EUROPEAN JOURNAL OF CANCER 47 (2011) 1763–1771

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Research outcomes and recommendations for the assessment of progression in cancer clinical trials from a PhRMA working group

A.M. Stone [a,*], W. Bushnell [b], J. Denne [c], D.J. Sargent [d], O. Amit [b], C. Chen [e], R. Bailey-Iacona [f], J. Helterbrand [g], G. Williams [h]

[a] AstraZeneca, Alderley Park, Macclesfield SK10 4TG, UK
[b] GSK, Collegeville, PA, USA
[c] Eli-Lilly, Indianapolis, IN, USA
[d] Mayo Clinic, Rochester, MD, USA
[e] Merck, North Wales, PA, USA
[f] AstraZeneca, Wilmington, DE, USA
[g] Genentech, South San Francisco, CA, USA
[h] Williams Cancer Drug Consulting, Wayne, PA, USA

ARTICLE INFO

Article history:
Received 29 October 2010
Accepted 16 February 2011
Available online 22 March 2011

Keywords:
Progression
Randomised clinical trials
Solid tumours
Bias

ABSTRACT

Purpose: Progression free survival (PFS) is increasingly used as a primary end-point in oncology clinical trials. This paper provides recommendations for optimal trial design, conduct and analysis in situations where PFS has the potential to be an acceptable end-point for regulatory approval.

Patients and methods: These recommendations are based on research performed by the Pharmaceutical Research and Manufacturers Association (PhRMA) sponsored PFS Working Group, including the re-analysis of 28 randomised Phase III trials from 12 companies/institutions.

Results: (1) In the assessment of PFS, there is a critical distinction between measurement error that results from random variation, which by itself tends to attenuate treatment effect, versus bias which increases the probability of a false negative or false positive finding. Investigator bias can be detected by auditing a random sample of patients by blinded, independent, central review (BICR). (2) ITT analyses generally resulted in smaller treatment effects (HRs closer to 1) than analyses that censor patients for potentially informative events (such as starting other anti-cancer therapy). (3) Interval censored analyses (ICA) are more robust to time-evaluation bias than the log-rank test.

Conclusion: A sample based BICR audit may be employed in open or partially blinded trials and should not be required in true double-blind trials. Patients should be followed until progression even if they have discontinued treatment to be consistent with the ITT principle. ICAs should be a standard sensitivity analysis to assess time-evaluation bias. Implementation of these recommendations would standardize and in many cases simplify phase III oncology clinical trials that use a PFS primary end-point.

* *Corresponding author:* Tel.: +44 1625 515969; fax: +44 1625 518537.
E-mail address: Andrew.stone@astrazeneca.com (A.M. Stone).
0959-8049/$ - see front matter
doi:10.1016/j.ejca.2011.02.011

1. Introduction

Progression of a patient's cancer is a devastating event leading to further treatment, complications and in most cases eventual death. Evaluating tumour progression in oncology clinical trials has intuitive appeal; however defining disease progression within clinical trials is not without complication or controversy.

The two most common end-points for assessing tumour progression are time to tumour progression (TTP) and progression free survival (PFS). Both end-points measure the time from randomization until objective tumour progression but PFS also includes death due to any cause. PFS is the preferred end-point for most regulatory settings because death from any cause is relevant to describing or predicting clinical benefit.[1,2] In solid tumours, progression is typically determined using the RECIST criteria,[3] which has standardised many of the technical aspects of tumour burden assessment.

In recent years there has been an increase in the use and regulatory acceptance of PFS as the primary end-point in cancer clinical trials. In a 1991 FDA-NCI white paper discussing tumour end-points for drug approval, neither TTP nor PFS were discussed.[4] From 1990 until 2002 there was a single FDA approval relying primarily on TTP or PFS without also relying on a survival benefit.[5] In contrast, over the past seven years the FDA has approved at least nineteen drug applications that were primarily based on a PFS end-point (Table 1).

PFS is considered an acceptable end-point by regulatory authorities in many countries; the United States (US) and European Union have regulatory guidance pertaining to the use of this end-point.[2,6,7] The FDA has promoted a series of open workshops on end-points for different tumours,[8] and has also addressed PFS in the setting of Oncologic Drug Advisory Committee (ODAC) meetings for lung cancer and colon cancer in 2003 and 2004 respectively.[1,9]

A key advantage of the PFS end-point is that since progression occurs months or years before death, the time required for the necessary number of events to provide the required statistical power is shorter than for an overall survival (OS) end-point. Further, the effect sizes that may be expected (in terms of hazard ratios) are larger compared to those expected for OS.[10] Perhaps most critically, the treatment effect

Table 1 – FDA Cancer Drug Approvals since November 2002 based primarily on PFS/TTP.[a]

Year	Drug (application, approval type)[b]	Treatment indication (line of therapy)[c]	Design[d]	End-point	PFS or TTP Findings		
					Hazard ratio (confidence interval)	*p* value	Medians (Mos.)
2002	Imatinib mesylate (S,AA)	CML, 1L	AC	PFS	0.18 (0.12, 0.28)	<0.001	5.5, 2.8
2004	Gemcitabine (S, RA)	Breast cancer, 1L	AC-AO	PFS	0.65 (0.52, 0.81)	<0.0001	5.2, 2.9
2005	Sorafenib (N, RA)	Renal cell cancer, 2L	PBO/BSC	TTP	0.44 (0.35, 0.55)	<0.0001	5.5, 2.8
2006	Sunitinib (N, RA)	GIST, 2L	PBO/BSC	TTP	0.33 (0.23, 0.47)	<0.0001	6.3, 1.5
2006	Lenalidomide (S, RA)	Myeloma, 2L	AC-AO	TTP	0.36 (0.26, 0.49)	<0.0001	8.6, 4.6
		Myeloma, 2L	AC-AO	TTP	0.39 (0.27, 0.47)	<0.0001	NR, 4.6
2006	Gemcitabine (S, RA)	Ovarian cancer, 2L	AC-AO	PFS	0.72 (0.57, 0.90)	0.0038	8.6, 5.8
2006	Panitumumab (N, AA)	Colon cancer, 2L	PBO/BSC	PFS	0.54 (0.44, 0.66)	<0.0001	3.0, 2.0 (mean)
2006	Rituximab (S, RA)	Low grade lymphoma, 1L	AC-AO	PFS	0.44 (0.29, 0.65)	<0.0001	28.8, 16.8
		'Maintenance'	PBO/BSC	PFS	0.36–0.49	NP	NP
2007	Sunitinib (S, RAC)	Renal cell cancer, 1L	AC	PFS	0.42 (0.32, 0.54)	<0.0001	10.9, 5.1
2007	Lapatinib (N, RA)	Breast cancer, 2L	AC-AO	PFS	0.57 (0.43, 0.77)	0.0001	6.3, 4.3
2007	Liposomal doxorubicin (S, RA)	Myeloma, 2L	AC-AO	TTP	0.55 (0.43, 0.71)	<0.0001	9.3, 6.5
2007	Alemtuzumab (S, RAC)	B-CLL, 1L	AC	PFS	0.58 (0.44, 0.77)	0.0001	14.6, 11.7
2007	Ixabepilone (N, RA)	Breast cancer, 2L	AC-AO	PFS	0.69 (0.48, 0.83)	<0.0001	5.7, 4.1
2008	Bevacizumab (S, AA)	Breast cancer, 1L	AC-AO	PFS	0.48 (0.39, 0.61)	<0.0001	11.3, 5.8
2008	Deniliukin diftitox (S, RAC)	CTCL, 2L	PBO/BSC	PFS	0.27 (0.14, 0.54)	0.0002	7.2, 2.7
2008	Bendamustine (N, RA)	CLL, 1L	AC	PFS	0.27 (0.17, 0.43)	<0.0001	18, 6
2008	Everolimus (N, RA)	Renal cell cancer, 2L	PBO/BSC	PFS	0.33 (0.25, 0.43)	<0.0001	4.9, 1.9
2009	Bevacizumab (S, RA)	Renal cell cancer, 1L	AC-AO	PFS	0.60 (0.49, 0.72)	<0.0001	10.4, 5.5
2009	Pazopanib (N, RA)	Renal cell cancer, 1L	PBO/BSC	PFS	0.46 (0.34, 0.62)	<0.001	9.2, 4.2
2010	Rituximab (S, RA)	CLL, IL	AC-AO	PFS	0.56 (0.43, 0.71)	<0.01	39.8, 31.5
		CLL, 2L	AC-AO	PFS	0.76 (0.60, 0.96)	0.02	26.7, 21.7

CML: chronic myelogenous leukaemia; GIST: Gastrointestinal stromal tumour; CTCL: Cutaneous T-cell lymphoma, CLL: Chronic lymphocytic leukaemia, TTP: time to progression, PFS: progression free survival, NR: not reached, NP: not provided in label.

[a] Data in table from respective FDA package inserts. Medians were converted to months.

[b] S: efficacy supplement; N: new drug application; AA: accelerated approval; RA: regular approval; RAC: conversion of accelerated to regular approval.

[c] 1L: first line therapy; 2L: second or subsequent lines of therapy.

[d] AC: active control (e.g., drug A versus drug B); AC-AO: active control add-on design (A + B versus A); PBO/BSC: placebo control or best supportive care.

observed in studies where PFS is the primary variable can provide an unambiguous measure of treatment effect, because the PFS end-point is generally achieved before patients receive subsequent alternative therapies.

There are however, several important criticisms concerning the use of PFS as a primary end-point in oncology clinical trials. The assessment of progression is not entirely objective and can be subject to bias. This is particularly a concern in open label studies, but is also a concern in double blind studies where toxicities can create partial unblinding. There is concern that a demonstrated effect on PFS may not translate into an effect on OS,[11] and further that the effect observed is dependent on the data handling and analysis methods employed. Additionally, it is not readily understood how a PFS effect seen in a clinical trial translates to a direct benefit for an individual patient.

To assess PFS, patients are generally followed until either progression or death. Patients who, at the data cut-off time for analysis, are known to be alive and whose disease has not progressed are censored by necessity, generally at their last assessment at which they were known to be progression-free. Additional patients may be censored in the statistical analysis: patients who stop randomised therapy, miss assessment visits or start to take additional anti-cancer therapy. Censoring this latter group of patients can be problematic, as such censoring can be 'informative'. Informative censoring occurs if patients who are censored are at a higher or lower risk of progression than patients who are not censored. Critically, informative censoring can induce bias,[11–14] particularly if not balanced across treatment arms. Such censoring has important implications for trial conduct highlighting the need for complete patient follow-up.[7,11,15] The importance of censoring is recognised by the presence of detailed methodological regulatory guidance on this issue.[2,7]

A further challenge to the PFS end-point is that unlike overall survival, the true date of progression is not precisely known, it can only be estimated based on the visit framework of a clinical trial. The date of progression is usually assigned as the date when progression is observed radiologically. Data analysed in this way are termed interval censored and results from such analysis can be biased in the presence of an asymmetrical visit schedule.[16]

Though the use of PFS as a primary end-point in oncology clinical trials is increasing, published research on the details of its operating characteristics, which would lead to an increased confidence in its use, are limited.[17–20]

This paper presents a summary of original research and new analyses, conducted through a large coalition of industry, academia and the FDA. Specific topics considered are: (1) the differential effects of variability and bias and its implications for the need for Blinded Independent Central Review (BICR), (2) a re-analysis of individual patient data from 28 randomised phase III trials applying different censoring rules, (3) an investigation into the robustness of interval censored analysis approaches. We also discuss the findings of the companion paper by Amit to summarize the overall conclusions reached by the working group. A set of recommendations is provided that it is believed will promote greater consistency in how trials employing the PFS end-point are designed, analysed, and interpreted.

2. Patients and methods

The Pharmaceutical Research and Manufacturers Association (PhRMA) formed a working group (WG) whose primary purpose was to conduct research to optimise approaches to the design and analysis of trials with PFS end-points. The focus of the WG is to provide data-driven recommendations and reduce the uncertainties associated with the use of PFS in those settings where it is appropriate. There were approximately 35 participants representing 17 pharmaceutical companies, the FDA, academic institutions and imaging contract research organizations within the WG. Inclusion in the WG is open and voluntary.

The effect of increasing levels of measurement variability on the attenuation of a PFS treatment effect was studied by simulating data from a tumour growth model. Firstly, true PFS times were generated from exponential distributions with true medians of 3 and 6 months for Arms 1 and 2, and hence a true HR(Arm2:Arm1) of 0.5. For each individual PFS time, a separate underlying tumour growth model[21] was created with $LD_j = (LD_0) * (\exp(-bt_j) + b * a_i * t_j)$ where, LD_j = Longest Diameter at visit j, LD_0 = baseline longest diameter, t_j = time, in months, of jth visit, the parameter b was fixed at 0.4 and the value of the parameter a for the ith patient, a_i, was calculated so that the PFS time derived from the tumour growth model (20% increase in LD_j from nadir) matched the simulated PFS time. Variability in the baseline longest diameter was added by assuming LD_0 was log-normally distributed so that $\mathrm{Ln}(LD_0) \sim N(1.29, \mathrm{sqrt}(0.205))$, corresponding to the mean baseline value of 4 cm and SD of 1.9 cm reported in a reproducibility study in non-small-cell lung cancer.[22] In order to replicate a clinical trial, it was assumed that patients were assessed every 1.5 months and LD_j values were recorded for each visit. Consequently, the nadir of visit based values does not match true nadir and this results in the simulated median visit based PFS being extended, on average, within each arm. Log-normally distributed measurement error was then incorporated by multiplying LD_j by random values from a normal distribution with zero mean and SDs of 0.077, 0.155 and 0.232, respectively; 0.077 and multiples were chosen as this measurement error was observed in the reproducibility study reported by Zhao.[22] One-thousand simulations were performed with 300 patients per arm and Efron[23] tie-handling.

This paper also presents a re-analysis of individual patient data from 28 trials provided by 12 companies/institutions which examined the impact of different censoring mechanisms. Selection of the trials for inclusion was made by the individual companies/institutions. To address this issue individual patient data was reanalysed from these trials according to a pre-defined analysis plan describing details for an Intention-To-Treat (ITT) analysis, as well as analyses that censor patients according to four defined rules.

- ITT – all recorded PFS events are included regardless of stopping randomised therapy or subsequent therapy.
- PDT – same as ITT, except censor patients who receive subsequent anti-cancer therapy prior to progression at the latest prior assessment.

- DISC – same as ITT, except censor patients who prematurely discontinue randomised therapy due to toxicity or other, non-progression related reasons at the latest prior assessment.
- MV – same as ITT, except censor patients who progress or die (in the absence of progression) after two or more missed visits, at the latest prior assessment.
- ALL – same as ITT, except censor patients who are censored in either PDT, DISC or MV at the earliest censoring time.

We examined whether different censoring rules tended to provide different estimates of the hazard ratio (HR), and the extent to which the percentage of ITT events censored by each rule differed between arms and the effect of this on estimation of the HR. A simple measure of whether the censoring was informative was calculated based on the assumption of an exponential distribution for PFS times. The extent to which this indicated the presence of informative censoring and the effect on the analysis of this differing censoring between treatment arms was examined.

Finally, simulations were performed to compare the robustness of the log-rank test and an interval censored approach to departures from equal between-arm assessment frequency. The log-rank test was performed by using a Cox-regression model with Efron[23] tie-handling; the interval censored method was the Finkelstein score test.[24] Ten-thousand simulations were run assuming a true HR(experimental:control) of 0.6 with a median PFS of 24 weeks for the control group.

3. Results

3.1. *Variability versus bias and the use of Blinded Independent Central Review*

In a randomised clinical trial it is critical to estimate the effect of treatment free from bias. When considering the optimal trial design there needs to be a clear distinction between bias and variability:

- Bias relates to systematic errors that result in over- or under-estimating the true effect of treatment on PFS. This paper focuses on measures for trial design, conduct and analysis that minimise bias.
- Variability relates to measurement error and uncertainty concerning the estimated PFS time for each individual.

Table 2 – Attenuation of treatment effect due to variability.

Standard deviation (cm)	Observed hazard ratio	Attenuation[a] (%)	Observed median[b] PFS (arm1, arm2)
0	0.504	0	4.3, 6.8
0.077	0.511	1	3.5, 6.3
0.155	0.524	4	3.1, 5.8
0.232	0.546	8	3.0, 4.7

[a] % Attenuation calculated as 100 * (observedHR – 0.504)/(1 – 0.504)
[b] Median PFS is the mean of the 1000 medians generated in the simulation study.

Random variability in PFS assessment will attenuate the treatment effect and could contribute to failing to identify an effective therapy, but critically does not result in bias that over-estimates the treatment effect. The effect of variability is described in Table 2, where increasing levels of variability were added to the tumour growth model recently applied in non-small-cell lung cancer.[21] The degree of attenuation is fairly modest (<10%) based on the extent of measurement error observed with expert readers[22] although such attenuation could lead to a loss of power by as much as 10%. These data highlight the importance of investigator training and consistency of imaging modality and reader in application of PFS assessment.

Bias in contrast, leads to an inaccurate estimate of the effect of a treatment. Bias can be detected by using a BICR. Critical to the interpretation of the results from a BICR is distinguishing bias from variability. Commonly quoted levels of disagreement in the progression dates between local and central reviewers primarily reflect the extent of variability. However, as detailed in the companion paper,[25] only if the discordance rate between local and BICR assessment occurs at a differential rate between study arms is the BICR indicative of bias.

Hazard ratios reported from BICR and local evaluations have been found to be highly consistent,[19,25,26] even in open-label trials, suggesting that local evaluations tend not to be subject to bias and that, therefore, the application of a BICR conducted on all patients may not significantly add value to a proper interpretation of PFS trial results. These findings, when coupled with the concern over informative censoring[19] indicate that (1) BICRs may not be necessary in truly blinded trials, and (2) bias may be adequately detected by performing a BICR on a random sample of patients in open or partially blinded trials.

3.2. *Patient follow-up and approaches to censoring*

The characteristics of the trials included are summarised in Table 3. The hazard ratios from the analyses based on the four censoring rules were generally consistent with the ITT approach; the hazard ratio from the 'Censor – All' analysis was within 0.1 of that from the ITT analysis in 21 of the 28 studies. Importantly, although consistent, the ITT analysis generally resulted in smaller treatment effects (HRs closer to 1) than those obtained by applying the various censoring rules (Fig. 1).

The difference in HRs between the ITT and the alternative analyses was smallest when censoring rates were similar between arms: although the relationship between the HR and differential rates of censoring between treatment arms (Online Figure) was only significant when censoring for all reasons was considered ('Censor – All', $p = 0.01$). Consistent with other studies,[13] we found that patients censored under 'PDT', 'DISC' and 'ALL' appear to be at an approximately 30% greater risk for progression relative to the study population as a whole. Different levels of informative censoring between treatment arms was clearly associated with the size of discrepancy between the ITT and censored analysis HR (Fig. 2); when censoring for all reasons considered ('Censor – All', $p < 0.001$), censoring for subsequent therapy ($p = 0.003$) and censoring patients who prematurely discontinue randomised

Table 3 – Overview of PhRMA censoring survey data.

Study characteristics	Surveyed studies ($N = 28$)[a]
Tumour type	
Colorectal	10
Breast	7
Lung	5
Hematologic[b]	2
Prostate	2
Renal	2
Design	
Open label/double-blind	23/5
Add on/substitution[c]	15/10
Superiority/non-inferiority[d]	23/4
Primary end-point (PFS or TTP/OS)	21/7
Sample size (total)	
200–399	5
400–599	10
600–799	7
⩾800	6

Description of analyses and extent of censoring

Censor	Median percentage (range) of ITT[e] events censored		Median percentage (range) of reduction in ITT[e] follow-up	
	Control	Experimental	Control	Experimental
PDT[f]	9 (0–32)	8 (0–32)	7 (0–39)	7 (0–26)
DISC[g]	18 (2–58)	17 (1–52)	16 (1–45)	12 (0–45)
MV[h]	5 (0–21)	5 (0–18)	5 (0–31)	6 (0–22)
ALL[i]	25 (6–59)	23 (3–57)	24 (3–51)	20 (3–50)

Data supplied by Abbott ($n = 2$), Amgen ($n = 2$), AstraZeneca ($n = 1$), BMS ($n = 1$), Eli-Lilly ($n = 3$), Genentech ($n = 5$), GSK ($n = 3$), NCCTG ($n = 4$), Pfizer ($n = 1$), Roche ($n = 4$), Sanofi-aventis ($n = 1$), Takeda ($n = 1$).

[a] Some studies ($n = 2$) had more than 2 arms, and are counted as separate studies with each comparison presented separately.

[b] Multiple Myeloma and Non-Hodgkins Lymphoma.

[c] Data not provided for three trials.

[d] Data not provided for one trial.

[e] ITT – All PFS events included regardless of stopping randomised therapy or subsequent therapy.

[f] PDT – Censor patients who receive subsequent anti-cancer therapy prior to progression at latest prior visit.

[g] DISC – Censor patients who prematurely discontinue randomised therapy due to toxicity or other, non-progression related reasons at latest prior visit.

[h] MV – Censor patients who progress, or die (in the absence of progression), after two or more missed visits at latest prior visit.

[i] ALL – Censor patients who are censored in either PDT, DISC or MV at earliest censoring time.

therapy ($p = 0.001$) but not when censoring for missed visits. There was no evidence that the association between the ITT HR and the HR based on alternative censoring definitions was dependent on sample size, study type (substitution versus add-on) or tumour type.

3.3. *Interval censoring*

In any clinical trial each patient's 'true' progression time is known only to have occurred at some time between two assessments. Despite the presence of multiple published techniques to perform analyses of such interval censored data[24,27,28] in practice, clinical trial data are not routinely analysed using interval censored analysis (ICA) approaches. We performed a simulation study to compare the performance of an ICA to the log-rank test. The findings presented here are a summary of the more extensive analysis by Sun.[29] As shown in Table 4, when the assessment times were similar between the two treatment arms, both the log-rank test and the ICA analysis provided unbiased results. However, when patients were assessed twice as often in the control arm, the log-rank test performed poorly as expected (with a very high type I error rate and biased estimate of treatment effect) while the ICA approach remained remarkably robust.

4. Discussion

This paper presents three key findings based on the work of the PhRMA PFS WG: the need for and benefit of independent radiology review, appropriate censoring and missing data conventions, and alternative analysis methods. The goal of this exercise was not to expand the use of PFS as a primary end-point, but to increase the confidence in the results of clinical trials which use PFS as a primary end-point by identifying the optimal methodological approaches.

The re-analysis of 28 oncology clinical trials explored the conventions surrounding censored observations and revealed that the most extreme application of censoring rules led to approximately 25% of the events being excluded (censored), and that excluding PFS events from analysis due to censoring

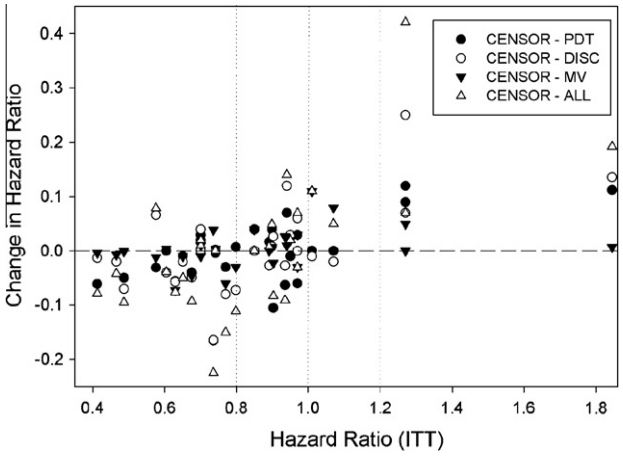

Fig. 1 – Change in ITT Hazard Ratio from applying censoring rules in analysis. Each point represents 1 of the 28 trials analysed using one of the 4 censoring rules (i.e. 112 points in total). Change in hazard ratio equals HR from censored analysis minus HR from ITT analysis.

tends to be anticonservative in the sense that it tends to result in a larger estimate of the treatment effect (i.e. further away from a HR of 1) than that from the ITT analysis.

The ITT analysis requires follow-up of all non-progressing patients, including those who received other anti-cancer therapy. Clearly, the decision to start new therapy is likely to be related to the patient's current prognosis or ability to tolerate therapy, and censoring the patient, either explicitly in the analysis, or implicitly by not collecting progression data from subsequent assessments, can lead to bias.[13,17] Since it cannot be known whether informative drop-out will occur, the most prudent (and conservative) approach would be to seek consent to follow all patients to objective progression, to be consistent with the ITT principle and to permit an ITT analysis to be performed.[7] As subsequent assessment data will likely have limitations, it is appropriate to question whether the use of anti-cancer therapy prior to progression in part contributed to any advantage for the experimental therapy and therefore, an analysis censoring such patients should be performed as a sensitivity analysis.

The analysis in the companion paper by Amit of 27 clinical trials which employed BICR demonstrated the strong correlation in HRs estimated by BICR and local evaluation (LE). This is consistent with the findings of others,[19,26] and indeed with similar reports from other therapeutic areas.[30,31] This analysis provides strong evidence for the reliability of the treatment from the local evaluation of progression. One can conclude from this analysis that BICR and LE are likely measuring the same phenomenon. Neither of these measures should be considered a gold standard; the BICR lack data from patients who progress by LE (and hence are informatively censored in any analysis), whereas the LE could be subject to real or inadvertent bias. At the overall study level, however, multiple analyses have now demonstrated a highly consistent estimate of the trial-level HR for the two approaches.

A sample based method for BICR has been proposed with well characterised operating characteristics. This method can directly measure the differential and directional discordance which can indicate a bias in the investigator assessment of progression. Extensive simulations explored the extent to which discordance due to random measurement

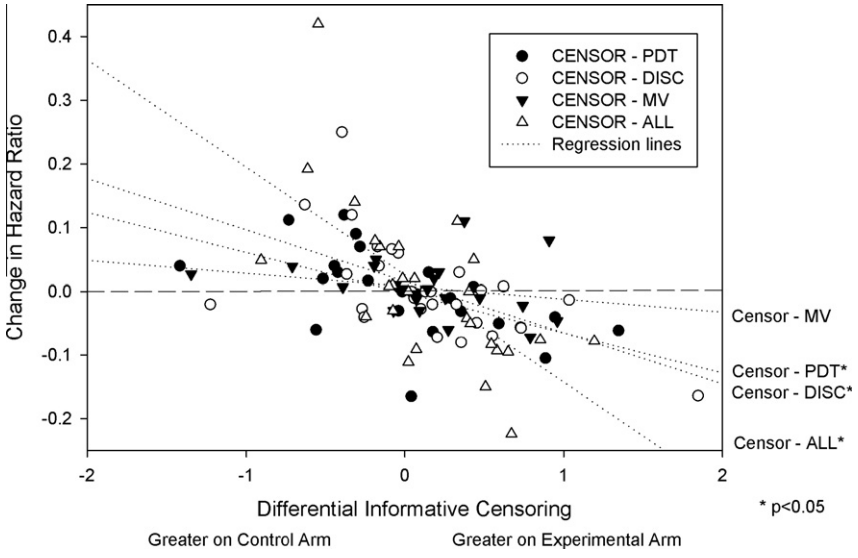

Fig. 2 – Factors associated with differences between ITT HR and censoring analyses HR. Differences in extent of informative censoring. Change in hazard ratio equals HR from censored analysis minus HR from ITT analysis. Censoring event rate ratio (CERR) equals event rate (no. of events/total patient follow up) calculated in period after censoring (due to MV, PDT, DISC or ALL) divided by event rate before censoring for each treatment arm. Differential informative censoring(*x* axis) = log ratio (exp:control) of within arm CERR. Each point represents 1 of the 28 trials analyzed using one of the 4 censoring rules (i.e. 102 points, 10 missing due to non-estimable CERR). Regression line is simple linear regression within each group, statistical analyses test whether the slope is zero and hence whether change in hazard ratio is related to the extent of differential informative censoring.

Table 4 – Comparison of log-rank test and interval censored analysis approach with unequal between arm assessment frequency with true HR = 0.6.

Event rate	Log-rank		Interval censored		True[a]
	T1 error[b] (%)	HR	T1 error[b] (%)	HR	T1 error[b] (%)
Equal assessment between arms[c]					
Constant	2.7	0.60	2.8	0.60	2.7
Differential assessments between arms[d]					
Constant[e]	35.0	0.55	2.3	0.61	2.4
Increasing[e]	59.2	0.53	2.6	0.61	2.7
Decreasing[e]	32.3	0.55	2.8	0.60	2.5

[a] True is a benchmark assuming exact PFS time known.
[b] T1 Error: proportion of simulations where 1-sided $p < 0.025$.
[c] Equal assessment: patients assessed every 8 weeks in both groups.
[d] Differential assessment: patients are assessed every 8 weeks and 16 weeks in the control and experimental group respectively.
[e] Constant, increasing and decreasing event rates correspond to exponential, weibull 1.5 and weibull 0.67 distributions respectively.

error exists, and demonstrated that if this discordance is equal between treatment groups, it does not result in a biased estimate of treatment effect. However adding additional variation to the progression time, which reflects poor study conduct, does result in a modest attenuation of the treatment effect (HR closer to 1.0). The presence of variability should not lead to patients being assessed more frequently than in routine clinical practice; as long as patients are assessed at a frequency that is no more than half of the control group median, there is a <2% loss in power.[20,29] Rather, an emphasis on accurate assessment is important.

The primary finding from the simulation studies evaluating interval censoring methods was that, even in the presence of extreme imbalance of visit schedules between treatment groups, the method developed by Finklestein[24] is robust to these deviations. The availability of this methodology should provide researchers with the confidence that a statistical tool exists to identify the benefit of therapy despite the presence of unscheduled assessments.

There are limitations to the findings presented. First, the sets of trials used to evaluate BICR and the various censoring methods were not selected at random, but were those contributed by the participating companies/institutions. However in both cases the set of clinical trials selected was not limited to certain tumour types, and considered to be representative and of adequate size. Secondly, the simulations conducted during the development of the sample based BICR rely on certain assumptions to mimic the system of evaluating PFS using BICR as closely as possible. The working group made every effort to be transparent with the simulations assumptions and methods.

A future focus of the PFS WG will be to determine objectively the presence (or lack thereof) of clinical benefit related to the PFS end-point, by documenting the consequences of progression in terms of deterioration in symptoms, quality-of-life and time-to-death.

4.1. *Recommendations*

The following recommendations for the design, conduct and analysis of trials using PFS as the primary end-point are made based on the research conducted by the WG.

- In clinical trials designed to establish the treatment effect using PFS as the primary end-point, the focus should be on the estimation of between-arm treatment effect as assessed through the HR, with the awareness that the assessment of individual patient progression times is subject to measurement error.
- The ITT principle should be employed in the primary analysis; patients should be followed until progression even if they have discontinued treatment or started a new anti-cancer therapy.
- Based on the strong correlation in HR between BICR and LE (Amit [25]), independent review of progression events should not be required in double blind trials.
- In open label trials or double blind trials with partial unblinding due to toxicity, a sample based BICR audit should be evaluated.
- When a sample-based BICR is employed, discordance between LE and BICR should be summarised by treatment group. If discordance between treatment groups differs meaningfully [25], then the degree of directional discordance should be considered and a full BICR of all subjects in the trial may be warranted.
- Because increased variation in the estimation of progression times can attenuate the treatment effect, rigorous training and monitoring should be in place at all investigational sites.
- Patients should be assessed at the same frequency in each treatment arm and interval censoring methods should be included as a sensitivity analysis. Once validated software is widely available, consideration should be given to the use of ICA methods to replace the log-rank test and Cox regression[32] as the primary tool for analysing PFS data.

Conflict of interest statement

Stone A.M., Bailey-Iacona R. – Employment and stock ownership, AstraZeneca.

Bushnell W., Amit A. – Employment and stock ownership, GSK.

Denne J. – Employment and stock ownership, Eli-Lilly.

Sargent D.J. – Consulting on design and analysis for multiple companies, none of which presents a conflict of interest with the work presented herein.

Chen C. – Employment and stock ownership, Merck.

Helterbrand J. – Employment and stock ownership, Genentech.

Williams G. – Consulted widely with many different pharmaceutical companies paid on an hourly basis but that none of these relationships represent a conflict with regard to the topic discussed in this paper.

Acknowledgements

The authors would like to acknowledge the contribution of the following members of the PhRMA PFS Expert Team: Pete Barker, AstraZeneca: Hans Burger, Roche: Paul Bycott, Pfizer: Bill Capra, Genentech: Michelle Casey, GSK: Bee Chen, Novartis: Tai-Tsang Cheng, BMS: Steve Dahlberg, Amgen: Gary Gordon, Abbott: Alison Goudie, AstraZeneca: Shenyang Hong, MedImmune: William John, Eli-Lilly: Richard Kay, RK Statistics: James Love, Boehringer Ingelheim: Frank Mannino, GSK: Ron Menton, Wyeth: Dmitri Pavlov, Pfizer: Jane Qian, Abbott: Martin Roessner, Sanofi-Aventis: Satrajit Roychoudhury, Novartis: Nicola Schmitt, AstraZeneca: Hongliang Shi, Takeda: Zhenming Shun, Sanofi-Aventis: Harry Staines, Boehringer Ingelheim: Xing Sun, Merck: Tao Wang, Pfizer: Xinju Wei, Schering-Plough.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ejca.2011.02.011.

REFERENCES

1. *FDA Oncologic Drugs Advisory Committee: endpoints for colorectal cancer regulatory approval*. http://www.fda.gov/ohrms/dockets/ac/04/transcripts/4037T2.DOC; 4th May, 2004.
2. *FDA guidance for industry clinical trial endpoints for the approval of cancer drugs and biologics*. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071590.pdf; May 2007.
3. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009;**45**:228–47.
4. O'Shaugnessy JA, Wittes RA, Burke G, et al. Commentary concerning demonstration of safety and efficacy of investigational anticancer agents in clinical trials. *J Clin Oncol* 1991;**9**:2225–32.
5. Johnson JR, Williams G, Pazdur R. End points and United States Food and Drug Administration approval of oncology drugs. *J Clin Oncol* 2003;**21**:1404–11.
6. *CHMP Guideline on the evaluation of anticancer medicinal products in man*. <http://www.ema.europa.eu/pdfs/human/ewp/020595en.pdf>.
7. *Methodological considerations for using progression-free survival (PFS) as primary endpoint in confirmatory trials for registration*. http://www.emea.europa.eu/pdfs/human/ewp/2799408en.pdf; January 2008.
8. *FDA cancer drug approval endpoints workshops*. <http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/CancerDrugs/ucm094586.htm>.
9. *FDA Oncologic Drugs Advisory Committee: Endpoints in clinical cancer trials and endpoints in lung cancer clinical trials*. <http://www.fda.gov/ohrms/dockets/ac/03/transcripts/4009T1.DOC>; 16th December 2003.
10. Broglio KR, Berry DA. Detecting an overall survival benefit that is derived from progression-free survival. *J Natl Cancer Inst* 2009;**101**:1642–9.
11. Fleming TR, Rothmann MD, Lu HL. Issues when using progression-free-survival when evaluating oncology products. *J Clin Oncol* 2009;**27**:2874–80.
12. Di Rienzo AG. Non parametric comparison of two survival-time distributions in the presence of dependent censoring. *Biometrics* 2003;**59**:497–504.
13. Rothmann M, Kati K, Lee KY, et al. Examining the extent and impact of missing data in oncology clinical trials. In: *Proceedings of the joint statistical meetings*; 2009, p. 4014–9.
14. Bhattacharya S, Fyfe G, Gray RJ, et al. Role of sensitivity analyses in assessing progression-free survival in late-stage oncology trials. *J Clin Oncol* 2009;**27**:5958–64.
15. Flyer P, Hirman J. Missing data in confirmatory clinical trials. *J Biopharm Stat* 2009;**19**:969–79.
16. Williams G, He K. Operational bias in assessing time to progression (TTP). *Proc Am Soc Clin Oncol* 2002 [abstr 975].
17. Carroll KJ. Analysis of progression free survival in oncology trials: some common statistical issues. *Pharm Stat* 2007;**6**:99–111.
18. Dancey JE, Dodd LE, Ford R, et al. Recommendations for the assessment of progression in randomised cancer treatment trials. *Eur J Cancer* 2009;**45**:281–9.
19. Dodd LE, Korn EL, Freidlin B, et al. Blinded independent central review of progression-free-survival in phase III clinical trials: Important design element or unnecessary expense? *J Clin Oncol* 2008;**26**:3791–6.
20. Stone AM, Wheeler C, Carroll KJ, et al. Optimizing randomised phase II trials assessing tumour progression. *Contemp Clin Trials* 2007;**28**:146–52.
21. Wang Y, Sung C, Dartois C, et al. Elucidation of relationship between tumor size and survival in non-small-cell lung cancer patients can aid early decision making in clinical drug development. *Clin Pharmacol Ther* 2009;**86**:167–74.
22. Zhao B, James LP, Moskowitz CS, et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. *Radiology* 2009;**252**:263–72.
23. Hertz-Picciotto I, Rockhill B. Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics* 1997;**53**:1151–6.
24. Finkelstein D. A proportional hazards model for interval-censored failure time data. *Biometrics* 1986;**43**:645–854.
25. Amit O, Mannino F, Stone AM, et al. Blinded independent central review of progression in cancer clinical trials: results from a meta-analysis. Eur J Cancer.
26. Tang PA, Pond GR, Chen EX. Influence of an independent review committee on assessment of response rate and progression-free survival in phase III clinical trials. *Ann Oncol* 2010;**21**:19–26.
27. Sun J, Zhao Q, Zhao X. Generalized log-rank tests for interval-censored failure time data. *Scand J Stat* 2005;**32**:49–57.
28. Zhao Q, Sun J. Generalized log-rank test for mixed interval-censored failure time data. *Stat Med* 2004;**23**:1621–9.
29. Sun X, Chen C. A comparison of commonly used analysis methods for interval-censored time-to-event data from clinical trials. *Stat Biopharm Res* 2010;**2**:97–108.

30. Mahaffey KW, Harrington RA, Akkerhuis M, et al. Systematic adjudication of myocardial infarction end-points in an international clinical trial. *Curr Control Trials Cardiovasc Med* 2001;**2**:180–6.
31. Pogue J, Walter SD, Yusuf S. Evaluating the benefit of event adjudication of cardiovascular outcomes in large simple RCTs. *Clin Trials* 2009;**6**:239–51.
32. Cox DR. Regression models and life-tables. *J R Stat Soc* 1972;**34B**:187–220.