EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

13 December 2012
EMA/CHMP/27994/2008

Appendix 1 to the guideline on the evaluation of anticancer medicinal products in man

Methodological consideration for using progression-free survival (PFS) or disease-free survival (DFS) in confirmatory trials

Draft Agreed by Oncology Working Party	**September 2011**
Adoption by CHMP for release for consultation	15 December 2011
Start of public consultation	December 2011
End of consultation (deadline for comments)	31 May 2012
Agreed by Oncology Working Party and Biostatistics Working Party	November 2012
Adopted by CHMP	13 December 2012
Date for coming into effect	1 July 2013

Keywords	Cancer, malignancy, biomarker, targeted drugs, pharmacogenomics

7 Westferry Circus ● Canary Wharf ● London E14 4HB ● United Kingdom
Telephone +44 (0)20 7418 8400 **Facsimile** +44 (0)20 7418 8416
E-mail info@ema.europa.eu **Website** www.ema.europa.eu

Introduction

The use of progression-free survival (PFS) or disease-free survival (DFS) as endpoint in clinical efficacy trials presents several methodological issues which need to be addressed prospectively.

This appendix provides some general regulatory guidance on issues to consider relating to definitions, frequency and methods of assessment, ascertainment bias, handling of deviations and missing data, and radiology review. Guidance on the choice of primary endpoint, and appropriateness of using PFS/DFS as primary endpoint, is addressed in the main body of the guideline.

Endpoint definition

PFS is traditionally defined as the time from randomisation (or registration, in non-randomised trials) to objective disease progression, or death from any cause, whichever occurs first. DFS is defined as the time from randomisation (or registration, in non-randomised trials) to objective disease recurrence or death from any cause, whichever occurs first. The time of the progression or recurrence event is determined using the first date when there is documented evidence that the criteria have been met, even in situations where progression is observed after one or more missed visits, treatment discontinuation, or new anti-cancer treatment.

Whenever possible, the definition of progression should follow established response evaluation criteria (e.g., RECIST, WHO criteria, EBMT criteria, RANO). Clear definitions on non-radiologic criteria, if applicable, should be provided. Depending on the type of agent, the site and type of lesion, and the objectives of the trial, modified criteria might have become established in a specific situation and be considered to be more appropriate. For instance, additional objective clinical and biochemical or radiological criteria may be used to assess progression. In all cases, it is important that the criteria for definition of a progression event are as objective as possible, and that the definitions be clearly and prospectively defined in the protocol.

As PFS is defined as a composite of different events (e.g., new lesions, progression of existing lesions or disease, death), it is recommended to report separate analyses for individual types of events using descriptive summary tables and, where appropriate, competing-risks approaches to explore treatment effect on the various types of events.

For certain types of agents that might interfere with the methods of detection (e.g., anticancer agents that through different mechanisms of action could interfere with the contrast enhancement of lesions on imaging or some agents, such as immunomodulators, that do not only interfere with the contrast enhancement but can determine a peculiar pattern of response, e.g., apparent progression with enlargement of lesions followed by response) different methods or endpoints need to be considered.

A ‘time to event’ approach is appropriate to define an endpoint for statistical analysis. Other approaches based on proportion of patients experiencing an event at a particular timepoint might have merit in some cases but have limitations and a sponsor considering use of a fixed timepoint approach is recommended to consider CHMP Scientific Advice.

Data capture

Information collected in CRF should be in full accordance with the protocol and should focus on the data necessary to implement all the planned analysis, in accordance with ICH topic E9.

Interval-detected progressions

Generally, the exact time of progression will not be known. Instead, progression will be known to have occurred during a particular time interval, e.g., between two follow-up visits to detect if a progression has occurred. Generally, for the purpose of the primary analysis, interval censoring is ignored and the analysis is carried out on the times of detected recurrence.

Deviations from the timings of scheduled evaluation will occur in practice, with progression events being detected during the interval (interval-detected progressions) as opposed to at the time of the scheduled follow-up visit (screening-detected progressions). The mixture of the two types of progressions is of concern due to the potential for introducing a detection bias leading to incorrect conclusions about the treatment differences. For instance, if there are important differences in terms of toxicity or symptom palliation between treatment groups, progressions will be detected earlier for the treatment group with higher toxicity or symptoms. Investigators may also examine more frequently patients on the control arm (or delay evaluations for patients on experimental arm) in view of an inherent bias in favour of the experimental arm. Clinical trials that are not adequately blinded are particularly at risk of ascertainment bias when a change in the clinical status of the patient prompts an unscheduled assessment of disease status.

Problems of bias due to unscheduled evaluations should be minimised by proper trial design and conduct. Clinical trials should be adequately blinded whenever possible. The schedule of assessment should be carefully considered. If the time between scheduled evaluations is short relative to the median time to progression, there will be few interval-detected progressions and unscheduled recurrences will not be a major concern.

If progression is detected during an unplanned evaluation, between two scheduled evaluations, for the purpose of the primary analysis, the date of progression should be assigned based on the documented time of progression and not, for example, based on scheduled time of evaluation. This approach is preferred as it is closer to the intention-to-treat principle and may be less prone to informative censoring (Stone et al., 2011). Alternative analyses based on scheduled time of evaluation and using interval censoring should be included as supportive analyses. A descriptive analysis about compliance with scheduled evaluations by treatment arm should be provided.

Various approaches have also been proposed on how to handle unexpected differences in the patterns of follow-up in supportive analyses, aiming to minimise bias whilst preserving accuracy of the estimated time of progression, and consideration should be given to the pre-specification of such analyses. *Post hoc* data analyses are, however, of limited value in compensating for detection bias.

Informative censoring

Observation of the PFS event for all randomised subjects will rarely be available in practice, leading to *censored* survival data. Commonly used methods for comparing the survival times between groups are only valid if the censoring is not related to any factors associated with the actual survival time (i.e., the censoring is said to be “uninformative”). Conversely, *informative* censoring may lead to incorrect conclusions about the extent of the treatment difference. There is no satisfactory way to correct for informative censoring, which should be minimised by adequate design and conduct of the study. Similar considerations apply to analysis of OS data.

The assumption of uninformative censoring generally holds for “administrative” censoring (patients with complete follow-up, for whom no event has been observed at the time of data cut-off for the analysis). However, this may not be the case for the set of patients with incomplete follow-up at the time of data cut-off for the analysis (“premature” censoring). For this set of patients the assumption of uninformative censoring should be examined systematically, using standard survival analysis approaches (e.g., investigating the time-to-censoring e.g. using Kaplan-Meier curves; examining patterns of censoring across covariates; exploring the association between censoring and covariates that may be associated with the PFS event, such as tumour burden, and any differences across treatment arm). The extent of informative censoring should be discussed in the clinical trial report.

Non-compliance with protocol-treatment may occur, for example, when subjects receive the wrong study medication (or none at all), withdraw from treatment prior to scheduled completion or change treatment before evidence of progression.

Events of withdrawal from study therapy prior to adjudicated progression are likely to be informative and the adequacy of censoring these events in the statistical analysis should always be questioned. There is no way to handle this problem that is optimal for all situations, but the principles of intention-to-treat should be followed as far as possible when defining the analysis set for the primary analysis of PFS/DFS. In particular, for all randomised patients, outcome data should be collected according to the intended schedule of assessment and the date of progression or recurrence should be assigned based on the time of the first evidence of objective progression or recurrence regardless of violations, discontinuation of study drug or change of therapy. If, for a particular study, a different approach is considered to be more appropriate, a justification is expected and CHMP Scientific Advice agreement is recommended at the planning stage.

Even if foreseen in the study protocol, it may at times be difficult to collect reliable data on progression for patients withdrawn from study therapy. For this, and for other reasons, there is a need to predefine and justify methods for handling missing data, including rules of censoring. These methods should be chosen so as to minimise bias and loss of information, while being adequate for the aim of the trial. This may include approaches that consider withdrawal or change of therapy prior to adjudicated progression / recurrence as events in an analysis of PFS/DFS. Potential biases should always be addressed and sensitivity analyses should be undertaken using different approaches. Supportive analyses may include for example an approach that assigns the progression date to the date of the scheduled clinic visit, interval-censored analysis, single time point analysis, with progression being assigned at one pre-specified time after randomisation, sensitivity analyses assuming that censored subject are at high-risk or low-risk of an event.

Primary and sensitivity analyses

For the purpose of the primary analysis, the date of progression should be assigned based on the documented time of progression and not, for example, based on scheduled time of evaluation.

The strategy for the primary analysis should be clearly written before the trial start. It is important that due consideration is given to the statistical analysis plan, including sensitivity analyses to address the handling of deviations and missing data, at the planning stage of the trial.

In blinded trials, conducting a blind review at the end of the trial may offer a valuable opportunity to review the data handling methods selected and the range of analyses proposed so that unforeseen issues can be addressed. Whilst a review at the end of an open-label trial can be conducted (before applying the randomisation code to the datasets), resulting amendments or updates to the statistical analysis plan are likely to be viewed with some scepticism because it is difficult to exclude the possibility that they are data-driven. For such trials, utmost diligence is required when writing the study protocol and statistical analysis plan as amendments to important aspects of the analysis made in the knowledge of accruing data would give rise to concern. How to deal with and document these data analysis issues should follow general guidance provided in the note for guidance on statistical principles for clinical trials (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use 1998). Caution should also be applied for trials that are conducted in a nominally blinded manner, but where adverse event profiles lead to functional unblinding.

At present, from a regulatory perspective there are several possible approaches that can be recommended for sensitivity analyses (see, e.g., Food and Drug Administration, 2007). The range of sensitivity analyses should be sufficient to demonstrate that the trial results are robust and will depend on the clinical situation and expected nature of the trial data observed (e.g., patterns of patient withdrawals). Any differences in conclusions from the range of analyses presented will need

to be explained. The importance of different analyses and analysis sets will also depend on the design of the trial (superiority or non-inferiority).

Sensitivity analyses should be planned to address any important assumptions in the methods used, including handling of deviations and missing data, uninformative censoring, proportional hazards, handling of unscheduled evaluations, as applicable.

Sensitivity analyses should be described in the protocol or the statistical analysis plan and any changes must be justified in the study report.

Interim analyses

Interim analyses are routinely employed in oncology trials to monitor safety, assess 'futility' and to consider whether there is sufficient evidence of efficacy to stop the trial early. The timing, objectives and conduct of interim analyses should always be justified in line with regulatory guidance, but in general monitoring of safety is supported and assessment of futility is not controversial. More challenging is the interim analysis designed to stop the trial early for demonstrated efficacy. Whilst interim analyses with this purpose are accepted in principle, there are particular considerations with PFS as primary endpoint and therefore interim analyses for PFS are not encouraged. If nevertheless these are deemed necessary and justified, the following specific issues should be addressed:

1. Datasets need to be sufficiently mature to ensure robust conclusions, about the ITT trial population and about subgroups of particular importance (internal consistency)
2. Often there will be only one confirmatory (pivotal) trial with resulting requirements for the level of evidence to be available
3. Data on OS, on safety and on other secondary endpoint might be immature or insufficient for a regulatory decision on benefit-risk.

Hence the timing and objectives of an interim analysis should not be planned considering only the detection of statistical significance in PFS. A proper justification will include consideration of the maturity of PFS and OS data and evidence available in subgroups, on safety and on secondary endpoints (see also *Follow-up and treatment after progression*, below). These considerations may render an interim analysis impractical.

Frequency and methods of assessment

The methods and frequency of tumour assessment should be the same across study arms, even when treatment cycles are of different lengths.

Evaluation of PFS requires that all sites of possible disease specific to that tumour type be assessed at baseline and that involved sites be systematically assessed during follow-up together with other sites, as clinically and radiologically indicated, ideally using the same methods. Similarly, evaluation of DFS may require that likely sites of disease be systematically assessed at follow-up assessment. Where there are reasons to suspect that certain drugs might influence the pattern of metastasis, additional relevant sites should be pre-specified for systematic assessment.

The frequency of assessment should therefore be adequate to detect the expected treatment effect. The timing of the assessment and the optimal frequency for assessing progression needs to be determined on a trial by trial basis, taking into account the aims of the trial and the treatment schedules and the specific pattern of progression of the disease. For example, it is expected that the first visit should be timely if the median PFS is short. A balance needs to be found between the need to assess progression precisely and the need to minimise exposure of patients to invasive and resource-intensive diagnostic procedures. Whilst increasing frequency reduces the chance for assessments between scheduled visits, it will increase the chance that scheduled visits will be missed entirely especially if the frequency is inconsistent with routine clinical practice. From a

statistical perspective, increasing frequency of assessment does little to increase statistical power unless the rate of progression is very rapid (Wallenstein et al., 1993; Stone et al., 2007)

Adherence to protocol-defined schedules is essential and deviations should be reported. Compliance with the visit schedule should be descriptively investigated at the time of the analysis and any impact on the trial results should be explored.

Blinded independent central review (BICR)

Evaluation of progression may be subject to measurement error, particularly in advanced disease where many lesions need to be followed. In general, efforts should be made to minimise the measurement error. If significant measurement error still occurs despite every reasonable effort to avoid it, from a regulatory perspective, this may still not be a major concern when assessing relative efficacy provided that it occurs equally across treatment arms and that the effect of treatment is sufficiently large. However, if the measurement error differs across treatment arms this may lead to difficulties in interpreting the results of the analysis. Similarly, if the treatment effect is small or moderate, a large measurement error may hamper the benefit-risk assessment. Thus, every reasonable effort should be made to minimise the measurement error through adequate standardisation of methods and training of investigators conducting the local evaluation.

Evaluation of disease progression by investigators can be subject to systematic bias in favour of one of the treatment arms, leading to incorrect treatment comparisons. For these reasons adequate masking techniques should be used whenever possible. Investigator bias is generally not an issue in properly double-blinded randomised trials. However, cancer drug trials are notoriously problematic when it comes to blinding due to the characteristics effects of different drugs. Indeed, frequently it may be impossible to mask treatment assignment completely, for example, due to the different toxicity profiles of the treatments. Studies against best care may be at particular risk of this kind of bias.

One strategy to try to detect and reduce this bias is to conduct a complete BICR of all relevant data for all patients. This strategy is recommended when important investigator bias is expected or in case of moderate expected treatment size of effect. BICR will be more meaningful in situations where the majority of events will be captured based on imaging as opposed to clinical progression.

However, if important investigator bias is present, even complete BICR of progression may still not prevent informative censoring because patients are taken off protocol at the time of locally evaluated progression and no further laboratory, imaging or clinical evaluation data may be available after this time point. One way to obviate this is to conduct real-time BICR. Another way to lessen this problem is to collect additional scans after locally designated progression. However, this may not be practical as patients may be lost to additional follow-up after local progression (Dodd et al. 2008).

Bias in the local investigator assessments can be investigated by looking at the direction of any discordance between investigator assessments and BICR. Statistics proposed for this purpose include early discrepancy rate (EDR) and late discrepancy rate (LDR), which are based on the frequency that the local evaluation declares progression respectively earlier or later than BICR.

One possible strategy to avoid complete BICR in certain situations is to perform BICR based on a sample only (“audit”). If a review of discordance statistics supports the absence of any bias in the local investigator assessment a complete BICR might be avoided. If bias cannot be excluded based on the audit, a complete BICR can subsequently be implemented to provide a basis for another analysis (Amit et al. 2011). Other strategies include blinded local or country-specific radiology reviews. As there is currently no extensive experience on the practical implementation of these approaches, regulatory guidance on the appropriateness of any such approaches should be sought on a case-by-case basis before implementation to discuss, in particular, ensuring integrity of the

study, how the sample will be generated and the statistics and metrics to be used for deciding whether or not an important directional discordance can be excluded.

In general, where complete BICR is appropriate, the primary analysis can be planned to be based on the outcome assigned through independent evaluation. If important investigator bias can reasonably be excluded, investigator evaluation can be planned to be used for the primary analysis. Regardless of the strategy, the role of the outcome assigned through BICR, and any decision rules regarding the extent of BICR, should be pre-specified in the protocol.

In general, the confidence in the quality of the trial will increase if the trial results from the BICR do not differ from the investigator assessments to any important degree.

The procedures for independent review shall be defined prospectively and described in the clinical trial documentation.

Size of effect

The size of effect should be quantified by plotting the estimates of the survivor functions for PFS, estimating the hazard ratio, estimating median time-to-event and other percentiles (e.g. upper quartile, lower quartile), and estimating the percentage of patients event free at particular time-points (e.g. % patients event free at 1-year), based on semi-parametric procedures. Although from a clinical perspective the median PFS is considered the preferred summary measure of the location of the distribution of PFS survival times, over-reliance on differences in medians should be avoided because this will generally be less informative than considering the survival curve as a whole. In any case, the choice of the summary measure should be justified and pre-specified.

Because the non-parametric estimates of the survivor functions are step-functions, parametric proportional hazards modelling and other “smoothing” techniques may also be useful in exploratory analyses.

The clinically relevant difference to be detected is often not a trivial issue in case of PFS, since this endpoint can be largely driven by laboratory, radiological or clinical evaluations that are not of immediate clinical relevance to patients. The difference needs to be justified prospectively based on clinical and epidemiological grounds, including for example, demonstration of surrogacy for OS, expected effect in terms of symptoms, change in treatment, emotional impact, or health-related quality of life.

At the time of analysis, supportive descriptive summary tables and competing risks analyses should be available for the different component of the composite PFS endpoint (e.g., new lesions, increase in size, death and cause-related deaths) and on supportive endpoints such as OS, symptom control, health-related quality of life (as appropriate).

Follow-up and treatment after progression

A large effect in terms of PFS is generally expected to be associated with an effect on OS. If this is not the case, a rational explanation should be provided.

When comparing treatments in terms of PFS it is important to consider that treatment with an experimental agent, even if advantageous in terms of PFS, may however be associated with poorer OS. This may be due, for instance, to long term-toxicity, different resistance profiles to treatments used after progression, or to biological changes leading to increased metastatic potential.

Thus, whenever possible when PFS is the primary endpoint, complete follow-up of all patients should be available until death and there should be sufficient reassurance that there is no detrimental effect in terms of OS (see main body of the guideline). In case of further treatments, and particularly where lack of efficacy of further treatments might be a concern, outcome to subsequent treatments in terms of objective response rate, PFS after next line of treatment (PFS2) should also be available where practicable. PFS2 is defined as the time from randomisation (or

registration, in non-randomised trials) to second objective disease progression, or death from any cause, whichever first. Patients alive and for whom a second objective disease progression has not been observed should be censored at the last time known to be alive and without second objective disease progression. In situations where OS and PFS2 cannot reliably be determined, it may be possible to rule out significant lack of efficacy of further treatments by looking at outcome in terms of end-of-next-line-treatment. For this analysis, an event is defined as end or discontinuation of next-line treatment, second objective disease progression, or death from any cause, whichever first.

One-way cross-over to the experimental arm after progression is likely to hamper any subsequent comparisons in terms of OS and other long-term secondary endpoints. Thus, this type of cross-over should generally be avoided in order to meet the objectives of the trial. If nevertheless it is considered necessary, there should be sufficient confidence that the available data in terms of PFS, OS, and any other important secondary endpoints will be convincing enough from a scientific and regulatory point of view to meet the objectives of the trial and to ensure that adequate conclusions can be drawn. In such situations, the analysis of OS can be done on the basis of planned secondary analyses or planned co-primary analyses.

References

Amit, O., et al. (2011), 'Blinded independent central review of progression in cancer clinical trials: Results from a meta-analysis', *Eur J Cancer*.

Dodd, L. E., et al. (2008), 'Blinded independent central review of progression-free survival in phase III clinical trials: important design element or unnecessary expense?', *J Clin Oncol,* 26 (22), 3791-6.

Food and Drug Administration (2007). Guidance for Industry - Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use 'E9: Statistical Principles for Clinical Trials', (ICH E9) http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf.

Stone A, Wheeler C, Carroll K et al. Optimizing randomized phase II trials assessing tumor progression. Contemporary Clinical Trials 2007;28:146–152

Stone, A. M., W. Bushnell, et al. (2011). Research outcomes and recommendations for the assessment of progression in cancer clinical trials from a PhRMA working group. Eur J Cancer **47**(12): 1763-1771.

Wallenstein S, Wittes J. The power of the Mantel–Haenszel test for grouped failure time data. Biometrics 1993;49:1077–87