.0	0.175	0.134
3	1	0.1	0.3	1.0	9.5	1.0	2.9	0.037	0.044
	2	0.1	0.3	1.0	13.0	1.3	3.9	0.036	0.036
	3	0.1	0.3	1.0	7.4	0.7	2.2	0.038	0.047
	4	0.1	0.3	1.0	12.4	1.2	3.7	0.030	0.038
	5	0.1	0.3	1.0	10.8	1.1	3.2	0.028	0.038
	6	0.1	0.3	1.0	6.9	0.7	2.0	0.033	0.047
4	1	0.3	0.1	0.2	9.4	2.8	0.9	0.885	0.588
	2	0.3	0.1	0.2	13.1	3.9	1.3	0.904	0.689
	3	0.3	0.1	0.2	7.5	2.3	0.7	0.893	0.524
	4	0.3	0.1	0.2	12.4	3.7	1.2	0.894	0.666
	5	0.3	0.1	0.2	10.9	3.3	1.1	0.894	0.632
	6	0.3	0.1	0.2	6.8	2.0	0.7	0.872	0.504
5	1	0.10	0.1	1.0	9.4	0.9	1.0	0.234	0.173
	2	0.15	0.1	1.0	13.1	2.0	1.3	0.460	0.366
	3	0.30	0.1	1.0	7.4	2.2	0.7	0.800	0.501
	4	0.10	0.1	1.0	12.5	1.2	1.3	0.232	0.194
	5	0.15	0.1	1.0	10.8	1.6	1.1	0.445	0.328
	6	0.30	0.1	1.0	6.8	2.0	0.7	0.758	0.478
6	1	0.10	0.3	1.0	9.5	0.9	2.8	0.042	0.040
	2	0.15	0.3	1.0	13.1	1.9	3.9	0.079	0.068
	3	0.30	0.3	1.0	7.4	2.2	2.2	0.133	0.134
	4	0.10	0.3	1.0	12.4	1.2	3.7	0.043	0.040
	5	0.15	0.3	1.0	10.8	1.6	3.3	0.079	0.064
	6	0.30	0.3	1.0	6.9	2.1	2.1	0.130	0.134
7	1	0.125	0.1	1.0	9.5	1.2	1.0	0.340	0.237
	2	0.185	0.1	1.0	13.1	2.4	1.3	0.607	0.466
	3	0.355	0.1	1.0	7.4	2.6	0.7	0.886	0.554
	4	0.100	0.1	1.0	12.4	1.2	1.2	0.248	0.197
	5	0.150	0.1	1.0	10.7	1.6	1.1	0.459	0.320
	6	0.300	0.1	1.0	6.9	2.1	0.7	0.776	0.489
8	1	0.167	0.1	1.0	9.5	1.6	0.9	0.457	0.337
	2	0.192	0.1	1.0	13.1	2.5	1.3	0.681	0.478
	3	0.500	0.1	1.0	7.4	3.7	0.7	0.979	0.630
	4	0.091	0.1	1.0	12.4	1.1	1.3	0.299	0.173
	5	0.156	0.1	1.0	10.8	1.7	1.1	0.493	0.340
	6	0.439	0.1	1.0	6.8	3.0	0.7	0.924	0.592
9	1	0.267	0.1	1.0	9.4	2.5	1.0	0.814	0.526
	2	0.292	0.1	1.0	13.0	3.8	1.3	0.921	0.681
	3	0.600	0.1	1.0	7.5	4.5	0.7	0.994	0.654
	4	0.191	0.1	1.0	12.4	2.4	1.3	0.737	0.471
	5	0.256	0.1	1.0	10.8	2.7	1.1	0.841	0.553
	6	0.539	0.1	1.0	6.9	3.7	0.7	0.981	0.621
10	1	0.267	0.3	1.0	9.4	2.5	2.8	0.162	0.124
	2	0.292	0.3	1.0	13.0	3.8	3.9	0.181	0.125
	3	0.600	0.3	1.0	7.5	4.5	2.2	0.195	0.175
	4	0.191	0.3	1.0	12.4	2.4	3.7	0.150	0.086
	5	0.256	0.3	1.0	10.8	2.8	3.3	0.170	0.112
	6	0.539	0.3	1.0	6.8	3.7	2.1	0.193	0.181
11	1	0.267	0.1	0.2	9.4	2.5	0.9	0.816	0.551
	2	0.292	0.1	0.2	13.2	3.8	1.4	0.919	0.683
	3	0.600	0.1	0.2	7.5	4.5	0.7	0.994	0.665
	4	0.191	0.1	0.2	12.4	2.4	1.3	0.738	0.479
	5	0.256	0.1	0.2	10.7	2.7	1.1	0.836	0.564
	6	0.539	0.1	0.2	6.8	3.7	0.7	0.983	0.629
12	1	0.267	0.1	5.0	9.4	2.5	0.9	0.813	0.495
	2	0.292	0.1	5.0	13.1	3.8	1.3	0.923	0.667
	3	0.600	0.1	5.0	7.5	4.5	0.7	0.993	0.650
	4	0.191	0.1	5.0	12.4	2.4	1.3	0.736	0.438
	5	0.256	0.1	5.0	10.8	2.7	1.1	0.837	0.537
	6	0.539	0.1	5.0	6.8	3.7	0.7	0.981	0.612

Scenario 2 shows that the design is materially less likely to approve the treatment when the toxicity rate is as high as 30%, despite the high efficacy rate. In this scenario, efficacy is high but toxicity is on the cusp of being acceptable. The design is very likely to reject the treatment when toxicity is as high as 30% and efficacy is as low as 10%, as seen in scenario 3. Despite leveraging information to approve the treatment with small cohort sizes when performance is good, BEBOP does not show a predisposition to approve the treatment when performance is poor. BEBOP is more likely than the cohort-specific beta-binomial models to reject the treatment in scenario 3 because it uses information from all 60 patients to estimate the toxicity rate.

Scenario 4 shows performance when efficacy events are highly negatively associated with toxicity. Here, the ability of patients to achieve efficacious outcomes are strongly hindered if they experience a toxicity event. In every other regard, the parameterisation of scenario 4 is the same as scenario 1. The performance of both methods are actually fractionally better in every cohort when events are associated.

Scenario 5 and 6 show performance when the efficacy probability improves with the rate of PD-L1 expression and pre-treatment status is uninformative (see Table 7). In scenario 5, efficacy is relatively poor in cohorts 1, 2, 4 and 5 but attractive in cohorts 3 and 6, and toxicity is low throughout. BEBOP approves the treatment in cohorts 3 and 6 with 75-80% probability, once again materially better than the beta-binomial models. In scenario 6, the odds of toxicity are much higher but the efficacy odds are the same. Both designs are now much less likely to approve the treatment.

Scenario 7 uses piecewise parallel efficacy curves, as shown in charts in Table 7. Translated to the probability scale, TN patients are 3-6% more likely to experience an efficacious outcome. Treatment is tolerable throughout. Once again, BEBOP is likely to correctly approve treatment in cohorts 3 and 6.

Other than the efficacy probabilities in cohorts 1-3, this scenario is otherwise the same as scenario 5. With the slight increase in efficacy probability, we expect the design to approve more often in cohort 3. It is pleasing, however, to see that the approval probability has slightly increased in cohort 6 too, even though it is unchanged. This is because the design has increased its expectations on the rate of efficacy in all high PD-L1 patients, based on the outcomes in the TN cohort.

Scenario 8 uses the efficacy rates in each cohort that [1] observed in their study, and a uniform toxicity probability of 10%. The notable aspect of this scenario is that there is an apparent interaction yielded by simultaneous low PD-L1 and pre-treated status so that the PD-L1-efficacy curves are not piecewise-parallel, as depicted in Table 7. BEBOP is overwhelmingly likely to approve treatment in cohorts 3 and 6 where efficacy is high. Interestingly, BEBOP is now more likely than not to approve treatment in cohort 2 as well, an improvement over the beta-binomial model of approximately 20%. It manages this, despite an average cohort size of 13.1 patients and efficacy rate only 9.2% above the critical value of 10%, because it leverages the outcomes observed in other cohorts.

In [1], stable disease was the best objective response level in 21.8% of patients. In our study, we have included stable disease in the efficacy event because it demonstrates disease containment, itself an achievement in PS2 NSCLC patients, thus we have reason to expect modestly higher efficacy rates than seen in [1]. That expectation is partly diminished because PS2 patients will be more frail than PS0 and PS1 patients and potentially less likely to achieve response. Scenario 9 uses the efficacy probabilities of Garon, increased in each instance by 10%. The objective here is to analyse how BEBOP performs when efficacy is high. Again, toxicity is left uniform at 10%. The design is now 90% likely to approve the treatment in three cohorts and at least 80% likely to approve in five out of six. In the remaining cohort, with efficacy probability 19.1%, BEBOP is over 70% likely to approve the treatment. This improvement over the approval rate in cohort 2 in scenario 8, despite the similar efficacy probability, comes from the sharing of information.

Scenario 10 is the same as scenario 9 with otherwise high probability of toxicity. Now, BEBOP is unlikely to approve the treatment, demonstrating the value of having simultaneous approval criteria for efficacy and toxicity.

Scenarios 11 and 12 are the same as scenario 9 with otherwise strong negative and positive association between efficacy and toxicity events. We see that the approval probabilities are largely unchanged.

The attractive operating characteristics of BEBOP confirm that 60 is an appropriate sample size in PePS2.

9.4 Additional Analyses

9.4.1 Alternative specifications for efficacy and toxicity models

We have specified a form for π_E that models efficacy curves for the previously-treated and not previously-treated groups that are piece-wise parallel on the log-odds scale, as used in scenario 7 of the simulation study and depicted graphically in Table 7. This specification could be invalidated by the presence of a strong interaction effect between PD-L1 and pre-treatment statuses on the odds of efficacy. This is the focus of simulation scenarios 8-12, where the BEBOP design copes admirably well. Similarly, we have modelled toxicity as constant in probability across the cohorts. If either of these assumptions seems to be invalidated by the outcomes we collect, we would consider a secondary analysis with alternative specifications for the efficacy and / or toxicity models given in Section 9.2.1. The chosen form for the alternative models would reflect the nature of the outcomes observed and the inferences of such would be compared to those of the primary analysis, specifically with regard to differences in the model structures.

9.4.2 Alternative analysis of the efficacy outcomes

Wason and Seaman [7] present the augmented-binary (Aug-Bin) method for analysing change in tumour size as a continuous variable and show their method to be a more efficient than analysing the dichotomised response variable via RECIST in a setting with two post-baseline assessments of disease. Wason and Lin are currently researching the method in a "best response" scenario using an arbitrary number of disease assessments, as used in PePS2. If their method is published and implemented in R, we plan to use it for a secondary analysis of the efficacy outcomes of PePS2. The applicability of their method in PePS2 is challenged by our cohort structure. One feasible approach is to implement their method in each cohort singly. However, we saw the cost of ignoring the cohort structure via the beta-binomial models in our simulation study. An alternative approach is to implement their method separately in the three PD-L1 groups. This would be permissible if the effect of pre-treatment on efficacy is small. A third approach is to implement their method on the whole PePS2 population. Both BEBOP and the Aug-Bin method aim to increase the efficiency of phase II clinical trials. It may be illuminating to compare the different ways in which Aug-Bin can be applied in a biomarker-driven phase II trial, and how the efficiency compares to that of BEBOP, an alternative that retains dichotomised outcomes yet incorporates predictive baseline information. This planned analysis is completely separate to the primary analysis, will be taken for academic interest, and is contingent on the extended Aug-Bin method being published and programmed in R.

9.4.3 Relationship of PD-L1 proportion score with primary outcomes

Garon et al [1] presented and validated the PD-L1 categorisation used in Table 1. As much is feasible with our sample size, we will analyse the effect of PD-L1 score on our primary outcomes. Some questions that will motivate exploratory analysis are:

1. Are the Garon PD-L1 categories valid in our PS2 patient population?
2. Are alternative categories suggested by our data?
3. Can the continuous PD-L1 score (rather than its dichotomisation) be used to model probability of efficacy?

The answer to 1) will be revealed as part of planned primary analysis. An answer to 2) using dichotomised outcomes could be revealed by using classification trees and n-fold cross-validation. Further research would be needed to answer 2) from the perspective of survival. Question 3 can be answered using plots and logistic models. For instance, simple plots of PD-L1 score vs best response / PFS / OS; or more complex joint-models of tumour size and survival, adjusted for PD-L1 score could be revealing.

Once again, this analysis is exploratory and supplementary to the planned primary analysis.

9.5 Subgroup Analysis

The primary analysis of the co-primary outcomes adjusts for categorical variables reflecting PD-L1 score and previous treatment, as demonstrated in Table 2.

Summaries of all outcomes may be calculated and presented in the subgroups defined by the stratifying variables listed in Table 2. Analysis methods will mirror those of the main population.

10. RULES OF INFERENCE

On occasion, data required in the analysis will be missing. If data are not available and cannot be obtained by query, the following rules of inference will be used to maximise data coverage.

10.1 Previous Cancer Treatment

In the vwPreviousCancerTreatment view, if the CancerTrtBoolID field is missing and TypeID_txt is provided, the patient will be inferred to have received previous cancer treatment. This would be used

to infer that previous cancer therapy has been received when a specific therapy is identified but confirmation that the patient received previous cancer treatment is omitted.

10.2 Cycle number

Sometimes the numerical cycle number is not specified but it may be inferable from the time-point label (e.g. 'Treatment cycle 6').

10.3 Sum of target lesion diameters and response with respect to target lesions

On the form detailing the RECIST assessments by CT scan, there are fields to report the response category with respect to the target lesions (RTL), and the sum of target lesion longest diameters (STLLD). When RTL is reported and STLLD is not, the missing value may be imputed to be the value closest to baseline consistent with the response categorisation. Effectively, this means that STLLD will be imputed to be: 0 when RTL = CR; 0.7 * baseline, when RTL = PR; 1.2 * baseline when RTL = PD. Likewise, RTL may be imputed when it is missing and STLLD is reported by comparing STLLD to baseline. In any given analysis, the number of imputed values will be reported and justified. If we use this method, we envisage it will be to maximise information on supporting graphical analyses like waterfall and spider plots.

11. STATISTICAL SOFTWARE

We have implementations of our PePS2 BEBOP design written in Python and Stan [3]. The Python implementation solves posterior integrals directly using Monte Carlo integration. Stan is a Bayesian programming language. The Stan implementation uses No U-Turn Sampling to sample elements from the posterior distribution. The two methods agree on the six-parameter model presented. The Python method cannot realistically be used on problems with more than six parameters because of the non-linear increase in difficulty of directly calculating integrals of higher dimension (the *curse of dimensionality*). The Stan method was developed with the prospect of analysing models with more than six parameters.

12. STORAGE AND ARCHIVING

Snapshots of the data used in interim analyses for TSC meetings will be stored beneath:

T:\Trials Work\EDD\PePS2\PePS2Analysis\TSC\

A snapshot of the data used for publications will be stored beneath:

T:\Trials Work\EDD\PePS2\PePS2Analysis\Publications\

A snapshot of the data used in the final analysis will be stored beneath:

T:\Trials Work\EDD\PePS2\PePS2Analysis\EndOfTrial\

13. REFERENCES

1. Garon, E. B., Rizvi, N. a, Hui, R., Leighl, N., Balmanoukian, A. S., Eder, J. P., … KEYNOTE-001 Investigators. (2015). Pembrolizumab for the treatment of non-small-cell lung cancer. *The New England Journal of Medicine*, *372*(21), 2018–28. http://doi.org/10.1056/NEJMoa1501824
2. Herbst, R. S., Baas, P., Kim, D. W., Felip, E., Pérez-Gracia, J. L., Han, J. Y., … Garon, E. B. (2016). Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): A randomised controlled trial. *The Lancet*, *387*(10027), 1540–1550. http://doi.org/10.1016/S0140-6736(15)01281-7

3. Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2016). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *VV*(Ii).
4. Brock, K. Yap, C. Llewellyn, L. Smith, H. McNab, G. Middleton, G & Billingham, L. A Design for Phase II Clinical Trials in Stratified Medicine with Efficacy and Toxicity Outcomes and Predictive Variables. (forthcoming publication)
5. Thall, P. F., & Cook, J. D. (2004). Dose-Finding Based on Efficacy-Toxicity Trade-Offs. *Biometrics*, *60*(3), 684–693.
6. Herrick, R., Norris, C., Cook, J. D., & Venier, J. (n.d.). EffTox. MD Anderson Center. Retrieved from https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=2
7. Wason, J. M. S., & Seaman, S. R. (2013). Using continuous data on tumour measurements to improve inference in phase II cancer studies. *Statistics in Medicine*, *32*(26), 4639–4650. http://doi.org/10.1002/sim.5867

14. APPENDIX

Table 7 - Curves showing PD-L1 cohort vs log-odds of efficacy, for previously-treated and –untreated patients, in simulation scenarios.