Independent radiologic review of the Gynecologic Oncology Group Study 0218, a phase III trial of bevacizumab in the primary treatment of advanced epithelial ovarian, primary peritoneal, or fallopian tube cancer: methodological comment

G. Daniele[1], M. Di Maio[1], M.C. Piccirillo[1]

Introduction

Important features of progression free survival (PFS) as an endpoint of randomized trials – such as independence from salvage therapy and subsequent therapies, shorter follow-up and lower number of events required on average – make it preferable to overall survival (OS) in some cases. However, despite the efforts made to harmonize the way disease progression is reported, the assessment of PFS is still based on personal evaluation of radiologic images. The potential for subjective assessment of progressive disease (PD) leading to differential judgment across the two arms of a trial, raises major concerns regarding biased evaluation of PFS, especially in unblinded trials. In view of this, regulatory authorities require registration studies with PFS as primary endpoint to include a blinded independent central review (BICR) of the radiologic examination of each patient. However, the objectives and the methodology of BICR, as well as any potential advantage, are still the subject of discussion among the different bodies.

Bevacizumab has been approved, in combination with carboplatin-paclitaxel chemotherapy, as first-line treatment for advanced ovarian cancer, based on a PFS advantage over carboplatin-paclitaxel chemotherapy alone, in the recently published Gynecologic Oncology Group (GOG) 218 study [1]. As required by regulatory bodies, GOG 218 incorporated BICR of radiology scans of all patients enrolled. The results of this analysis have been recently published [2].

Summary of the two publications

The GOG 218 trial was a phase III, randomized, three-arm, double-blind, placebo-controlled trial evaluating the potential benefit, in term of PFS, of adding short-term (concurrent with chemotherapy for 6 cycles) or long-term bevacizumab (concurrent with chemotherapy, plus maintenance for up to 22 cycles) to carboplatin-paclitaxel chemotherapy, as first-line treatment for patients with stage III-IV ovarian cancer [1]. Computed tomography (CT) or magnetic resonance (MR) was performed at baseline and after cycles 3, 6, 10, 14, 18, 22 and then every 3 months for 2 years, then every 6 months for 3 years, and annually thereafter. Imaging was interrupted after PD was detected for each patient, except when PD was declared only based on elevated serum CA-125. In that case, the investigator had to obtain a scan within 2 weeks of detection of the increased CA-125. Independent review started 2 years after the enrolment commenced, thus scan collection was both retrospective and prospective. For those patients progressing based solely on CA-125 criteria, data were censored to the last tumour assessment when they were known to be PD-free. To be eligible for BICR, patients should have been on study treatment for at least 9 weeks. Two independent radiologists performed the imaging review. A third radiologist served as adjudicator in case of discordance between the previous two. A clinician (medical oncologist) made the final adjudication of response and progression status based on the final radiologic evaluation and clinical data. Median PFS assessed by BICR were estimated with the Kaplan-Meier method, and the log-rank test was used to compare PFS rates between the arms and the Cox model was used to estimate the stratified hazard ratios (HRs) according to stage and GOG-Performance

[1]Clinical Trials Unit, Istituto Nazionale Tumori IRCCS "Fondazione G. Pascale", Napoli, Italy.

Correspondence to: Dr. Gennaro Daniele, Clinical Trials Unit, Istituto Nazionale Tumori IRCCS "Fondazione G. Pascale", Via M. Semmola, 80131 Napoli, Italy.
Phone: +39 081 5903571 – Fax: +39 081 7702938
E-mail: g.daniele@istitutotumori.na.it

CANCER BREAKING NEWS 2014;2(1):35-37

Status. Discordance measures were calculated according methodology described by Amit et al. [3]. Overall, scans of 91% of patients enrolled in GOG 218 were considered by the BICR; among these, 97.2% of patients had all the scans required by the protocol. Compared with the primary analysis PFS data, censored for CA-125-only PDs (median PFS of 12 months for control [chemotherapy plus placebo] and 18.2 months for the long-term bevacizumab arm), the BICR evaluated median PFS was 13.1 months in the control arm and 19.1 months in the long-term bevacizumab arm. HRs in favour of long-term treatment were 0.624 (95% CI 0.520–0.749; p<0.001) and 0.623 (95% CI, 0.503–0.772; p<0.001) in the local evaluation (LE) and BICR analyses, respectively. Short-term bevacizumab treatment was not significantly different from the control arm in both analyses. Moreover, an overall concordance between LE and BICR-assessed PD status (76.8%) and PD date (73.3%) was observed. However, the degree of concordance in PD date assessment was not uniform across the arms, with a higher frequency of PD declared earlier by BICR in the control arm (24.9%) and short-term arm (20.5%) compared with the long-term arm (10.6%), resulting in a higher degree of LE-BICR concordance, in the latter *versus* the control arm (84.6% *vs* 68% respectively). The authors rightly conclude that the effect size demonstrated both at LE and with BICR is a reliable estimate of the benefit resulting from adding long-term bevacizumab to chemotherapy in the selected population of patients with chemotherapy-naïve advanced ovarian cancer.

Methodological comment

Although OS is the most reliable and ethically acceptable endpoint in the clinical development of cancer drugs, the large number of patients to be studied and the long follow-up required have contributed to increased use of PFS as the primary endpoint in many trials required for the registration of anti-cancer drugs. Indeed, since disease progression usually precedes the death of a patient, the follow-up time required to register the number of events needed for the analysis is shorter and the number of patients is smaller than with OS. Moreover, prolonging the time a patient is progression free could represent an important outcome *per se* and it is not influenced by subsequent treatments. However, some methodological hurdles hinder the complete reliability of PFS as a relevant proof of treatment effect. Particularly, there are three sources of bias that could occur when measuring and comparing PFSs between two arms (one experimental and one control) in a clinical trial. The first bias happens when the intervals between subsequent tumour assessments are systematically longer in one arm *versus* the other. The second bias is when the rate of patient dropout is significantly different between the two arms. The third, the evaluation bias, is based on the investigator feeling that the experimental treatment is better than the control, and therefore is more likely to declare PD in the control than in the experimental arm. Although the first two biases can be prevented at protocol development and data monitoring stages, respectively, the third is difficult to control, especially in unblinded trials. The case of the investigator wishing to shift the patient to the experimental arm in an unblinded crossover trial – when the investigator knows his/her patient is being treated with control – could be a typical case in which evaluation bias might occur. Regulatory agencies and methodologists advocate the use of BICR, central review of all the radiologic examinations in a trial by an independent central radiologist committee, as the major tool to prevent evaluation bias. The modalities and the significance of the BICR are still under discussion by regulatory bodies and drug developers. In fact, implementing central collection of all radiologic scans in a trial is logistically cumbersome and very expensive. Although some new web-based systems can aid in implementing this process in a more efficient way, these largely depend on the IT infrastructure of peripheral sites. Therefore, central collection of scans for each patient is still almost invariantly retrospective, and this could represent a major issue in considering BICR as a reliable measurement of PFS. Indeed, after a patient is considered to have PD at LE, the treatment is changed and his/her scans after detection of PD are no longer available to the BICR. In cases where the BICR was unable to confirm PD assessed locally, data for that patient are censored at that point and no more information is available for him/her. In this case, the censoring is informative (producing a lack or distortion of information) and might lead to a bias. This is the case in which the blinding – considered to be the solution for the bias – is a source of bias for its own sake. Obviously, censoring would not be a problem if the number of censored patients is very similar between the two arms. However, it could be argued that, in the case of different efficacies between treatments, censoring would happen significantly more often in the arm with lower efficacy. Indeed, as the number of events registered increases, so does the chance of discordance.

Finally, another point should be considered when evaluating the benefits of BICR. Implementation of full BICR is constantly associated with discrepancies. These discrepancies between BICR and LE assessments can be due to measurement variations among the local and central physicians that do not represent a bias, since they are equally, at least in principle, distributed between the arms. As an indirect demonstration of this latter concept, only in few

cases [4] discrepancies at patient level have led to different conclusions about treatment effect between LE and BICR, whilst in the vast majority of cases, discrepancies did not hamper the overall results of the trials [3, 5, 6]. For all these reasons, complete BICR has been discouraged [6] and an audit strategy, based on a randomly selected sample of patients, has been proposed. Overall, two audit strategies have been proposed: the first, promoted by a group of researcher from the National Cancer Institute (NCI) [5, 7] is a two-step strategy aimed at demonstrating that local PFS estimation of the treatment effect (the HR) is not significantly biased. In the first step, a BICR-based HR is calculated for the audited sample of patients. In the case that the two HRs are significantly different, a significant risk of bias exists and there is need for a full BICR (the second step). This strategy applies only in cases where the HR is clinically meaningful and shows a statistically significant advantage in favour of the experimental arm. The second approach [3] is the one Burger et al. applied in the study commented on here [2]. This method uses differential discordance as a measure of the evaluation bias. In particular, two measures are calculated: the early discrepancy rate (EDR; the rate of PD declared earlier locally *versus* BICR) and the late discrepancy rate (LDR; the rate of PD declared later by LE *versus* BICR). The differential discordance for each of these measures is the difference between the two arms (i.e. experimental-control). A negative differential discordance in EDR (higher number of "early" PDs in the control arm) and/or a positive one for LDR (higher number of "late" PDs in the experimental arm) suggests a potential risk of evaluation bias favouring experimental treatment and needs to be further evaluated by a full BICR. In the analysis presented by Burger et al. [2], both small, positive differential EDR (0.07) and negative LDR (–0.09) provide reassurance on the reliability of the LE and BICR assessment of PFS as an estimate of treatment effect.

A recently published paper assessed which of the two approaches is the best performing in 26 randomized superiority trials [8]. Overall, it emerged that both are reliable strategies for performing BICR audits. However, the first method, developed by an NCI researcher, seems to perform better in most situations in distinguishing between trials with or without a potential evaluation bias. This method can be hampered by the sample size required; this is directly dependent on the effect size observed at LE. In contrast, the second method, although intuitive, is limited by the accurate choice of the thresholds for EDR and LDR to be considered positive for the risk of evaluation bias. The choice of this threshold might be the subject of discussion by regulatory agencies, thus requiring complete BICR more frequently [9]. Moreover, this method counts discrepancies, but does not account for how far apart they are. In cases where the intervals were long and asymmetric between the arms, such discrepancies would potentially have led to an evaluation bias. In the case of Burger et al., however, it has been said that in the control arm, PD was declared earlier by BICR than LE in a higher percentage of cases, it has been also said that the time from the two assessments declaring PD was in most cases ≤12 weeks (i.e. one re-evaluation interval). This is reassuring as it suggests against the presence of a significant bias.

References

1. Burger RA, Brady MF, Bookman MA, et al. Incorporation of bevacizumab in the primary treatment of ovarian cancer. N Engl J Med. 2011;365(26):2473-83.
2. Burger RA, Brady MF, Rhee J, et al. Independent radiologic review of the Gynecologic Oncology Group Study 0218, a phase III trial of bevacizumab in the primary treatment of advanced epithelial ovarian, primary peritoneal, or fallopian tube cancer. Gynecol Oncol. 2013;131(1):21-6.
3. 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;47(12):1772-8.
4. 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(1):19-26.
5. 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(22):3791-6.
6. Ford R, Schwartz L, Dancey J, et al. Lessons learned from independent central review. Eur J Cancer. 2009;45(2):268-74.
7. Dodd LE, Korn EL, Freidlin B, et al. An audit strategy for progression-free survival. Biometrics. 2011;67(3):1092-9.
8. Zhang JJ, Zhang L, Chen H, et al. Assessment of audit methodologies for bias evaluation of tumor progression in oncology clinical trials. Clin Cancer Res. 2013;19(10):2637-45.
9. Pignatti F, Hemmings R, Jonsson B. Is it time to abandon complete blinded independent central radiological evaluation of progression in registration trials? Eur J Cancer. 2011;47(12):1759-62.