BUC & BEARDSLEY
919 EIGHTEENTH STREET, N.W.
SUITE 600
WASHINGTON, D.C. 20006-5503

WRITER'S TELEPHONE
202-736-3600

TELEPHONE
202-736-3600
FACSIMILE
202-736-3608

June 26, 2007

John K. Jenkins, M.D., F.C.C.P.
Director, Office of New Drugs (HFD-20)
Center for Drug Evaluation and Research
Food and Drug Administration
White Oak Bldg. 22, Room 6304
10901 New Hampshire Avenue
Silver Spring, MD 20993

Re: Amendment to
Complaint and Request for
Correction Pursuant to Federal
Data Quality Act
NDA 21-649
Genasense (oblimersen) for advanced melanoma

Dear Dr. Jenkins:

On behalf of our client, Genta Incorporated ("Genta"), Buc & Beardsley submitted a complaint and request for correction under the Federal Data Quality Act ("FDQA") dated May 16, 2007. Genta is the sponsor of the above-referenced NDA. Additional information germane to Genta's FDQA complaint recently became available and is being submitted as an amendment to Genta's complaint and request for correction.

As noted in our May 16 letter, FDA has continued to perpetuate on its website numerous materials that contain FDA's flawed statistical model, the application of that model to Genta's data on PFS, and the erroneous conclusion that Genasense does not improve PFS compared to the control arm in the clinical trial in question.

Attached as Exhibit A hereto is a copy of relevant portions of Genta's June 22, 2007 submission to the EMEA entitled "EMEA H/C/000171 Supplementary Information: Detailed Grounds for Re-examination of the CHMP Negative Opinion." As stated in our May 16 letter, the EMEA's 180-day assessment report relied heavily on FDA's flawed model and erroneous conclusion. Genta's new submission to the EMEA further demonstrate the flaws in FDA's model and conclusion.

In addition, Genta recently learned that an article entitled "Analysis of progression-free survival in oncology trials: Some common statistical issues" appeared in the April/June 2007 issue of Pharmaceutical Statistics (copy attached hereto as Exhibit B).[1] This article contains a

[1] *Pharmaceu. Statist.* 2007: **6**: 99-113.

reference to the webpages that are the subject of Genta's complaint (see footnote 14 and related text in the article). The appearance of this reference in a recent publication further demonstrates that the dissemination of these flawed materials continues as a result of their continuing presence on FDA's website without correction, and therefore continues to harm Genta.

Genta's May 16, 2007 complaint has been posted on the Department of Health and Human Services website[2] and we request that this amendment be promptly posted as well. Genta also restates its request that you consider this complaint, as amended, on an expedited basis in order to limit the damage that is being done to Genta.

Sincerely,

/S/

Nancy L. Buc
Deborah Livornese

cc: Ms. Laurie Lenkel
Sheldon T. Bradshaw, Esq.
Jane A. Axelrad, Esq.

[2] *See,* http://aspe.hhs.gov/infoquality/request&response/32a.pdf.

Exhibit A

June 22, 2007 Submittal to CHMP

Genta

Genta Development Ltd.
12A Upper Berkeley Street
London W1H 7QE

EMEA H/C/000711

Supplementary Information:

Detailed Grounds for Re-examination of the CHMP Negative Opinion

Genasense® (oblimersen sodium)

Table of Contents

June 22, 2007

Supplementary Information for Detailed Grounds for Re-examination of the CHMP Negative Opinion

1. Additional evidence supporting absence of bias in progression-free survival (PFS) analysis in Study GM301

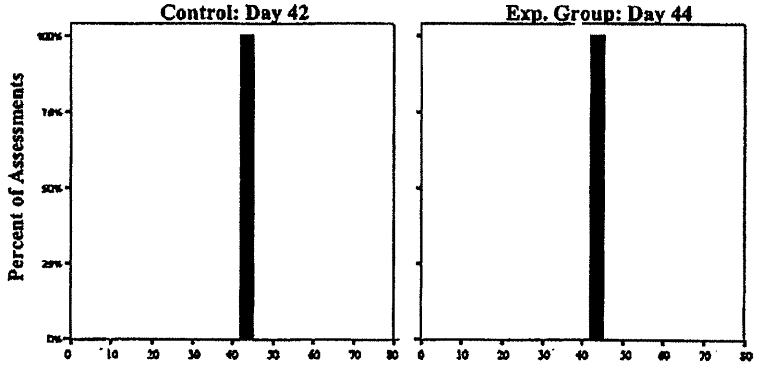
A. Genta simulations of PFS

The FDA simulation that evaluated potential bias in the PFS result was based on the assumption that all assessments in each arm occurred on exactly the same day and that all assessments in the Genasense arm occurred 2 days later (had "a 2-day delay") than all assessments in the DTIC arm (e.g., the first assessment occurred for all patients in the DTIC arm on Day 42 and for all patients in the Genasense arm on Day 44; the second assessment on Day 84 and Day 86, respectively, etc.). Assuming that the distribution of PFS was exponential with a median PFS of 50 days in both treatment groups, it was then demonstrated that this delay would lead to a substantial bias in the comparisons of PFS.

Genta has now repeated those simulations using more realistic assumptions regarding the assessment times, but making similar assumptions regarding the distribution of PFS in order to make a valid comparison. The Genta simulations assume a normal distribution of assessment times with means of 42 and 44 days for the first assessment, means of 84 and 86 days for the second assessment, etc., for the DTIC and Genasense arms, respectively, and a standard deviation of 10 days.

The only difference between the FDA and Genta models is the distribution of assessment timing. In the FDA model, no variation was permitted; in contrast, the Genta model allowed a normal distribution reflective of the actual pattern of assessments observed in Study GM301.

Figure 1: FDA assumption of time to first assessment (model assumptions: all assessments occur on the same day; standard deviation = 0)

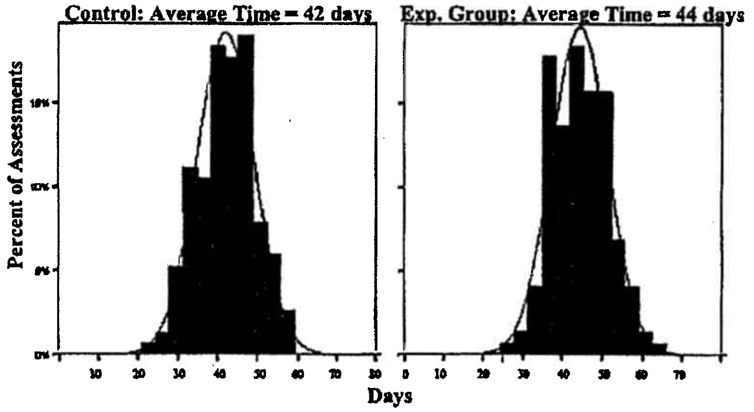


Control: Day 42

Exp. Group: Day 44

Figure 1 shows the distribution of assessment times assumed by the FDA for the first assessment. Similar distributions were assumed for all subsequent assessments.

Figure 2: Genta assumption of time to first assessment (model assumptions: normal distribution; standard deviation = 10 days)



Control: Average Time = 42 days

Exp. Group: Average Time = 44 days

Figure 2 shows the distribution of assessment times for the first assessment that was used as the basis for the Genta simulations. Similar normal distributions were assumed for all subsequent assessments.

Figure 3: Study GM301 - Actual distribution of time to first assessment

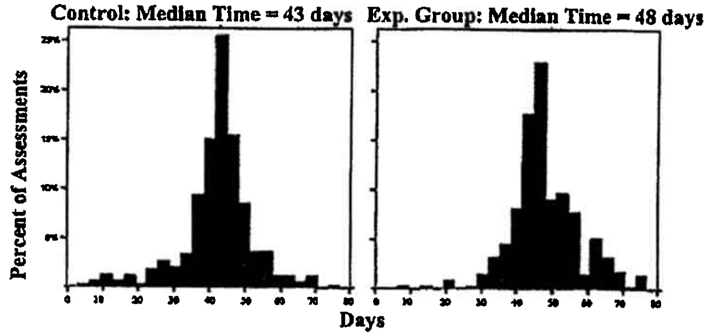


Figure 3 shows the actual distribution of the first assessment times in Study GM301; similar distributional shapes were seen for subsequent assessments. It is apparent from these figures that the Genta simulations are based on a realistic modeling of the actual pattern of assessment timing in Study GM301 (compare Figure 3 with Figure 2). In contrast, the FDA simulations are based on unrealistic assumptions (compare Figure 3 with Figure 1).

Figure 4: Kaplan-Meier Curves for PFS

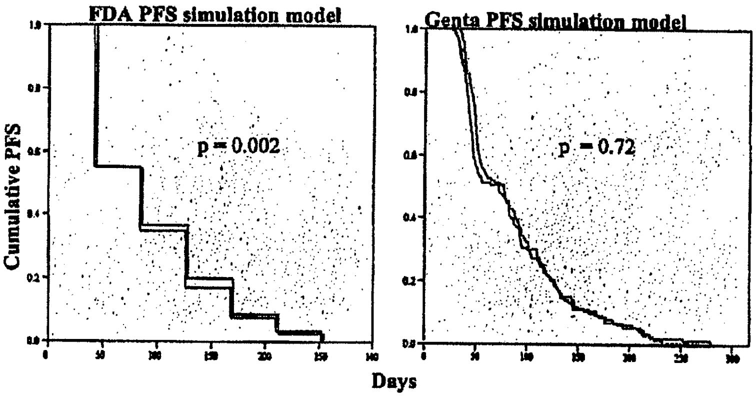


Figure 4 shows two sets of simulated Kaplan-Meier curves for PFS. Note that in both sets of curves, the two arms are virtually superimposable. The left set is the result of a typical simulation using the FDA model for assessment times. Despite the fact that these curves are superimposable, they show a statistically significant difference (p = 0.002). The right set of curves is the result of a typical simulation using the Genta model for assessment times and the curves are not significantly different (p = 0.72). The statistical significance for the left set of curves is a direct result of the unrealistic assumptions and peculiar behavior of the FDA model.

Using the assumptions previously described, the FDA undertook a series of 5000 simulations and showed that the average p-value across these simulations was 0.004. They also showed that 98% of the 5000 simulations would have resulted in a statistically significant p-value and a false conclusion of a treatment difference. They therefore concluded that asymmetry in assessment timing would give a highly significant result and lead to an incorrect conclusion regarding treatment effect.

The Genta series of 5000 simulations had an average p-value of 0.49 with only 5.7% of the simulations resulting in false inference. The conclusions to be drawn from the Genta simulations are exactly the opposite of those based on the FDA simulations. The Genta simulations show there is very little bias when realistic assumptions consistent with the actual assessment times in Study GM301 are used. A total absence of bias would result in an average p-value of 0.50, and only 5% of the simulations would result in a false conclusion. The treatment difference in PFS in Study GM301 ($p = 0.0003$ at 6 months and $p = 0.0007$ at 24 months) cannot possibly be explained by a bias in assessment times. These highly significant results indicate a clear benefit for Genasense in prolonging PFS.

B. PFS sensitivity analyses

Four sensitivity analyses were performed by the Applicant to address concerns regarding potential bias in PFS results due to differences in assessment times between treatment groups. The results are shown in Table 1 below. The hazard ratios from these four sensitivity analyses range from 0.71 to 0.84 and are very consistent with the primary analysis of PFS (hazard ratio = 0.75). Results of all four sensitivity analyses were statistically significant ($p < 0.05$), indicating the robustness of the primary analysis of PFS.

Table 1: Sensitivity analyses to address the potential for bias related to assessment time differences for PFS

Method	Hazard ratio	Log-rank p value
(1) Assumed progression present at scheduled visit when visit occurred late and included censored subjects	0.79	0.003
(2) Used scheduled visit dates instead of actual assessment dates	0.71	< 0.0001
(3) Assumed progression present at scheduled visit when visit occurred late and excluded censored subjects	0.80	0.004
(4) Used nominal cycle number	0.84	0.048

Details of each of the sensitivity analyses follow:

Sensitivity Analysis # 1 - assumed progression was present at the scheduled visit when the visit occurred late and included censored subjects.

During the treatment phase, if PD (or the last non-PD assessment) was determined at a visit that occurred later than scheduled, the date of PD (or censoring) was "brought back" to 42 days after the most recent previous assessment date. During the follow-up phase, if PD (or the last non-PD assessment) was determined at a visit that occurred later than scheduled, the date of PD (or censoring) was "brought back" to 60 days after the most recent previous assessment date.

Sensitivity Analysis # 2 - used scheduled visit dates instead of actual assessment dates.

Beginning with the actual date of first dose, all assessments were assumed to have occurred on the scheduled assessment date (e.g., every 21 days during the treatment phase and every 60 days during the follow-up phase). If PD occurred between the scheduled date of assessments, the date was "taken forward" to the next scheduled visit date.

Sensitivity Analysis # 3 - assumed progression was present at the scheduled visit when the visit occurred late and excluded censored subjects.

This is the same as analysis # 1 except that it includes only subjects with PD. Non-PD assessment times (censored values) were left unchanged.

Sensitivity Analysis # 4 - used the nominal cycle number.

This analysis removes time as a factor. Each determination of PD or censored value uses the cycle number in which it occurred instead of a date.

C. Other exploratory analyses inconsistent with PFS bias

C.1. First assessment events of progression

When counting first assessment outcomes irrespective of time, there are more events of progression in the DTIC arm.

Table 2: Number of patients with progression of disease at first assessment

	Genasense + DTIC	**DTIC**	
	N=323	**N=313**	
Assessment of progression	**n (%)**	**n (%)**	**p value**
Negative	161 (50)	132 (42)	
Positive	162 (50)	181 (58)	0.05

This analysis eliminates the influence of time in the first assessment. If the treatment differences in PFS were entirely due to bias, then the overall numbers of patients progressing at the first assessment would be the same in the two groups. This is not the case: in the DTIC arm, 58% (181/313) of patients progress at the first assessment compared to 50% (162/323) of patients in the Genasense arm.

C.2. Overall survival benefit in patients who progress after the respective median ("better-prognosis patients")

The median time to progression in the DTIC arm was 1.6 months compared to 2.6 months in the Genasense arm based on an intent-to-treat analysis.

Figure 5: Kaplan-Meier survival curves for patients who progress before the median in the DTIC and Genasense groups

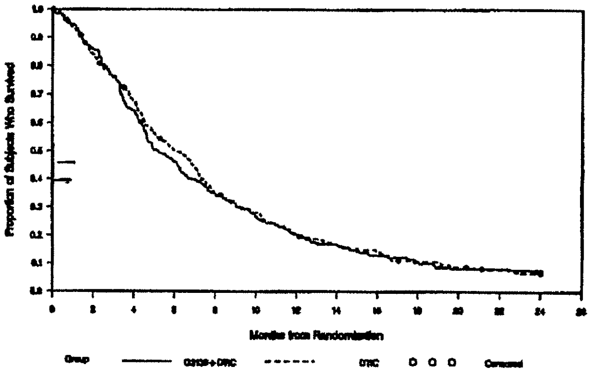


Figure 5 shows the Kaplan-Meier survival curves for patients progressing before the median in the DTIC and Genasense groups. These curves are not significantly different (p = 0.79).

Figure 6: Kaplan-Meier survival curves for patients who progress after the median in the DTIC and Genasense groups

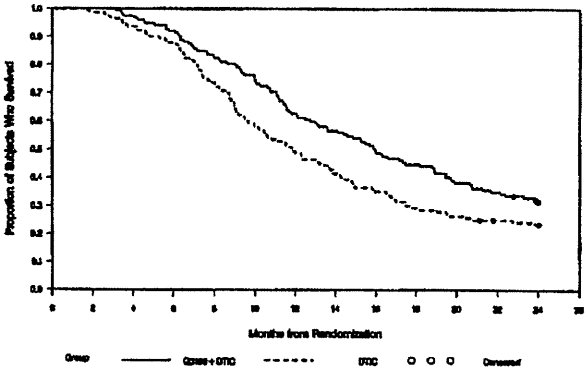


Figure 6 shows the Kaplan-Meier survival curves for patients progressing after the median in the DTIC and Genasense groups. These curves differ significantly (p = 0.017) and indicate that the better-prognosis patients in the Genasense arm survive, on average, 4 months longer than the better-prognosis patients in the DTIC arm. Thus, a 1-month average improvement in time to progression with Genasense translated into a 4-month average improvement in overall survival.

If the benefit for Genasense in terms of PFS were spurious and due to bias caused by asymmetry in assessment times, then the observed improvement of one month for progression would not translate into an overall survival benefit.

C.3. PFS results in normal vs. elevated LDH populations

Figure 7: Progression-free survival by baseline LDH

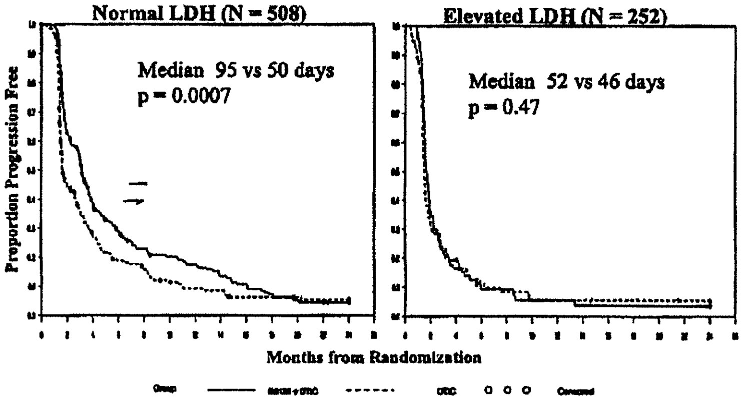


Figure 7 shows the Kaplan-Meier PFS curves for the DTIC and Genasense treatment groups among those patients with normal baseline levels of LDH (LDH ≤ 1.1 * ULN), as well as the corresponding curves for the elevated LDH population (LDH > 1.1 * ULN). The curves for the elevated LDH population are virtually identical; in contrast, a large and statistically significant difference (p=0.0007) is seen for the normal LDH population.

If these differences in PFS were due simply to bias in assessment timing, similar effects should be seen in both subgroups. The fact that different levels of effect occur in the normal and elevated LDH populations does not support the presence of bias.

2. Redacted for FDQA Submittal

3. Redacted for FDQA Submittal

4. Summary (response to detailed grounds for negative opinion)

Statistical significance was achieved for secondary endpoints at the 6-month follow up analysis (p=0.0003 for PFS, p=0.019 for confirmed response and p=0.057 for durable response). In particular PFS was highly significant although overall survival was not. It was accepted at that time that the survival data were simply not mature enough to give any meaningful insight regarding longer-term survival outcome. Any standard methodology for dealing with multiplicity, for example using a Bonferroni adjustment, would still have concluded a highly significant benefit for PFS and strong trends for confirmed response and durable response at this 6-month analysis.

With longer follow up (24-month minimum for all cases as pre-specified in the protocol), the survival data were fully mature; only 14 of 771 cases were lost to follow-up in the 24-month data on survival. The information at this time point is mature, complete and final and overrides that in the incomplete 6-month dataset. An additional analysis with longer follow-up does not raise multiplicity concerns; no operational bias could have occurred between the analyses as study accrual was complete, and treatment was already administered to all patients at the time of the initial analysis. No alternative treatment is known to impact survival. Thus, nothing could have altered the course of the study. The 24-month dataset is the most complete and the most informative. Therefore, it is also the most relevant. This is true for survival as well as for the secondary endpoints. However since the data on PFS and confirmed response were already fully mature at 6-months, little changes with regard to those endpoints; they are presented at 24 months simply for completeness. Durable response continued to fully mature after 6-months and the 24-month result is completely consistent with that seen at 6-months. Thus, considering the secondary endpoints at 6-months or 24-months creates no concrete methodological issue with regard to multiple analyses since the data are essentially the same.

At 24 months, overall survival provides a strong trend in favor of a benefit for Genasense. In fact the 24-month data and the 6-month data are very consistent with a hazard ratio of 0.89 at 6-months and a hazard ratio of 0.87 at 24-months. However, the maturity of the data for patients censored in the 6-month analysis has caused the p-value to become nearly significant at 24 months.

The issue of whether an adequate methodology to deal with multiple analyses was planned has no practical consequence for Study GM301 because of the consistency of all endpoints in terms of treatment effect and the high degree of statistical significance of PFS at both 6-months and 24-months. In addition, the CHMP guideline (CPMP/EWP/908/99) 'Points to Consider on Multiplicity Issues in Clinical Trials' states 'If a multiple test situation occurs which was not foreseen, a conservative approach will be necessary, e.g. Bonferroni or a related procedure'. It is relevant to note that adopting a Bonferroni correction for multiplicity would still result in a high level of statistical significance for PFS at 6-months and 24-months. Assigning a penalty to the results would not change the overall evidence in any way; several endpoints are highly significant and others show a positive trend in favor of Genasense. Finally, all of these endpoints should be regarded as necessary to describe the full treatment effect, and they provide a completely consistent picture.

Despite the fact that Study GM301 was an open-label study, the secondary endpoints are reliable. To reduce the potential for investigator bias, response and progression were determined by computer according to RECIST based on investigator measurements. All complete and partial responses underwent independent and blinded review using 2 different methodologies, which confirmed the veracity of these results. For PFS, a blinded independent review of an approximate 10% random sample was performed. This review confirmed the dates of progression to be correct, with a relatively equal distribution of variability between study arms. Study GM301 is now widely accepted as the reference standard in the literature regarding the activity of DTIC in melanoma.

The initial regulatory review of the strong PFS result was confounded by a flawed analysis performed by FDA reviewers utilizing Monte Carlo simulations. These simulations, which were based on a hypothesis that all patients were seen on the exact same day in each arm and exactly 2 days apart, implied that the highly significant PFS result was due to a bias in the timing of assessments. The sensitivity analysis presented in the MAA, which used a standard methodology, demonstrated that the results are robust. Very recently, a series of similar Monte Carlo simulations that utilized realistic assumptions was presented by the Applicant, and demonstrated the integrity of the PFS result. The potential for bias in these results appears to be very limited and cannot account for the magnitude of the observed PFS result (Genasense + DTIC, 78 days; DTIC, 49 days [p = 0.0007]).

Importantly, Study GM301 elucidated the role of LDH as a prognostic factor that exerts a strong independent effect on survival outcome. Small variations in baseline LDH, even within the normal range, have been shown to behave as a continuous variable interacting with treatment effect and permit the identification of a target population most likely to benefit from treatment. This population of patients with normal baseline LDH can be easily identified with a simple, routine, and widely available blood test. The prognostic relevance of LDH in this population has been separately confirmed using data from an unrelated EORTC study, the results of which were independently analyzed and reported.

The safety evaluation of the Genasense + DTIC combination did not reveal unusual toxicities in this population. The toxicities of interest during review included thrombocytopenia, neutropenia, and elevation in hepatic enzymes. The incidence of grade 3-4 infections and platelet transfusions associated with these cytopenias was not increased, and the hepatic enzyme elevations were reversible and, in many instances, associated with hepatic metastases. Given the paucity of treatments available for patients with advanced or metastatic melanoma and the serious toxicities associated with many of combinations of drugs and biologic agents commonly used in this disease, the risks associated with Genasense should be considered acceptable.

Advanced or metastatic melanoma represents an area of major unmet medical need; treatment options for these patients are very limited. Taking into account the significantly positive effects of Genasense on PFS and overall and durable response, a strong trend toward significance in overall survival, and the acceptable safety profile, the benefit/risk equation leads to a positive result.

Exhibit B

Article from Pharmaceutical Statistics

PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2007; **6**: 99–113
Published online 22 January 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.251

MAIN PAPER

Analysis of progression-free survival in oncology trials: Some common statistical issues

Kevin J. Carroll*,†
AstraZeneca Pharmaceuticals, Global Clinical Information Science, Alderley Park, Macclesfield, UK

With the advent of ever more effective second and third line cancer treatments and the growing use of 'crossover' trial designs in oncology, in which patients switch to the alternate randomized treatment upon disease progression, progression-free survival (PFS) is an increasingly important endpoint in oncologic drug development. However, several concerns exist regarding the use of PFS as a basis to compare treatments. Unlike survival, the exact time of progression is unknown, so progression times might be over-estimated and, consequently, bias may be introduced when comparing treatments. Further, it is not uncommon for randomized therapy to be stopped prior to progression being documented due to toxicity or the initiation of additional anti-cancer therapy; in such cases patients are frequently not followed further for progression and, consequently, are right-censored in the analysis. This article reviews these issues and concludes that concerns relating to the exact timing of progression are generally overstated, with analysis techniques and simple alternative endpoints available to either remove bias entirely or at least provide reassurance via supportive analyses that bias is not present. Further, it is concluded that the regularly recommended manoeuvre to censor PFS time at dropout due to toxicity or upon the initiation of additional anti-cancer therapy is likely to favour the more toxic, less efficacious treatment and so should be avoided whenever possible. Copyright © 2007 John Wiley & Sons, Ltd.

Keywords: *oncology; event count analysis; progression-free survival informative censoring; interval censoring*

INTRODUCTION

Progression-free survival (PFS) is an important endpoint in oncologic drug development. With the

*Correspondence to: Kevin J. Carroll, AstraZeneca Pharmaceuticals, Global Clinical Information Science, Alderley Park, Macclesfield, UK.
†E-mail: kevin.carroll2@astrazeneca.com

advent of a new generation of biologically targeted cytostatic anti-cancer agents, drug developers and researchers can no longer rely on uncontrolled phase II trials and response rate to screen new medicines for clinical utility [1–3]. For drugs designed to stabilize disease, the most sensible phase II approach has been argued to be randomized trials with PFS as the primary endpoint [4,5]. Further, with the advent of ever more effective second and third line cancer treatments and the growing use of 'crossover' designs, in which patients switch to the alternate randomized treatment upon disease progression, detecting an improvement in survival in confirmatory phase III trials has been recognized as an increasingly difficult goal [6–8]. The recently issued EMEA anti-cancer guideline acknowledges these issues and states that either survival or PFS can be used as a primary endpoint in pivotal trials seeking approval for a new drug; when PFS is used and justified as the primary endpoint, survival should be a stated secondary endpoint with follow-up sufficient 'to ensure that there are no relevant negative effects on this [survival] endpoint' [9]. In light of the issues, there has recently been a number of open discussions initiated by the US Food and Drug Administration (FDA) to examine the utility of progression and other measures as endpoints for oncologic drug approval [6–8,10]. In particular, at the Oncologic Drugs Advisory Committee (ODAC) discussion in December 2003 on approval endpoints in non-small cell lung cancer, the vote was 18 'yes', 0 'no' and 1 abstention for the use of PFS as an endpoint to support accelerated approval in the advanced disease setting [6]. Similarly, at the ODAC discussion in May 2004 on approval endpoints in colorectal cancer, the vote was 8 'yes', 5 'no' for the use of PFS as an endpoint to support full approval in the advanced setting [8]. More recently, PFS has been used as the sole basis to provide full drug approval for sorafenib in the treatment of advanced renal cell carcinoma and panitumumab in the treatment of advanced colorectal cancer [11,12].

However, despite such support for the use of PFS as a primary endpoint to support drug approval, key concerns remain regarding the use of PFS to compare treatments for relative effectiveness:

(i) Unlike survival, the exact timing of progression is unknown. Discrete clinic visit schedules for disease assessment means that progressions that occur in between visits are commonly assigned to the visit at which progression was detected, leading to over-estimation of the time to progression [13]. Consequently bias may be introduced in the comparison of treatments as was suggested in the FDA's review of oblimersen sodium [14]. This concern has led to a consensus emerging that clinic visits need to be frequent and identically scheduled between treatments to ensure an accurate determination of the time of progression and a fair comparison of treatments.

(ii) It is not uncommon for randomized therapy to stop (say, due to toxicity) or for additional anti-cancer therapies to be initiated prior to progression being documented. Handling of such patients in the analysis is problematic; recent FDA draft guidelines have suggested that progression time should be censored at the time of the intermediate event [15]. However, this view does not appear to be entirely shared by EU regulatory authorites based on the recently issued appendix to the CHMP's anti-cancer guideline [9,16].

This paper discusses these issues, their practical implications and importance when comparing treatments, and explores if there are ways in which they might be addressed or ameliorated. The remainder of the paper is therefore structured as follows: Section 2 describes a typical oncology trial design with PFS as the primary endpoint. Section 3 examines the practice of assigning the time of progression to the visit at which it was detected and Section 4 examines censoring PFS time on drop-out due to toxicity or the initiation of additional anti-cancer therapy. Section 5 then closes the paper with recommendations for trial design and analysis and a brief discussion of some other, key issues.

DOI: 10.1002/pst

TYPICAL ONCOLOGY TRIAL DESIGN

Suppose two treatments, experimental (E) and control (C), are to be compared in terms of PFS time in a clinical trial powered to detect an underlying hazard ratio, E:C, of size θ, with a 1-sided Type I error rate of α and power $1-\beta$ so that a total of d events are required [17]. Assuming event rates of λ_E and λ_C, uniform accrual over a period of R months and a minimum follow-up period of F months (giving a maximum follow-up of $R+F$ months, which is hereafter referred to as the 'trial follow-up period'), a total of $2N$ patients are to be randomized on a 1:1 basis [18–20]. Assume also that disease status is assessed at regular, scheduled clinic visits, every V months, say. For simplicity, further assume that V is chosen such that $\frac{R}{V}$ and $\frac{F}{V}$ are both integer so that $\frac{F}{V}$ is the minimum and $\frac{R+F}{V}$ the maximum number of scheduled assessments per patient and a clinic visit always takes place at the end of the trial follow-up period.

In advanced disease, PFS time is defined as the interval from randomization to the first of either disease progression or death from any cause. The equivalent measure in adjuvant settings is disease-free survival (DFS), being the interval from randomization to the first of either disease recurrence or death from any cause. The discussion that follows is framed in terms of PFS but can equally be applied to DFS. Since disease is normally assessed at regular, scheduled clinic visits, the exact time of progression is typically unknown. The time of progression is therefore usually assigned to the date of the clinic visit at which it is detected. Patients who are lost to follow-up prior to progression or who reach the end of the trial follow-up period without progression are right censored in the analysis. Patients may also stop randomized therapy during the trial follow-up period prior to reaching a confirmed progression event due toxicity or the addition of further anti-cancer therapy. Such patients are commonly not followed further for progression status, being censored at the time of the event associated with the cessation of randomized therapy.

THE IMPACT OF NOT KNOWING THE EXACT TIMING OF PROGRESSION

When the time of progression is assigned to the visit at which progression was first detected, the extent to which bias is introduced can be gauged directly for exponentially distributed lifetimes using maximum likelihood methods (see Appendix A). If T_i denotes the observed PFS time, event or censored, for the ith patient then

$$\overline{T} = \frac{\sum_{i=1}^{N} T_i}{d} = \frac{1}{\text{observed event rate}}$$

is approximately

$$N\left(\frac{V}{1-\mathrm{e}^{-\lambda V}}, \frac{V^2\mathrm{e}^{-\lambda V}}{d(1-\mathrm{e}^{-\lambda V})^2}\right)$$

or, more conveniently,

$$\ln\left(\overline{T}\right) \sim N\left(\ln\left(\frac{V}{1-\mathrm{e}^{-\lambda V}}\right), \frac{\mathrm{e}^{-\lambda V}}{d}\right)$$

If comparing E to C, then the observed hazard ratio

$$\hat{\theta} = \frac{\text{observed event rate}_E}{\text{observed event rate}_C} = \frac{\overline{T}_C}{\overline{T}_E} \tag{1}$$

is a biased estimate since

$$E[\hat{\theta}] = \frac{V_C(1-\mathrm{e}^{-\lambda_E V_E})}{V_E(1-\mathrm{e}^{-\lambda_C V_C})} \neq \frac{\lambda_E}{\lambda_C} \tag{2}$$

Note that bias is introduced even if visits are scheduled symmetrically between treatments ($V_E = V_C$) and, as one would expect, the degree of bias depends upon the ratio of the interval between visits and the expected PFS time (that is on $\lambda_E V_E$ and $\lambda_C V_C$). The bias in the hazard ratio erodes the power of the standard log rank test to

$$\phi^{-1}\left[(z_\alpha + z_\beta)\omega - z_\alpha\right] \tag{3}$$

where $\omega = \text{abs}\left(\frac{\ln(\hat{\theta})}{\ln(\theta)}\right)$ and $\phi^{-1}(.)$ is the inverse of the cumulative normal distribution; to restore power, the target number of events would have to increase to

$$d' = \frac{d}{\omega^2} \tag{4}$$

(see Appendix B). Table I illustrates the degree of bias that can be introduced by assigning progression time to the clinic visit at which it was detected and the consequences on power.

Table I shows that as the interval between visits lengthens and the number of clinic visits declines, then the hazard ratio is increasingly biased toward the null. Consequently, power falls and, in the examples given, the number of events required to maintain 90% power increases by between 7% and 16% even when visits are as frequent as every month.

Assuming a common visit schedule between treatments, it is of interest to note that to ensure retention of at least $100(1-\gamma)\%$ power $\gamma > \beta$, visits must be scheduled approximately every V' months where

$$V' = (\text{median PFS on } C) \times \frac{2}{\ln(2)}\frac{1}{\theta}\left(\frac{\theta^k - \theta}{1 - \theta^k}\right) \quad (5)$$

and $k = \text{abs}\left(\frac{z_\alpha + z_\gamma}{z_\alpha + z_\beta}\right)$. This result follows since $\frac{V}{1-e^{-\lambda V}} \approx \frac{1}{\lambda} + \frac{V}{2}$ for small λV so that $\hat{\theta} \approx \frac{\theta(1+\tau)}{1+\tau\theta}$, where $\tau = \frac{V\lambda_C}{2}$ (see Appendix B). Table II provides V' for varying θ and median PFS values.

Table II suggests that, for hazard ratios between 0.80 and 0.667, the interval between visits can afford to be no more than about $\frac{1}{2}$ the median PFS time on control to ensure power does not fall below 80%. With a larger hazard ratio of 0.50, the interval between visits can be longer, up to approximately $\frac{2}{3}$ the median PFS time on control.

The Type I error is not inflated providing the scheduling of visits is the same on E and C. However, Table III illustrates the degree to which the Type I error can be increased when clinic visits are asymmetric between treatments.

While it is unlikely that clinic visits would intentionally be scheduled asymmetrically, Table III serves to illustrate the importance in practice of closely matching visit schedules when performing routine log rank analyses of PFS time.

Possible design and analysis strategies when the exact timing of progression is unknown and assigned to the clinic visit at which it was detected

When assigning the time of progression to the visit at which it is detected, bias is introduced and

Table I. Bias and loss of power associated with assigning time of progression to the scheduled clinic visit at which it was detected.

Hazard ratio, θ	Median PFS on E (months)	Median PFS on C (months)	Interval between clinic visits, V, (months)	Expected HR[a], $\hat{\theta}$	Log rank power[b] (%)	Relative increase in d to compensate for loss in power[c]
0.667	6	4	0.5	0.677	87.8	1.07
			1	0.686	85.4	1.16
			2	0.705	80.0	1.34
			4	0.740	67.2	1.81
0.75	8	6	0.5	0.755	88.5	1.05
			1	0.761	86.9	1.11
			2	0.771	83.3	1.23
			4	0.792	75.0	1.51
0.80	12	9.6	0.5	0.803	89.1	1.03
			1	0.806	88.1	1.07
			2	0.811	85.9	1.14
			4	0.822	81.1	1.30

[a] HR = hazard ratio via equation (2).
[b] Log rank power via equation (3); assuming trial originally powered at 90% ($\beta = 0.1$), 2.5% 1-sided α level to detect a HR size θ.
[c] Via equation (4).

Table II. Maximum inter-visit interval length to maintain at least 80% power[a] for varying θ and median PFS values.

Hazard ratio, θ	$\frac{2}{\ln(2)}\frac{1}{\theta}\left(\frac{\theta^k-\theta}{1-\theta^k}\right)$	Median PFS on C (months)	Visits at least every V' months
0.8	0.5058	4	2.0
		6	3.0
		9	4.6
		12	6.1
0.75	0.5219	4	2.1
		6	3.1
		9	4.7
		12	6.3
0.667	0.5520	4	2.2
		6	3.3
		9	5.0
		12	6.6
0.50	0.6315	4	2.5
		6	3.8
		9	5.7
		12	7.6

[a] Assuming trial originally powered at 90% ($\beta=0.1$), 2.5% 1-sided α level to detect a HR size θ.

power is decreased. Some options to address this problem are as follows:

(i) Do nothing. If clinic visits are scheduled symmetrically between treatments, ensure they occur at least every V' months as per equation (5) and accept some loss in power. This is the approach most commonly taken in the analysis of PFS times. Some other approaches that might be adopted are given below.

(ii) Increase the target number of events to $d'=\frac{d}{\omega^2}$. Note that since PFS is a mixture of assumed progression times and known times to death, this increase will be somewhat conservative.

(iii) Since $\hat{\theta}$ is known to be biased, use rather

$$\breve{\theta}=\frac{\ln\left(1-\frac{V}{\overline{T}_E}\right)}{\ln\left(1-\frac{V}{\overline{T}_C}\right)} \tag{6}$$

as the asymptotically unbiased maximum likelihood estimate of the hazard ratio with estimated variance

$$\hat{\mathrm{V}}\mathrm{ar}[\ln(\breve{\theta})]=\frac{V^2}{d_E\overline{T}_E^2\left\{\ln\left(1-\frac{V}{\overline{T}_E}\right)\right\}^2\left(1-\frac{V}{\overline{T}_E}\right)}+\frac{V^2}{d_C\overline{T}_C^2\left\{\ln\left(1-\frac{V}{\overline{T}_C}\right)\right\}^2\left(1-\frac{V}{\overline{T}_C}\right)} \tag{7}$$

(see Appendix A). The lack of bias in this estimate is illustrated by simulation in Table IV.

This approach represents an interval-censored analysis as described by Stone *et al.* [21] and Whitehead [22]. Note that, with SE $\breve{\theta}$ being close to that expected from a log rank analysis of actual PFS times $\left(\sqrt{\frac{4}{200}}=0.1414\right)$, this approach requires little if any increase in trial size. It is also important to note that both $\breve{\theta}$ and SE$\breve{\theta}$ vary little as V increases. This suggests, contrary to common belief, there is little to be gained by the imposition of very frequent clinic visits – when data are analysed on an interval-censored basis, frequent visit scheduling is unnecessary and would serve only to impose an unnecessary burden on patients and investigators alike. Note that while simulations in Table IV are based on exponentially distributed PFS times, distribution-free interval-censored analyses are possible via PROC LIFETEST in SAS [23] and Prentice and Gloeckler provide a method for interval-censored analyses via Cox regression [24]; both approaches require a common visit schedule between treatments.

(iv) If, despite protocol intent, clinic visits are not executed exactly as planned leading to variable spacing between visits and asymmetry in schedules between treatments, PROC LIFEREG can be used to estimate event rates on E and C assuming exponentially distributed PFS times (and alternative distributions), and thus provides an unbiased comparison of treatments. In practice, PFS times will not always follow an exponential distribution making interval-censored analyses in these circumstances difficult. However, Sun *et al.* give a generalized formulation of the log rank test applicable to interval-censored data that provides a score statistic to test equality of survival

DOI: 10.1002/pst

Table III. Inflation in Type I error resulting from asymmetric visit scheduling in a trial with 508 events (sized to detect an assumed hazard ratio of 0.75, 90% power, 2.5% 1 sided α).

Median PFS on E and C (HR = 1)	Interval between visits on C (months)	Interval between visits on E (months)	Expected HR[a], $\hat{\theta}$	Type I error[b] (1-sided)
4	0.5	1	0.959	0.069
	1	2	0.920	0.152
6	1	1.5	0.972	0.050
	1	2	0.945	0.092
	2	3	0.946	0.090
9	1	1.5	0.981	0.040
	2	3	0.963	0.062
	3	4	0.964	0.061
12	1	2	0.973	0.050
	2	3	0.972	0.050
	3	4	0.972	0.050
	4	6	0.946	0.090

[a] HR = hazard ratio via equation (2).

[b] Type I error $= \phi^{-1}\left[-1.96 - \dfrac{\ln\left(\hat{\theta}\right)}{\sqrt{\frac{4}{508}}}\right].$

Table IV. Hazard ratio estimates resulting from 1000 simulations of a trial with 200 patients (100 per arm) in which all patients achieve an event.

Hazard Ratio, θ	Median PFS on E (months)	Median PFS on C (months)	Interval between clinic visits, V, (months)	Expected value of $\hat{\theta}$[a]	$\hat{\theta}$[b]	$\breve{\theta}$[c]	SE $\ln\left(\breve{\theta}\right)$[d]
0.667	6	4	0.5	0.677	0.679	0.672	0.1438
			1	0.686	0.681	0.669	0.1418
			2	0.705	0.690	0.664	0.1388
			4	0.740	0.718	0.668	0.1428
0.75	8	6	0.5	0.755	0.751	0.746	0.1483
			1	0.761	0.759	0.750	0.1438
			2	0.771	0.764	0.748	0.1376
			4	0.792	0.782	0.751	0.1433
0.80	12	9.6	0.5	0.803	0.804	0.802	0.1425
			1	0.806	0.808	0.803	0.1440
			2	0.811	0.811	0.802	0.1377
			4	0.822	0.817	0.799	0.1474

[a] Expected value of $\hat{\theta}$ via equation (2) to illustrate the closeness of the simulation to the theoretical result.

[b] $\hat{\theta}$ = geometric mean of 1000 hazard ratios based on analysis of PFS time where timing of progression is assigned to the visit at which is was detected.

[c] $\breve{\theta}$ = geometric mean of 1000 simulated hazard ratios based on equation (6).

[d] SE $\breve{\theta}$ = standard deviation 1000 simulated hazard ratios based on equation (6).

DOI: 10.1002/pst

distributions [25]. While there is no means given for estimating an overall treatment effect such as the hazard ratio, at least a p-value can be obtained to assess the strength of evidence against the null.

(v) As an alternative to the analysis of PFS time, treatments could be compared on the basis of the overall number of PFS events occurring at any time during the trial follow-up period thus circumventing issues associated with the over-estimation of PFS time, scheduling of visits and any asymmetry between treatments.

Under proportionality, an analysis with a complementary log-log link function [21,24,26,27] would provide an unbiased estimate of the hazard ratio as

$$\tilde{\theta} = \frac{\ln(1-p_E)}{\ln(1-p_C)} \tag{8}$$

where p_E and p_C are the proportions of patients with a PFS event on treatment E and C. The estimated variance of $\ln\tilde{\theta}$ by Taylor series expansion is

$$\hat{\mathrm{V}}\mathrm{ar}[\ln(\tilde{\theta})] = \frac{p_E}{N(1-p_E)\{\ln(1-p_E)\}^2} + \frac{p_C}{N(1-p_C)\{\ln(1-p_C)\}^2} \tag{9}$$

It is interesting to note that $\tilde{\theta}$ and $\hat{\mathrm{V}}\mathrm{ar}[\ln(\tilde{\theta})]$ coincide with $\hat{\theta}$ and $\hat{\mathrm{V}}\mathrm{ar}[\ln(\hat{\theta})]$ when $V=R+F$, that is, when there is just one assessment of progression coinciding with the end of the trial period.

Further, by noting that $S_C(t)^{\theta} = S_E(t)$ where θ is the true hazard ratio and, with no censoring, that $S_E(R+F)=p_E$ and $S_C(R+F) = p_C$ so that $\tilde{\theta}=\theta$ and the number of events expected on E and C are $N[S_E(R+F)]$ and $N[S_C(R+F)] = Np_E$ and Np_C, it is possible to compare the power of the log rank test on exact PFS times with the power of an analysis with a complementary log–log link based only on the total number of events occurring over the trial follow-up period. Under these circumstances, the relative efficiency of the two tests is given by

$$\frac{\frac{1}{Np_E}+\frac{1}{Np_C}}{\mathrm{Var}[\ln(\tilde{\theta})]} = \left[\frac{1}{p_E}+\frac{1}{p_C}\right] \times \left[\frac{p_E}{(1-p_E)\{\ln(1-p_E)\}^2} + \frac{p_C}{(1-p_C)\{\ln(1-p_C)\}^2}\right]^{-1}$$

Assuming 90% power in the log rank test, Figure 1 plots the relative efficiency for values of $S_C(t)$ from 0.05 to 0.95.

In line with work earlier work, Figure 1 indicates that, under proportionality, a comparison on the overall number of PFS events over the follow-up period is associated with little loss of power relative to the log rank test on exact PFS times providing fewer than around 50% of patients have reached an event [28,29]. If fewer than 75% of patients have reach an event, the loss in power is, at most, 5%. It is not until 90% or more have reached an event that the power of the relative risk test dips below 80% to around 77%. For exponentially distributed times to event, since the probability of an event over the trial follow-up period $\approx 1-\mathrm{e}^{-\lambda(0.5R+F)}$, fewer than 75% events will in general be assured if the median follow-up at the time of the analysis is not more than two times the median PFS time [20].

This suggests that in those trial settings where progression of disease is the primary focus but significant concerns persist regarding the assumed time of progression, a supportive analysis based on the number of patients with a PFS event over the trial follow-up period can provide reassurance. This analysis is unbiased under proportionality and suffers relatively little loss of power under common trial circumstances. It also offers the opportunity to simplify clinical trial design. It might be possible, for example, to envisage a trial where progression is assessed as per clinical practice with a requirement for objective verification of any suspected disease progression. At a minimum and in addition to the baseline assessment of disease, a single mandatory assessment at the end of the trial follow-up period would be

DOI: 10.1002/pst

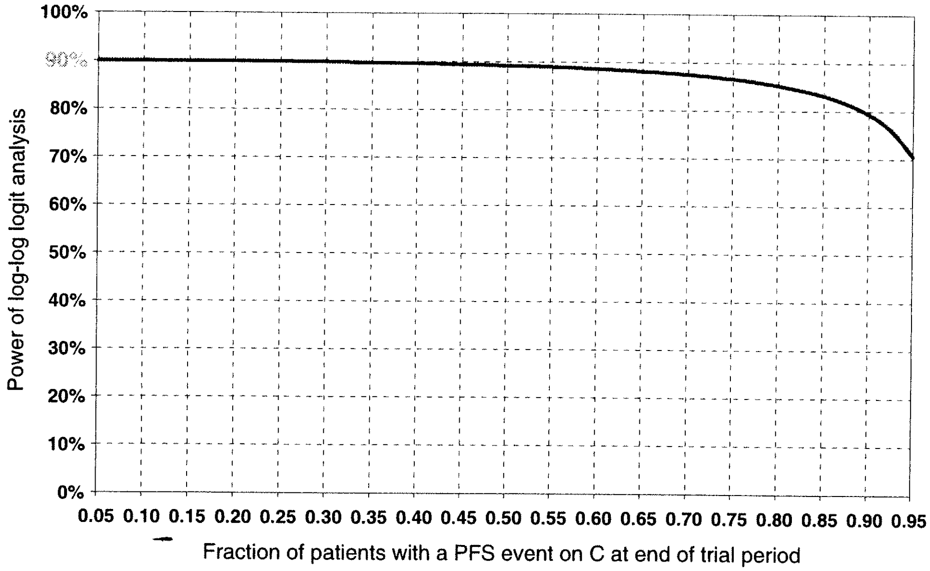


Figure 1. Power of a comparison based on the number of patients with a PFS event over the trial period relative to the log rank test on PFS time with 90% power.

required in patients who had not previously progressed in order to catch any missed progressions. PFS events would then be counted over the trial follow-up period and treatments compared via an analysis with a complementary log-log link function.

CENSORING ON DROPOUT DUE TO AE OR ADDITIONAL ANTI-CANCER THERAPY

FDA's recent Draft Guidance for Industry on Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics recommends that patients who stop taking randomized therapy prior to documented progression are censored at the time randomized treatment is stopped [15]. The rationale for this recommendation is not provided explicitly, but seems to be related to a concern that PFS times may be over-estimated otherwise. Patients who die in the absence of documented progression remain an event, irrespective of whether the death occurred whilst the patient was still receiving or some time after stopping randomized therapy.

This approach is, unfortunately, highly problematic since it ignores the issue of informative censoring. Patients who stop taking randomized therapy prior to documented progression frequently do so due to either toxicity of the drug or due to a deterioration in the status of their disease. In such cases, the treating physician often judges that immediate intervention, commonly in terms of the introduction of a new cancer treatment, is in the best interests of the patient without necessarily waiting for confirmatory, radiographic evidence of progressive disease.

Time to progression therefore cannot be censored in the analysis as the censoring mechanism is self evidently informative. In such circumstances, if the prevalence of censoring differs between arms, naive censoring could lead to extremely biased results and, ultimately, incorrect licensing decisions [30]. Figure 2 provides a simple illustration of the problem.

Suppose E is compared to C in a trial of 100 patients, 50 per arm. Suppose on C that 25 patients progress whilst taking drug at a mean

DOI: 10.1002/pst

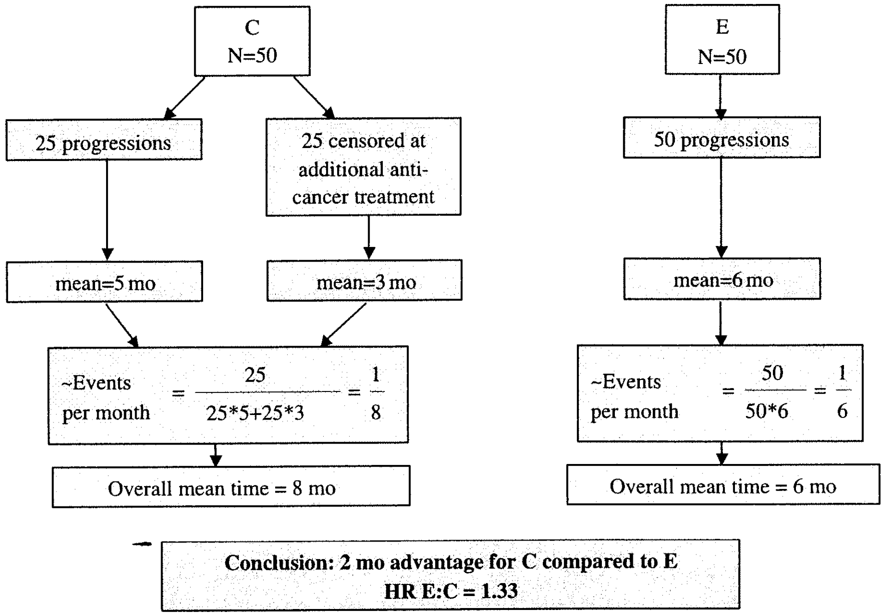

C
N=50
25 progressions
25 censored at additional anti-cancer treatment
mean=5 mo
mean=3 mo
~Events per month = 25 / 25*5+25*3 = 1 / 8
Overall mean time = 8 mo
E
N=50
50 progressions
mean=6 mo
~Events per month = 50 / 50*6 = 1 / 6
Overall mean time = 6 mo
Conclusion: 2 mo advantage for C compared to E
HR E:C = 1.33

Figure 2. Censoring on the addition of further anti-cancer therapy.

time of 5 months and the other 25 patients receive additional anti-cancer treatment prior to documented progression at a mean time of 3 months. Suppose on *E* that all 50 patients progress at a mean time of 6 months; no patients received additional anti-cancer treatment. It is obvious that *E* is the better treatment, with a longer time to progression and no need for additional anti-cancer treatment. However, suppose now that the data are subject to formal statistical analysis with the 25 patients on *C* who received additional anti-cancer treatment censored for progression. The progression event rate on drug *C* is therefore $\frac{1}{8}$ progressions per patient per month compared to $\frac{1}{6}$ progressions per patient per month on drug *E*, giving mean PFS times of 8 and 6 months for *C* and *E*, respectively, and a hazard ratio of 1.33, leading to a conclusion that, in fact, drug *C* is better than *E*. Clearly, there is a problem with a recommended statistical analysis when it leads to a conclusion that the less efficacious and more toxic treatment is better. If, however, those on *C* who received additional anti-cancer therapy are treated rather as failures as in Figure 3, a more sensible conclusion is reached that *E* is in fact better than *C*.

Hence, recommendations to censor patients who stop taking randomized treatment prior to documented progression, perhaps due to the use of additional anti-cancer therapy owing to a deterioration in their condition or due to toxicity, are inherently flawed and should be avoided. This practice, if adopted, not only results in informative censoring but also contravenes the basic principle of an intent-to-treat analysis which is the accepted standard for the comparison of treatments for survival. If a similar approach was applied to the analysis of survival, then only those deaths occurring on randomized treatment would be considered when comparing treatments with all other deaths censored. The interpretation of such an analysis is, at best, unclear and its relevance to the assessment of treatment policies questionable. Overall, it would seem better and more consistent to apply a common standard to important efficacy variables such as PFS and survival to allow both to be interpreted within the same framework. For

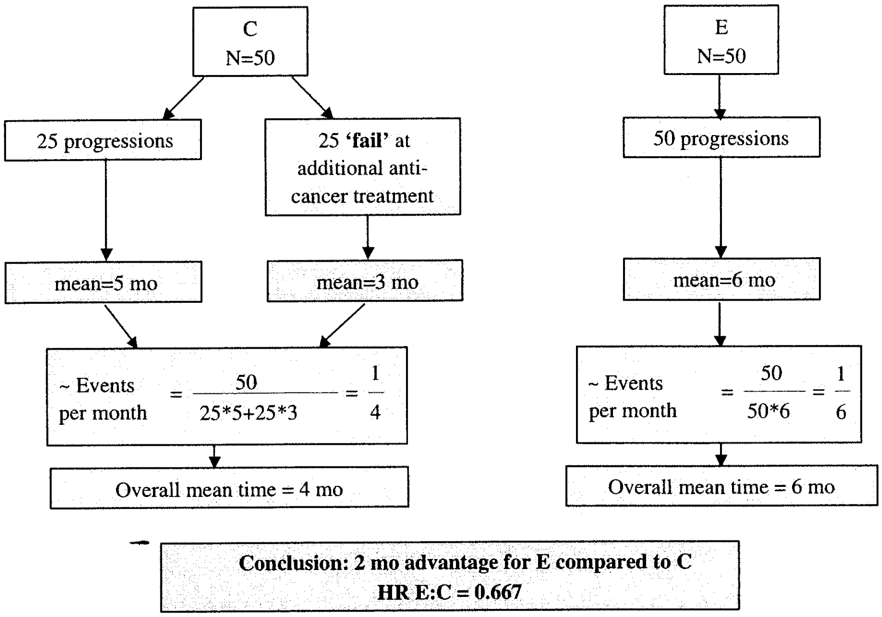

C
N=50
25 progressions
25 'fail' at additional anti-cancer treatment
mean=5 mo
mean=3 mo
~ Events per month = 50 / 25*5+25*3 = 1/4
Overall mean time = 4 mo
E
N=50
50 progressions
mean=6 mo
~ Events per month = 50 / 50*6 = 1/6
Overall mean time = 6 mo
Conclusion: 2 mo advantage for E compared to C
HR E:C = 0.667

Figure 3. Addition of further anti-cancer therapy considered as a 'failure'.

progression (like survival) this would mean the routine follow-up patients for documented evidence of progression irrespective of when and why they stop taking randomized treatment so that treatment policies could be compared on the basis of data that reflect actual clinical practice. This is essentially the same approach as forwarded in the recently published appendix to the CHMP anti-cancer guideline [16].

SUMMARY, RECOMMENDATIONS AND DISCUSSION

This paper has focused on some key statistical issues associated with the analysis of PFS in oncology trials. The routine practice of assigning the time of progression to the clinic visit at which it was first detected results in a downwardly biased estimate of the hazard ratio and, thus, reduces power. Further, if clinic visit schedules are not closely matching between treatments, the Type I error can also be increased. Fortunately, these issues can be addressed as follows:

- Size the trial to detect a true HR of size θ via the log rank test. Assume E events are required to provide power of $1-\beta$ with a 1-sided Type I error rate of α Plan for, and maintain during conduct, a common clinic visit schedule.
- Employ an interval-censored analysis of PFS.
- Alternatively, PFS times can be assigned to clinic visit at which they were first detected and PFS time analysed via the usual log rank test; however, to maintain power at $1-\beta$ the target number of events should be increased to d' (as defined in equation (4)).
- If, as is common in practice, the visit schedule is not as closely adhered to as intended, resulting in variability in the interval between visits and, possibly, between treatments also, a supportive analysis based on the number of PFS events over the trial period will provide for an

unbiased comparison between treatments. An analysis with a complementary log–log link will provide an estimate of the hazard ratio with reasonable power so long as no more than around 75% of patients have an event.

To illustrate the problems associated with assigning the time of progression to the visit at which it was first detected, it has been assumed event times are exponentially distributed since this allows the reader to most easily appreciate the extent to which bias can be introduced and how alternative estimators might be formulated to eliminate this bias. An area of further work might be to look at how rank based estimators of the hazard ratio (such as the Pike estimator or the exponent of the ratio of observed minus expected deaths to the variance from the log rank test) perform when PFS times are not known exactly [31,32]. It might also be of interest to examine other distributions for PFS times, such as the Weibull or log Normal and the performance of $\breve{\theta}$ when proportionality holds but the underlying distribution of PFS times is not exponential.

With respect to patients who stop randomized therapy during the trial follow-up period prior to reaching a confirmed progression event due to toxicity or the addition of further anti-cancer therapy, the common practice of censoring at the time of the intermediate event is highly problematic and is likely to favour the less efficacious, more toxic treatment. Adopting the ITT approach used in the analysis of survival, whereby all patients are followed for a documented evidence of progression irrespective of when and why they stop taking randomized treatment, would provide (i) a better basis for comparing treatment policies and (ii) data that more closely mimic actual clinical practice. If desired, a supportive analysis could still be conducted censoring dropouts in the absence of documented progression, though considerable care would be needed when interpreting the results.

The issues raised in this article are not the only concerns that impact the use of PFS as an endpoint to demonstrate drug effectiveness. Two key issues worth raising briefly are (a) whether an improvement in PFS is a clinical benefit in and of itself or is at least reasonably likely to predict clinical benefit in terms of symptomatic improvement and/or overall survival and (b) the need for independent review of radiological data relating to disease progression. With respect to the first of these issues, recent work in prostate and colorectal cancer has seen the question of surrogacy of PFS for survival carefully and formally examined using contemporary statistical methodology [33–35]. This work supports the use of PFS as a true surrogate endpoint in these disease settings and, in doing so, lends support to the view offered by Williams, that few in the oncology community doubt that delaying the growth of a cancer is of benefit to patients; rather, issues relate to whether progression can be reliably measured in trials and, if so, what a given improvement in progression means clinically [6, transcript p. 30].

The use of open trials in oncology raises the possibility of bias in the assignment of progression status by the treating investigator. As evidenced in both FDA and CHMP guidelines, this concern frequently results in a request from regulatory agencies for independent review of radiographic and imaging data in patients said to have progressed by the investigator [9,15,16]. While this may make sense in open, small scale trials with few investigational sites, the value of independent review in large-scale international trials with possibly hundreds of sites is questionable – when seeking a large effect on progression, a false claim would seem rather unlikely in the absence of a systematic intent to defraud across multiple countries and sites. A further difficulty introduced when incorporating an independent review, is how to handle patients where the investigator and independent review disagree on progression status and where the investigator believes the patient has progressed. In this situation, radiological assessment will cease and any censoring of this data will be informative as such patients will be closer on average to progressing than patients neither the investigator or independent reviewer believe have progressed.

Even when an independent review is deemed worthwhile, the common practice to review only data in patients who have progressed is unsatisfactory at best and misleading at worst. This approach will always lead to a less precise estimate of the treatment effect since the number of progression events can only go down. A more satisfactory approach would be to also take a random sample of non-progressing patients to estimate the fraction of patients without progression reclassified as progressive by independent review. The overall number of progression events could then be estimated under independent review and treatment groups compared accordingly. This approach was used in the review of progression events in the bicalutamide early prostate cancer programme where it was concluded that there was no evidence of bias in the investigator assessment of progression [36].

The trend toward the use of PFS as a primary endpoint to assess the effectiveness of new anticancer treatments is, on the whole, beneficial to drug development and consistent with the aim of FDA's Critical Path and EMEA's Road Map initiatives which actively seek ways to accelerate the drug development process [37,38]. It is hoped that this article will help to address some of the perceived statistical issues related to trial design and analysis and, in doing so, will help to alleviate the concerns and barriers that might otherwise discourage or even prevent the use of PFS as a primary endpoint in oncologic drug development.

ACKNOWLEDGEMENTS

The author would like to thank Andrew Stone, Biostatistics Group, AstraZeneca Pharmaceuticals and three anonymous reviewers for their helpful and constructive comments that served to improve the content and clarity of this paper.

REFERENCES

1. Stadler WM, Ratain MJ. Development of target-based antineoplastic agents. *Investigational New Drugs* 2000; **28**:7–16.
2. Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. Clinical trial designs for cytostatic agents: are new approaches needed? *Journal of Clinical Oncology* 2001; **19**:265–272.
3. Simon RM, Steinberg SM, Hamilton M, Hildesheim A, Khleif S, Kwak LW, Mackall CL, Schlom J, Topalian SL, Berzofsky JA. Clinical Trial Designs for the Early Clinical Development of Therapeutic Cancer Vaccines. *Journal of Clinical Oncology* 2001; **19**:1848–1854.
4. Eskens FA, Verweij J. Clinical studies in the development of new anticancer agents exhibiting growth inhibition in models: facing the challenge of proper study design. *Oncology Haematology* 2000; **34**:83–88.
5. Stone A, Wheeler C, Barge A. Improving the design of phase II trials of cytostatic anticancer agents. *Contemporary Clinical Trials* 2006. In press, Corrected Proof, Available online 14 July 2006 at www.scincedirect.com (last accessed 1 October 2006).
6. FDA Oncologic Drugs Advisory Committee: Endpoints in clinical cancer trials and endpoints in lung cancer clinical trials. 16th December 2003. http://www.fda.gov/ohrms/dockets/ac/03/transcripts/4009T1.DOC (last accessed 1 October 2006).
7. FDA Public Workshop on Clinical Trial Endpoints in Prostate Cancer, 21–22 June 2004. http://www.fda.gov/cder/drug/cancer_endpoints/default.htm#prostate (last accessed 1 October 2006).
8. FDA Oncologic Drugs Advisory Committee: Colorectal cancer endpoint discussion. 4th May 2004, transcript page 218. http://www.fda.gov/ohrms/dockets/ac/04/transcripts/4037T2.DOC (last accessed 1 October 2006).
9. CHMP Guideline on the evaluation of anticancer medicinal products in man. December 2005. http://www.emea.eu.int/pdfs/human/ewp/020595en.pdf (last accessed 1 October 2006).
10. FDA/CDER. FDA Project on Cancer Drug Approval Endpoints, 2003: http://www.fda.gov/cder/drug/cancer_endpoints/default.htm (last accessed 1 October 2006).
11. FDA/CDER New and Generic Drug Approvals: NEXAVAR (sorafenib) product labelling, December 2005. Label: http://www.fda.gov/cder/foi/label/2005/021923lbl.pdf (last accessed 1 October 2006). Approval letter: http://www.fda.gov/cder/foi/appletter/2005/021923ltr.pdf (last accessed 1 October 2006).
12. FDA/CDER New and Generic Drug Approvals: VECTIBIX (panitumumab) product labelling, September 2006. Label: http://www.fda.gov/cder/foi/label/2006/125147s0000lbl.pdf (last accessed 1 October 2006).
13. Williams G, He K, Chen G, Chi G, Pazdur R. Operational Bias in assessing time to progression (TTP). *Proceedings of the American Society of Clinical Oncology* 2002; abstr 975.
14. FDA Oncologic Drugs Advisory Committee: Genasense^{TM} (oblimersen sodium) Injection for

DOI: 10.1002/pst

Advanced Melanoma in Combination With Dacarbazine (DTIC). 4th May 2004. Statistical Review. http://www.fda.gov/ohrms/dockets/ac/04/briefing/4037B1_04_C-FDA-Statistical%20Review-RSR13.doc and http://www.fda.gov/ohrms/dockets/ac/04/slides/4037S1_04_FDA-Sridhara-Ridenhour.ppt (last accessed 1 October 2006).

15. FDA/CDER Guidance for Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics DRAFT GUIDANCE, 2005. http://www.fda.gov/cder/guidance/6592dft.htm (last accessed 1 October 2006).
16. CHMP Draft appendix 1 to the guideline on the evaluation of anticancer medicinal products in man. London, 27 July 2006. http://www.emea.eu.int/pdfs/human/ewp/26757506en.pdf#search = %22emea%20anti-cancer%20guideline%20appendix%22 (last accessed 1 October 2006).
17. Schoenfeld DA. The asymptotic properties of nonparamteic tests for comparing survival distributions. *Biometrika* 1981; **68**:316–318.
18. Rubinstein LV, *et al.* Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Diseases* 1981; **34**:469–479.
19. Yateman NA, Skene AM. Sample sizes for proportional hazards survival studies with arbitrary patient entry and loss to follow-up distributions. *Statistics in Medicine* 1992; **11**:1103–1113.
20. Bernstein D, Lagakos SW. Sample size and power determination for stratified clinical trials. *Journal of Statistical Computer Simulations* 1978; **8**:65–73.
21. Stone A, Wheeler C, Carroll K, Barge A. Optimizing randomized phase II trials assessing tumor progression. *Contemporary Clinical Trials* 2006. In press, Corrected Proof, Available online 19 May 2006 at www.scincedirect.com (last accessed 1 October 2006).
22. Whitehead J. The analysis of relapse clinical trials, with application to a comparison of two ulcer treatments. *Statistics in Medicine* 1989; **8**:1439–1454.
23. SAS/STAT® *User's Guide. Version 6* (4th edn), Vol. 2. SAS Institute Inc.: Cary, NC, 1989.
24. Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* 1978; **34**:57–67.
25. Sun J, Zhao Q, Zhao X. Generalised log-rank tests for interval-censored failure time data. *Scandinavian Journal of Statistics* 2005; **32**:49–57.
26. Cox DR, Oakes D. *Analysis of Survival Data*. Chapman & Hall: London, 1984.
27. Collett D. *Modeling Survival Data in Medical Research*. Chapman & Hall: London, 1994.
28. Cuzick J. The efficiency of the proportions test and the log rank test for censored survival data. *Biometrics* 1982; **38**:1033–1039.
29. Gail MH. Applicability of sample size calculations based on a comparison of proportions for use with the log rank test. *Controlled Clinical Trials* 1985; **6**:112–119.
30. DiRienzo AG. Nonparametric comparison of two survival-time distributions in the presence of dependent censoring. *Biometrics* 2003; **59**:497–504.
31. Berry G, Kitchin RM, Mock PA. A comparison of two simple hazard ratio estimators based on the logrank test. *Statistics in Medicine* 1991; **10**: 749–755.
32. Sellke T, Siegmund D. Sequential analysis of the proportional hazards model. *Biometrika* 1983; **70**:315–326.
33. Collette L, Burzykowski T, Carroll KJ, *et al.* Is prostate-specific antigen a valid surrogate end point for survival in hormonally treated patients with metastatic prostate cancer? Joint Research of the European Organisation for Research and Treatment of Cancer, the Limburgs Universitair Centrum, and AstraZeneca Pharmaceuticals. *Journal of Clinical Oncology* 2005; **23**:6139–6148.
34. Buyse M, Burzykowski T, Carroll K, *et al.* Progression-free survival (PFS) as a surrogate for overall survival (OS) in patients with advanced colorectal cancer: an analysis of 3159 patients randomized in 11 trials. *Proceedings of the American Society of Clinical Oncology* 2005; abstr 3513.
35. Molenberghs G, Buyse M, Geys H, *et al.* Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials* 2002; **23**:607–625.
36. FDA/CDER Public Workshop on Clinical Trial Endpoints in Prostate Cancer, 21–22 June 2004. Re-evaluation of Radiographic Outcomes: The Casodex Early Prostate Cancer Trial Program Experience. http://www.fda.gov/cder/drug/cancer_endpoints/ProstatePresent/Carroll.ppt (last accessed 1 October 2006).
37. FDA Challenge and opportunity on the Critical Path to New Medical Products. March 2004. http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html (last accessed 1 October 2006).
38. The European Medicines Agency Road Map to 2010: Preparing the Ground for the Future. March 2005. http://www.emea.eu.int/pdfs/general/direct/directory/3416303enF.pdf (last accessed 1 October 2006).

APPENDIX A

Suppose N patients are followed for some event of interest with the time to the event denoted as t.

Suppose t is an exponentially distributed random variable with parameter λ. Suppose the process is monitored at r equally spaced intervals of length V for a total follow-up time of rV. Events therefore occur in intervals $(0, V]$, $(V, 2V]$,...,$((r-1)V, rV]$ with the event for the ith patient occurring in $((k_i-1)V, k_iV]$. Events times are assigned to the start of the interval in which they are detected. There are d patients with an event. Furthermore, c patients are randomly censored prior to time rV with censoring times of k_gV for the gth patient. The $N-d-c$ remaining patients who are without an event at the end of the follow-up period are right censored at time rV. The likelihood function is therefore given by

$$L = \prod_{i=1}^{d} \left(\mathrm{e}^{-\lambda(k_i-1)V} - \mathrm{e}^{-\lambda k_i V}\right) \prod_{g=1}^{c} \mathrm{e}^{-\lambda k_g V} \prod_{j=1}^{N-d-c} \mathrm{e}^{-\lambda r V} \tag{A1}$$

$$\ell = \ln(L) = d\ \ln\left(\mathrm{e}^{\lambda V} - 1\right) - \lambda\left[\sum_{i=1}^{d} k_i V + \sum_{g=1}^{c} k_g V + \sum_{j=1}^{N-d-c} rV\right] \tag{A2}$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{dV\mathrm{e}^{\lambda V}}{\mathrm{e}^{\lambda V} - 1} - \left[\sum_{i=1}^{d} k_i V + \sum_{g=1}^{c} k_g V + \sum_{j=1}^{N-d-c} rV\right] \tag{A3}$$

$$\frac{\partial \ell}{\partial \lambda} = -d\left[\frac{\sum_{i=1}^{N} T_i}{d} - \frac{V}{1 - \mathrm{e}^{-\lambda V}}\right] \tag{A4}$$

where T_i denotes the observed time, event or censored, of the ith patient. Thus

$$\overline{T} = \frac{\sum_{i=1}^{N} T_i}{d} = \frac{1}{\text{observed event rate}}$$

is the MLE for

$$\frac{V}{1 - \mathrm{e}^{-\lambda V}} \text{ and } \mathrm{Var}\left[\overline{T}\right] = \frac{V^2 \mathrm{e}^{-\lambda V}}{d(1 - \mathrm{e}^{-\lambda V})^2}$$

If comparing two treatments, experimental (E) and control (C), the estimated hazard ratio

$$\hat{\theta} = \frac{\text{observed event rate}_E}{\text{observed event rate}_C} = \frac{\overline{T}_C}{\overline{T}_E}$$

is biased since

$$E\,[\hat{\theta}] = \frac{1 - \mathrm{e}^{-\lambda_E V}}{1 - \mathrm{e}^{-\lambda_C V}} \neq \frac{\lambda_E}{\lambda_C} \tag{A5}$$

Further,

$$\mathrm{Var}\,[\ln(\hat{\theta})] = \frac{\mathrm{e}^{-\lambda_E V}}{d_E} + \frac{\mathrm{e}^{-\lambda_C V}}{d_C} \tag{A6}$$

and, thus,

$$\hat{\mathrm{V}}\mathrm{ar}\,[\ln(\hat{\theta})] = \frac{1 - \frac{V}{\overline{T}_E}}{d_E} + \frac{1 - \frac{V}{\overline{T}_C}}{d_C} \tag{A7}$$

An approximately unbiased estimate of the HR is given by

$$\breve{\theta} = \frac{\ln\left(1 - \frac{V}{\overline{T}_E}\right)}{\ln\left(1 - \frac{V}{\overline{T}_C}\right)} \tag{A8}$$

with variance

$$\mathrm{Var}\,[\ln(\breve{\theta})] = \frac{\mathrm{e}^{\lambda_E V}\left(1 - \mathrm{e}^{-\lambda_E V}\right)^2}{d_E \lambda_E^2 V^2} + \frac{\mathrm{e}^{\lambda_C V}\left(1 - \mathrm{e}^{-\lambda_C V}\right)^2}{d_C \lambda_C^2 V^2} \tag{A9}$$

and, thus,

$$\hat{\mathrm{V}}\mathrm{ar}[\ln(\breve{\theta})] = \frac{V^2}{d_E \overline{T}_E^2\left\{\ln\left(1 - \frac{V}{\overline{T}_E}\right)\right\}^2\left(1 - \frac{V}{\overline{T}_E}\right)} + \frac{V^2}{d_C \overline{T}_C^2\left\{\ln\left(1 - \frac{V}{\overline{T}_C}\right)\right\}^2\left(1 - \frac{V}{\overline{T}_C}\right)} \tag{A10}$$

APPENDIX B

Suppose two treatments are to be compared in a clinical trial on a time to event endpoint using the log rank test. To test the hypotheses H_0: hazard ratio = 1 vs H_1: hazard ratio = θ (<1) with a 1-sided

DOI: 10.1002/pst

Type I error rate of α and power $1-\beta$ a total of

$$d = \frac{4\left(z_\alpha + z_\beta\right)^2}{\ln(\theta)^2} \qquad \text{(B1)}$$

events are required [18]. It therefore follows immediately that

(i) if the power to detect $\hat{\theta}$, $\hat{\theta} > \theta$, with d events is $1-\gamma$, then

$$z_\gamma = \left(z_\alpha + z_\beta\right)\omega - z_\alpha \qquad \text{(B2)}$$

so that

$$1 - \gamma = \phi^{-1}\left[\left(z_\alpha + z_\beta\right)\omega - z_\alpha\right] \qquad \text{(B3)}$$

where $\omega = \mathrm{abs}\left(\frac{\ln(\hat{\theta})}{\ln(\theta)}\right)$ and $\phi^{-1}(.)$ is the inverse of the cumulative normal distribution.

(ii) there is a simple relationship between θ and $\hat{\theta}$ such that

$$\hat{\theta} = \theta^k \qquad \text{(B4)}$$

where $k = \mathrm{abs}\left(\frac{z_\alpha + z_\gamma}{z_\alpha + z_\beta}\right)$

(iii) to maintain power of $1-\beta$ to detect $\hat{\theta}$ a total of

$$d' = \frac{d}{\omega^2} \qquad \text{(B5)}$$

events are required.

Suppose now that $\hat{\theta} = \frac{1-\mathrm{e}^{-\lambda_E V}}{1-\mathrm{e}^{-\lambda_C V}}$. If λV is small, then

$$\begin{aligned}\frac{V}{1-\mathrm{e}^{-\lambda V}} &\approx \frac{1}{\lambda}\left[\frac{1}{1-\frac{\lambda V}{2}+O(\lambda^2 V^2)}\right]\\ &= \frac{1}{\lambda}\left(1+\frac{\lambda V}{2}+O(\lambda^2 V^2)\right)\\ &\approx \frac{1}{\lambda}+\frac{V}{2}\end{aligned} \qquad \text{(B6)}$$

so that

$$\hat{\theta} = \frac{1-\mathrm{e}^{-\lambda_E V}}{1-\mathrm{e}^{-\lambda_C V}} \approx \frac{\frac{1}{\lambda_E}+\frac{V}{2}}{\frac{1}{\lambda_C}+\frac{\lambda_C}{2}} = \frac{\theta(1+\tau)}{1+\tau\theta} \qquad \text{(B7)}$$

where $\tau = \frac{V\lambda_C}{2}$. Substitution of (B7) in to (B4) reveals that if a common visit schedule is used when assessing PFS such that PFS times are assigned to the visit at which progression was first detected, to ensure retention of at least $1-\gamma$ power, $\gamma > \beta$, visits must be scheduled approximately every V' months where

$$V' = (\text{median PFS on } C) \times \frac{2}{\ln(2)}\,\frac{1}{\theta}\left(\frac{\theta^k-\theta}{1-\theta^k}\right) \qquad \text{(B8)}$$

DOI: 10.1002/pst