EUROPEAN JOURNAL OF CANCER 47 (2011) 1759–1762

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Editorial Comment

Is it time to abandon complete blinded independent central radiological evaluation of progression in registration trials?

Francesco Pignatti[a,*], *Rob Hemmings*[b,d], *Bertil Jonsson*[c,e]

[a]*The European Medicines Agency (EMA), London, United Kingdom*
[b]*Medicines and Healthcare Products Regulatory Agency (MHRA), London, United Kingdom*
[c]*Läkemedelsverket, Uppsala, Sweden*

ARTICLE INFO

Article history:
Available online 3 June 2011

Few subjects have generated as much regulatory discussion as the choice of primary endpoints in phase III trials in the advanced or metastatic cancer setting. Acknowledging that overall survival (OS) remains the most objective and clinically convincing endpoint to support a favourable decision on the benefit-risk balance and to change clinical practice, progression-free survival (PFS) has also been proposed as the primary endpoint to support drug approval. Although the interplay between PFS and OS remains unknown for most agents, and particularly for new ones, a primary rationale for using PFS is that this endpoint could be considered as a clinical benefit endpoint in itself, provided the treatment effect is sufficiently large. From the perspective of drug developers, the interest in PFS is because of the expectation that treatment effect will be numerically larger and quicker to observe, compared to OS. Furthermore, determination of PFS is not confounded by subsequent therapy and the increased interest in PFS as a primary endpoint likely reflects the fact that more active treatments have become available, making the OS comparison more complex due to the difficulty to control for the effect of subsequent therapies on OS. Regulators have recognised wider acceptance of PFS as the primary endpoint in registration trials, as evidenced by emerging guidance documents.[1,2] Nevertheless, the general acceptance of PFS as a primary endpoint from a regulatory perspective is still controversial because the clinical relevance of this endpoint is often debated, even in indications where the conventional response criteria (RECIST) are deemed appropriate.[3,4]

The use of PFS is also complicated by several non-trivial methodological issues associated with measuring progression in an unbiased way. In particular, attribution of progression based on local evaluation (LE) may be subject to measurement error and bias, so that regulators around the world have traditionally requested complete (100% of cases) blinded independent central radiological evaluation (BICR) of progression to address these issues.[1,2] This requirement has been relaxed somewhat for double-blinded trials although the real effectiveness of the blinding for cancer drugs can always be questioned. Whilst assessment of progression is associated with different degrees of subjectivity in different indications, this has not been systematically used to identify high-risk situations that demand BICR. Indeed, BICR has become a standard for most registration trials with PFS.

* *Corresponding author:* Address: 7 Westferry Circus, Canary Wharf, London E14 4HB, United Kingdom. Tel.: +44 20 7523 7031, mobile: +44 78 8426 5946; fax: +44 20 7418 8613.
E-mail address: francesco.pignatti@ema.europa.eu (F. Pignatti).
[d] EMA Committee for Medicinal Products for Human Use (CHMP) Member, Chairperson of the CHMP Scientific Advice Working Party.
[e] Chairperson of the CHMP Oncology Working Party.
0959-8049/$ - see front matter
doi:10.1016/j.ejca.2011.05.009

In a recent issue of the EJC, Stone et al. and Amit et al. put forward the view of a pharmaceutical industry consortium (PhRMA PFS Working Group) questioning the need for systematic complete BICR to ensure unbiased treatment comparisons.[5,6] BICR adds significant complexity and cost to the trial and the issue raised is part of a wider debate on the added value of BICR and the search for more efficient strategies to ensure that reliable PFS data are available at the time of licensing.[7] Indeed, it can be agreed with the authors that the primary role of BICR is usually to check for absence of bias in the LE assessment. The authors report the results of simulation studies and a meta-analysis of company trials where the authors found that on average, BICR did not lead to different conclusions than local investigator's evaluations, emphasising that most of the discordance is due to measurement error and not to bias. The important distinction between measurement error and bias should be recognised by trial sponsors and regulators alike. The authors suggest that the risk of bias could be effectively measured based on an independent review of only a random sample of patients ("audit"). They develop the concept of differential discordance as the key framework for evaluating bias in the local evaluation. This is based on the early discrepancy rate (EDR) and late discrepancy rate (LDR), which are based on the frequency that LE declares progression, respectively, earlier or later than BICR. If bias cannot be excluded based on the audit, a complete BICR can subsequently be implemented to provide a basis for another analysis.

The proposed audit to identify trials deserving full BICR is certainly of interest and the measures of discordance proposed, LDR and EDR, are intuitive. However, when it comes to detection and correction of bias in practice, the operational aspects and efficiency of the various statistics proposed remain to be fully explored and validated. For example, using the proposed threshold for differential discordance rate, the method is expected to predict a potential difference of hazard ratios of 20–30% between BICR and local investigator.[6] Smaller differences of, say, 15% might go undetected. In terms of regulatory decisions, such differences may alter the balance of benefits and risks. This invites some caution on the side of regulators before the operational characteristics of this classifier are agreed as being generally applicable. Other criteria that could be explored would be to exclude possible implications for the hazard ratio of greater than 0.1, or no more than, say, 30% of the estimated treatment effect. Again, the criteria selected merit case-by-case consideration. It is possible that, along the lines proposed by the authors, differential discordance based on high-risk samples (rather than random samples), together with other risk factors could be more valuable in determining which trials need complete BICR, and the operational characteristics could perhaps be set according to the sample selected. The ideal, of course, would be prospective identification of such high-risk situations, to allow for prospectively planned, real-time BICR when needed, and avoid the problem of informative censoring associated with retrospective BICR.[7]

Methodological issues relating to PFS also include the use of 'censoring' to handle some patient withdrawals, missed visits, changes in treatment etc. Historically, little attention has been paid to whether this censoring is 'informative', as triallists and regulators might consider for missing data problems in other therapeutic indications, and this should be improved. The authors note that patients censored have approximately 30% greater risk for progression than those continuing follow-up as per protocol and so the possibility of bias introduced by the manner in which these data are handled is entirely credible. Stone and colleagues find some support for imputation based on the actual date of detected progression (so-called intent-to-treat method, ITT) as opposed to other methods, particularly in the context of superiority trials. This is in line with the position expressed in the current European Medicines Agency (EMA) guideline, preferring the 'ITT' approach as a starting point for analysis and inference. Furthermore, if data to facilitate this analysis are collected, all other analyses involving censoring are also possible as supportive sensitivity analyses.[8]

Stone and colleagues also encourage us to depart from the traditional way of analysing PFS data that ignores the fact that the progression event is never actually observed but only known to have occurred (or not) in the interval between two evaluations, where the length of the interval may actually vary for each interval and patient. A correct way to analyse these data is as interval-censored survival data. The note is not meant to be merely "academic". It is based on the finding that the interval-censoring approach is more robust to bias that occurs when investigators may choose to increase or decrease the frequency of evaluation depending on the treatment group. There is a risk, for example, that toxicity may prompt for more screening detected progression in one of the treatment arms, introducing a bias, sometimes in favour of the experimental treatment group. Also, worsening of symptoms may trigger more unscheduled evaluations in the control group but not in the experimental group due to a greater expectation by investigators for the experimental treatment, thus introducing a bias in favour of the experimental group. Ignoring this problem has been the norm for pharmaceutical companies in the regulatory setting where the time of detection of progression is used as the time of progression. In fact, the comparison made by the authors (Stone et al.[5], Table 4) is somewhat extreme as alternative approaches to implementing the log-rank test could be considered in the event of unbalanced visit schedules to reduce the problem described. In particular, it could be considered whether the timing of events on the treatment arm with the increased visit schedule could be determined based on the less frequent visit schedule from the other treatment arm (e.g. events on both arms are pushed back to the same point). This approach has its own complications but, under this scenario, the bias reported would not appear so extreme (though estimates of efficacy, e.g. median time to progression, would be prolonged, which is also disadvantageous). Notwithstanding this, the simulations found that using another statistical method in the first place may largely avoid the issue of bias due to unscheduled evaluations. Interval-censoring survival analysis has only recently become more widely available in the relevant statistical software owing to some of the technical complexities of this type of analysis.[9] Although there are a number of open methodological questions about the analysis of interval-censored data, these statistical approaches may soon become part of the standard

set of supportive analyses of PFS to assess the impact of detection bias due to unscheduled assessments, where these cannot be avoided.

Potential limitations to the exercise conducted by the Working Group relate to the extent of the simulations conducted, the inclusion and exclusion of trials from the meta-analysis, and the discussions of conclusions in general terms without consideration of limitations when applied in specific situations. In fact, whilst it is always possible to extend simulation work to include different scenarios, it can be agreed that the work is adequate for the 'proof of concept' of the authors' recommendations. The meta-analysis can necessarily include only those trials that are published, or those on which data can be made available from the trial sponsor. Again, there might be concern that 'failed' trials are less likely to be included and that trials might have 'failed' for reasons related to the questions under investigation here (e.g. high discordance between LE and BICR). Nevertheless, this potential concern is not considered critical. Of greater importance is whether the recommendations can be applied equally to trials in all therapeutic indications. In fact, it is strongly recommended to carefully consider how the recommendations can be applied to any given future development programme and mechanisms are in place for specific scenarios to be discussed with regulatory authorities.

In conclusion, is it time to abandon systematic complete BICR as the standard for assessing progression in registration trials that cannot be conducted as reliably double-blinded trials? Probably not yet, but we definitely agree that it should be possible to develop a more efficient risk-based approach. The proposed audit to evaluate trials deserving full BICR is certainly a step in the right direction. Provided that robust predictors of the risk of bias can be developed and validated and that the specific issues pertaining to assessment of progression in individual therapeutic indications are considered, complete BICR may soon be avoided in the majority of trials. Regardless of the strategy chosen for progression adjudication, sponsors should aim for a meaningful difference from the start, combined with adequate training and quality assurance. Sufficient data should be available from the early clinical development to ensure that the trial is conducted in the population most likely to respond. Complex and costly BICR have little to add in case of minute, clinically irrelevant, differences.

Further reflection is required to define situations where complete BICR should still be routinely planned, considering, for example, whether measurement error (e.g. inter-reader variability) is likely to be high (e.g. pancreatic cancer, mesothelioma) which might give rise to concerns over possible bias; the role of imaging in the assessment of progression; the choice of control and whether one-way crossover is permitted (e.g. studies versus best supportive care with the possibility to switch to experimental treatment at time of progression); whether the trial includes a small number of large investigator sites such that a problem of bias in one centre could have a large effect on the estimated treatment effect. BICR will be more meaningful in situations where the vast number of events will be captured based on imaging as opposed to clinical progression. A definitive list of considerations is not yet available (and is not considered in detail by Stone et al.) but this highlights the need for consideration of the role and importance of BICR on a case-by-case basis. Where LE is the preferred option, retention of scans to facilitate BICR if required is still foreseen as an important recourse for trial sponsors.

Whether we should be looking at PFS as primary endpoint in the first place is of course an entirely different and more fundamental question. Advocates of PFS argue that progression is associated with poorer long-term prognosis and has a very strong emotional impact. It exposes patients to the anxiety of treatment failure, change to less effective and possibly more toxic treatments, worsening symptoms and an overall deterioration of the quality of life. They also claim that to study PFS may be inevitable in situations where subsequent treatments may have an impact on OS. Sceptics on the other hand invoke the dubious clinical interpretation of this endpoint, which is based on radiological definitions that were intended for phase II trials, and stress the need for confirmatory data in terms of OS. In addition, a sponsor arguing in favour of PFS as primary endpoint based on feasibility of OS is only sometime persuasive. This argument is at risk of being spoilt by sponsors who claim lack of feasibility seemingly just as a route to conducting smaller, shorter trials with arguably more predictable outcomes. The debate continues. In this respect, the focus of the PFS Working Group to further develop ways to measure the clinical relevance of progression deserves full attention.

Publication disclaimer

The views presented here are personal and should not be understood or quoted as those of the European Medicines Agency.

Conflict of interest statement

None declared.

Acknowledgements

With thanks to David Brown, MHRA and to Iordanis Gravanis, EMA, for helpful insights.

REFERENCES

1. Guideline on the evaluation of anticancer medicinal products in man; 2006. Available from: http://www.ema.europa.eu/ema/pages/includes/document/open_document.jsp?webContentId=WC500017748.
2. Guidance for industry, clinical trial endpoints for the approval of cancer drugs and biologics; 2007. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071590.pdf.
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. Tuma R. Progression-free survival remains debatable endpoint in cancer trials. *J Natl Cancer Inst* 2009;**101**:1439–41.
5. Stone AM, Bushnell W, Denne J, et al. Research outcomes and recommendations for the assessment of progression in cancer clinical trials from a PhRMA working group. *Eur J Cancer* 2011.
6. 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* 2011.
7. 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.
8. Appendix 1 to the guideline on the evaluation of anticancer medicinal products in man (CHMP/EWP/205/95 Rev. 3) methodological considerations for using progression-free survival (PFS) as primary endpoint in confirmatory trials for registration; 2008. Available from: http://www.ema.europa.eu/ema/pages/includes/document/open_document.jsp?webContentId=WC500017749.
9. Zhang Z, Sun J. Interval censoring. *Stat Methods Med Res* 2010;**19**:53–70.