



U.S. Food and Drug Administration

Notice: Archived Document

The content in this document is provided on the FDA's website for reference purposes only. It was current when produced, but is no longer maintained and may be outdated.

Ovarian Cancer Endpoints Workshop

April 26, 2006

MEETING SUMMARY

Bethesda North Marriott Hotel and Conference Center
Bethesda, MD

U.S. Food and Drug Administration
American Society of Clinical Oncology
Co-sponsored by the American Association for Cancer Research

Welcome and Introduction

Richard Pazdur, MD (Co-Chair), Director, OODP, Center for Drug Evaluation & Research (CDER), welcomed participants and said the meeting was one of a series of FDA-sponsored workshops evaluating endpoints for drug approval studies in the most common cancers. The purpose of this workshop was to have a wide-ranging discussion about the pros and cons and major areas of controversy with regard to endpoints for trials to support the approval of new drugs to treat ovarian cancer. Issues highlighted at the workshop would subsequently be discussed by the Oncologic Drugs Advisory Committee (ODAC), the FDA's statutory advisory body on issues related to oncology drugs. The workshop was not an advice-giving meeting; by law, FDA may take advice only from its statutory advisory committees.

Regulatory Background

Lee Pai-Scherf, MD, Medical Officer, Division of Biologic Oncology Products, summarized FDA requirements for new drug approval. Regular marketing approval of oncology drugs requires substantial evidence of efficacy from adequate and well-controlled clinical trials. Drugs also must be safe for their intended use. The safety requirement was promulgated in the Federal Food Drug and Cosmetic Act of 1938 and the efficacy requirement was codified in a 1962 amendment to that act. The methodology used to assess patient response to the study drug must be well defined and reliable.

Two approval pathways exist for oncology drugs. Under *regular approval*, efficacy must be demonstrated by clinical benefit (e.g., prolongation of life or better quality of life) or by an established surrogate for clinical benefit. *Accelerated approval* (AA) requires demonstration of efficacy based on a surrogate that is "reasonably likely" to predict clinical benefit. AA was added to the new drug application regulations in 1992 to allow drug approvals for serious or life-threatening diseases. The new drug must provide benefits over available therapy. Subsequent confirmation of clinical benefit is required and is part of the post-marketing process.

A clinical endpoint is a measurement or sign that directly measures how a patient feels, functions, or survives. A surrogate endpoint is a measurement or sign that is used as a substitute for a clinical endpoint. It is assumed that the surrogate is a reliable predictor of the primary endpoint of interest. Established surrogates (used for regular approval) include durable complete remission (CR) in acute leukemias and progression-free survival (PFS) in adjuvant therapy for breast cancer. An example of a surrogate endpoint that is reasonably likely to predict clinical benefit (i.e., used for AA) is durable tumor response in certain solid tumors.

From the early 1970s to the mid-1980s, FDA approved oncology drugs on the basis of tumor response rate (RR) alone. In the mid-1980s, on the advice of ODAC, FDA determined that RR

generally should not be the sole basis for approval. The potential benefit associated with a response did not necessarily outweigh the substantial toxicity of oncology drugs, and the correlation between RR and survival or clinical benefit was not well established. The new FDA position required an improvement in survival or in patients' symptoms for drug approval. In the early 1990s, other endpoints that potentially demonstrated clinical benefit were examined and accepted by FDA. Examples are disease-free survival (DFS) for hormonal adjuvant therapy for breast cancer and durable CR in leukemias.

Advantages and Disadvantages of Different Drug Approval Endpoints

Dr. Pai-Scherf outlined the pros and cons of several endpoints: overall survival (OS), PFS, RR, and measures of patient feeling or function.

Survival

Survival is the gold standard in oncology treatment. The benefits of survival as a study endpoint are that it is 100% accurate for the event and date and it is not subject to investigator bias. Drawbacks are that a survival endpoint requires a larger sample size and longer follow-up than other endpoints. In addition, crossover and secondary therapy may obscure the results.

Progression-Free Survival

The use of PFS as a study endpoint enables shorter follow-up and thus permits studies to be completed more quickly. Additionally, results are not obscured by secondary therapy. However, PFS carries the potential for bias because the outcome is sensitive to the timing (usually every 2 to 4 months) of the assessment.

Response Rate

Using an RR endpoint, the treatment is entirely responsible for any tumor reduction. Investigators must also consider duration of response in their assessment. RR can be reliably assessed in single-arm trials.

Patient-Reported Outcomes

Measures of patient feeling or function can be reported by patients or observed by third parties. The benefit of such measures is that they provide the patient's perspective on treatment. The use of patient-reported outcomes (PROs) as endpoints for drug approval has several drawbacks, however. Blinding of studies is required but is difficult to execute. Small score changes and missing data often lead to inconclusive findings. Adequate development and validation of PRO measurement instruments is critical. In addition, the statistical analysis must plan for multiple comparisons.

Past FDA Approvals for Ovarian Cancer Drugs

Dr. Pai-Scherf reviewed the drugs previously approved to treat ovarian cancer and the endpoints used in clinical trials of those drugs.

Cisplatin for first- and second-line therapy was approved in 1978 on the basis of two randomized clinical studies ($N = 52$ in both studies) comparing (1) cisplatin vs. cisplatin/adriamycin vs. thiotepa alone or plus methotrexate and (2) cisplatin alone vs. cisplatin/hydration/mannitol. Approval was based on RR. In the first study, RR was 42% vs. 67% vs. 36%; in the second study, RR was 42% for cisplatin alone vs. 63% for the combination therapy.

In 1991, carboplatin was approved for first-line treatment of advanced ovarian cancer (in established combination with other approved chemotherapeutic agents). Approval was based on

the results of two randomized controlled trials (RCTs) of carboplatin vs. cisplatin. Both drugs, in combination with cyclophosphamide, demonstrated equivalent OS between the two groups. There was limited statistical power to demonstrate equivalence in overall pathologic complete RR because of the small number of patients.

In 1998, paclitaxel/cisplatin received regular approval on the basis of two RCTs comparing paclitaxel/cisplatin with cisplatin/cyclophosphamide. Approval was based on a statistically significant improvement in OS in the paclitaxel/cisplatin arms. The studies also showed a significantly higher RR and longer time to progression (TTP) in the paclitaxel/cisplatin arms.

Altretamine received regular approval for second-line therapy in 1990. Approval was based on two single-arm studies showing RRs of 20% and 14%. The duration of response ranged from 2 to 36 months.

Paclitaxel received regular approval for second-line therapy in 1992 based on a Phase III study ($N=407$) of bifactorial design that compared 2 different doses (135 or 175 mg/m²) and two schedules (3-hour or 24-hour infusion). The RR was 16.2% (95% CI; 12.8–20.2%). The median duration of response was 8.3 months (range 3.2 to 21.6 months). The study had insufficient power to determine whether a particular dose and schedule produced superior efficacy; however, the 24-hour infusion was found to be more toxic than the 3-hour infusion.

Topotecan received regular approval for second-line therapy in 1996 on the basis of two studies. One, a randomized study of topotecan vs. paclitaxel, yielded an RR of 21% for topotecan and 14% for paclitaxel. The duration of response was 25.0 weeks for the topotecan arm and 21.6 weeks for the paclitaxel arm. TTP was 18.9 weeks for the topotecan arm and 14.7 weeks for the paclitaxel arm. OS was 63.0 weeks for the topotecan arm and 53.0 weeks for the paclitaxel arm. The other, a single-arm study ($N=111$), found an RR of 14% and a median response duration of 22 weeks.

Liposomal doxorubicin received AA for second-line therapy in 1999 on the basis of three single-arm studies yielding an RR of 13.8 %, with an average response duration of 39.4 weeks. The drug received regular approval in 2005 on the basis of a randomized comparison with topotecan. OS was 14.4 months for liposomal doxorubicin vs. 13.7 months for topotecan. TTP was 4.1 months for liposomal doxorubicin vs. 4.2 months for topotecan. RR was 19.7% for liposomal doxorubicin vs. 17.0% for topotecan; median response duration was 6.9 months vs. 5.9 months.

Summary

In summary, both RR and OS have been the basis of drug approvals for first-line therapies for ovarian cancer. For second- and third-line therapy, RR, TTP, and OS have been the basis for approvals.

Design Issues in Clinical Trials of Ovarian Cancer

J. Tate Thigpen, MD (co-chair), University of Mississippi School of Medicine, provided an overview of the management of ovarian carcinoma and offered recommendations for clinical trial design. He emphasized that celomic epithelial cancers account for 90% of all ovarian cancers and were the focus of this meeting.

Management of ovarian carcinoma depends on the extent of disease and prior therapy that the patient has received. The FIGO* staging system is used to classify the extent of disease and provide the basis for treatment considerations. Patients with newly diagnosed Stage I or II disease have limited ovarian carcinoma confined to the ovaries and pelvis. Patients diagnosed with Stage III or IV disease have advanced ovarian carcinoma that is intraperitoneal (IP) or involves distant metastases. Patients whose disease recurs or persists after prior therapy are referred to as having recurrent or persistent disease. Most patients (75%) present with disseminated (Stage III or IV) disease; 17% with Stage III disease, 58% with Stage IV.

The standard of care for advanced disease, set forth at a 2004 Gynecologic Cancer Intergroup (GCIIG) consensus conference, consists of maximum attempted surgical cytoreduction, chemotherapy following surgery, and a drug regimen of paclitaxel/carboplatin repeated every 3 weeks for 6 cycles. Results with this approach depend on the volume of residual disease. Those with large-volume residual disease (nodules >2 cm in diameter remaining after surgery) will achieve a clinical CR in 40 to 50% of cases and will have a median PFS of 17 months and median survival of 30 months. Those with small-volume residual disease (no nodule >2 cm in diameter after surgery) have a 95% probability of ending initial therapy with no clinical evidence of disease and will exhibit median PFS of 25 months and median survival of 60 months.

For Stage I or II, the general practice is to look at tumor features that predict prognosis and divide patients according to low and high risk for recurrence. Low-risk patients are those with low-grade disease whose cancer is intracystic, who have no extraovarian disease, have negative peritoneal cytology, and have no ascites. High-risk patients are those with intermediate- to high-grade disease whose cancer is extracystic and who have extraovarian disease, positive peritoneal cytology, and ascites. Treatment recommendations for limited disease include platinum-based therapy, total abdominal hysterectomy–bilateral salpingo-oophorectomy, and careful surgical exploration. Some patients with low-risk limited disease have no further therapy. Those with high-risk disease receive platinum-based therapy; adjunct platinum-based therapy cuts relapse in half. Patients with advanced disease receive surgical cytoreduction and paclitaxel/carboplatin.

Most patients (62%) will not achieve long-term control of the disease and will develop either recurrent or persistent disease. Management of such patients requires that they first be classified as having either *chemosensitive disease* (i.e., response to first-line therapy leading to a treatment-free interval of at least 6 months) or *chemoresistant disease* (i.e., progression during first-line therapy or best response to first-line therapy; stable disease; or recurrence within 6 months of completing first-line therapy). Those with chemosensitive disease are retreated with a platinum-based regimen; the expected RR is >60% and median survival is ≥ 30 months. Those with chemoresistant disease are treated with alternative drug therapy; expected RR is 12 to 32% and median survival is ≥ 8 months.

Dr. Thigpen listed the ovarian cancer drugs that have been approved by FDA along with drugs that are actively used but not approved. Research directions of immediate interest include dose intensity (IP therapy), the addition of further cytotoxic agents, the role of biologic agents, and the value of maintenance or consolidation therapy.

* Federation Internationale de Gynecologie et d'Obstetrique

Unique Features of Ovarian Cancer

Ninety percent of ovarian cancers arise from the celomic epithelium in the ovary and elsewhere in the peritoneum. The primary route of spread is IP seeding. Accurate staging and assessment require evaluation of the peritoneal cavity. No effective early diagnostic test exists, with the result that most patients present with advanced disease. Treatment requires multiple active systemic agents. Patients have a high RR to standard front-line chemotherapy. Post-recurrence and post-progression treatment have a significant impact on OS. CA-125 is widely used as a marker of both progression and response. Many clinical decisions are based on CA-125 values, a fact that must be taken into account in designing clinical trials.

Goals of Therapy

The main goal of therapy is cure. Clinical goals of therapy include prolonged survival, delayed progression of disease, reduced tumor burden, alleviation of symptoms, and minimization of the toxicity of therapy. Trial endpoints must reflect these goals.

Trial Endpoints That Reflect Goals of Therapy

Current endpoints include survival, PFS, objective response (measured by RECIST* and CA-125), pathologic complete response, and quality of life (QoL) (measured by instruments such as FACT-O, QLQ-C30, and QLQ-OV28).

Trial Settings to Which Endpoints Will Be Applied

Trial endpoints can be applied in five clinical settings: first-line therapy for new disease (advanced and limited), maintenance therapy, and recurrent/persistent disease (chemosensitive and chemoresistant).

First-Line, Advanced Disease

Ovarian cancer most commonly presents as advanced disease. Survival improvement generally is required for drug approval. The GCIG consensus conference concluded the following:

- “There is an impact of post-recurrence/progression therapy on overall survival.”
- “It is not possible to standardize post-recurrence/progression therapy at present.”
- “Although overall survival is an important end point, progression-free survival may be the preferred primary end point for trials assessing the impact of first-line therapy because of the confounding effect of the post-recurrence/progression therapy on overall survival.”
- “There should be clear definition of how to determine progression-free survival.”

PFS should be considered the primary endpoint for trials of advanced disease for the following reasons: (1) it avoids the confounding effect of additional therapy, (2) it provides a measure of clinical benefit (i.e., increased time off therapy without progression), and (3) it predicts survival improvement—that is, it is a surrogate for OS.

Results from seven clinical trials (GOG-97, 111, 152, 52, 158, 114, and 172) support the use of PFS as a primary endpoint. In addition, six trials conducted outside the United States (ICON 2, ICON 3, AGO/GINECO, AGO OVAR 3, OV 10, and EORTC Surg) show concordance between PFS and survival. Two trials (GOG-47 and GOG-132), however, showed discordant results. Those results can be explained by the impact of post-recurrence/progression therapy. In trials involving advanced disease, follow-up to determine progression must be uniform and of sufficient frequency and must account for the effect of CA-125 measurement on therapy.

* Response Evaluation Criteria in Solid Tumors

Limited Disease

No drugs have been approved specifically for the limited-disease population. The GCIG consensus conference concluded the following:

- “There is an impact of post-recurrence/progression therapy on overall survival.”
- “It is not possible to standardize post-recurrence/progression therapy at present.”
- The treatment goal for early ovarian cancer should be “recurrence-free survival” (RFS).

The ICON 1/ACTION trials, a pooled analysis of two trials involving 925 patients, the majority of whom were at high risk for recurrence, found an 11% improvement in RFS at 5 years and an 8% improvement in OS with platinum-based adjuvant chemotherapy.

Reasons to consider RFS as the primary endpoint for trials of limited disease are that (1) RFS avoids the confounding effect of additional therapy, (2) it provides a measure of clinical benefit (increased time off therapy without progression), and (3) RFS improvement predicts survival improvement. The same research caveats apply as for advanced disease: Follow-up to determine progression must be uniform and of sufficient frequency, and it must account for the effect of CA-125 measurement on therapy.

Maintenance/Consolidation Therapy

No drugs have been approved specifically for maintenance/consolidation therapy. The GCIG consensus conference recommended (in the only statement not approved unanimously) OS as the endpoint for maintenance following first-line therapy. A minority voted that in certain situations, PFS may be considered a primary endpoint in maintenance trials following first-line therapy. The conference also noted: “Since trials involving maintenance by definition have longer treatment on the experimental arm as compared with the control, the real question is whether the prolonged therapy improves survival.”

Only one positive study of a drug for maintenance/consolidation therapy (GOG-178) has taken place. PFS was the endpoint in that study; patients with advanced disease in clinical CR after front-line therapy who received 12 cycles of paclitaxel 175 mg/m²/3h experienced significantly longer PFS than patients who received 3 cycles of the regimen (28 months vs. 21 months). The difference was maintained well beyond cessation of maintenance. Survival data are not yet available; a confirmatory trial (GOG-212) uses survival as the primary endpoint and a no-maintenance control arm.

A recently activated trial (involving some of the groups voting that OS is the only valid endpoint) uses PFS as the primary endpoint. Insufficient data support any alternative to OS as the appropriate endpoint at the present time.

Recurrent/Persistent Disease

For recurrent or persistent disease, approvals have been based on RR, OS, and (more recently) PFS. Some studies that have been the basis for approval missed their primary endpoint but yielded significant differences in regard to other endpoints. The GCIG consensus conference concluded: “The choice of the primary end point needs to be fully justified with appropriate power calculations. Symptom control/QoL (for early relapse) and OS (for late relapse) may be the preferred primary end point although PFS should still be used in the assessment of new treatments.”

Reasons to consider PFS as the primary endpoint for trials of recurrent/persistent disease are that PFS avoids the confounding effects of additional therapy, provides a measure of clinical benefit (more time without increasing tumor burden), and appears to predict survival improvement. PFS predicted for survival improvement in two of three large Phase III trials for recurrent/persistent disease (ICON 4, AGO OVAR 2.5, and PLD vs. topotecan). In the third trial, 75% of patients received further therapy. Caveats are that follow-up to determine progression must be uniform and of sufficient frequency, and it must account for CA-125.

Recommendations

Dr. Thigpen made the following recommendations:

- For first-line therapy in advanced disease, the primary endpoints should be survival and PFS (which predicts survival, reflects clinical benefit, and avoids the confounding effects of further therapy). Supporting endpoints are response, complete response, and QoL.
- For limited disease, endpoints should be survival and DFS (which predicts survival, reflects clinical benefit, and avoids the confounding effects of further therapy).
- For maintenance/consolidation therapy, the primary endpoint should be survival. The case for an alternative endpoint is not clear, but PFS would avoid the confounding effect of further therapy and would reflect clinical benefit in the form of greater time without progressing tumor burden.
- For recurrent/persistent disease, endpoints should be survival and PFS (which predicts survival, reflects clinical benefit, and avoids the confounding effects of further therapy). Supporting endpoints are response, complete response, and QoL.

Issues for further discussion include the role for CA-125 in determining progression and response; clinical trial endpoints (particularly PFS and its assessment) for regulatory approval of first-line therapy for advanced ovarian cancer, along with maintenance and subsequent therapy; the role of patient-reported outcomes; and biomarker and endpoint research priorities.

Regulatory Use of CA-125 in Ovarian Cancer

Use of CA-125 for Response Evaluation in Ovarian Cancer

Robert C. Bast, MD, Translational Research, MD Anderson Cancer Center, Houston, TX, began by describing the characteristics of CA-125, which is a high-molecular-weight, heavily glycosylated mucin. It has recently been cloned and has been designated as MUC16. Like other tumor markers, it is released from dying cancer cells and is probably shed by proteolytic cleavage. The amount of CA-125 shed relates to surface expression and to protease activity.

Unpublished data by Jeff Boyd at Sloan-Kettering, combined with data from Dr. Bast's own laboratory, indicate that CA-125 shedding is modestly upregulated by stimulating protein kinase A, inhibiting protein kinase C, or inhibiting tyrosine phosphatase. Dr. Bast's data show that changes affect only about 20% of CA-125 molecules that are shed or expressed on the cancer cell surface. This is important because to use CA-125 as an endpoint for different inhibitors, one must ensure that the inhibitor does not consistently upregulate CA-125, leading to false negative results. Shedding is modestly downregulated (less than 20%) by epidermal growth factor and lysophosphatidic acid and is not affected by a variety of biological agents, including estrogen, progesterone, and androgen.

In healthy adult women, strong CA-125 expression is limited almost exclusively to the endometrium. Serum levels can be elevated by obstetric and gynecologic conditions including pregnancy, menstruation, endometriosis, adenomyosis, and uterine fibroids; those conditions are

not confounding factors in ovarian cancer because most patients have had hysterectomy and oophorectomy. Elevated CA-125 can result from inflammation of serosal surfaces and effusions in the pleural, pericardial, or peritoneal cavity. No expression of CA-125 occurs in about 20% of epithelial ovarian cancers.

Even though CA-125 at the tissue level is expressed by 80% of epithelial ovarian cancers, as a marker, it is elevated in sera from >90% of ovarian cancer patients because it is produced both by cancer cells and by activated mesothelial cells in the peritoneum reacting to serosal implants. It can be measured by double-determinant immunoassays with murine monoclonal antibodies.

In 1979, Dr. Bast's group, in collaboration with Robert Knapp, developed the OC125 antibody and first defined the antigen. Dr. Bast described the group's initial assay and said that with Tim O'Brien's development of the M11 antibody, it was possible to develop an assay using M11 to capture the antigen and OC125 to detect it; this is now the standard CA-125 II assay. An advantage of the CA-125 II is that its interassay coefficient of variation is <5%, permitting accurate monitoring. The half-life of CA-125 if all sources are removed is 7 to 14 days with a biphasic curve. An apparent half-life of >21 days indicates residual disease and poor prognosis. Values can be elevated after laparotomy but return to baseline within 1 month. Anti-murine immunoglobulin antibodies in patient sera can provide false-positive elevations with both CA-125 assays.

The strength of CA-125 for monitoring patients and for clinical trials is that it detects disease when nodules are too small to be detected by even the most sensitive CT scan or MRI. Almost half of patients who would be willing to participate in clinical trials do not meet the RECIST criteria; if a biomarker could be used, more women would be eligible for trials.

Applications of CA-125

From the earliest reports, it was shown that with successful treatment, CA-125 declined, then began to rise. CA-125 levels track tumor burden with approximately 90% accuracy during treatment of recurrent disease. Elevated CA-125 detects recurrence in approximately 80% of patients with a lead time of 2 to 3 months; however, not every patient has elevated CA-125 before detection of disease by more conventional methods. Biochemical recurrence detected by measurement of CA-125 alone can precede evidence of recurrence by diagnostic imaging or physical examination.

Dr. Bast and Dr. Rustin suggested three applications for CA-125: (1) as an endpoint in Phase II trials; (2) as an endpoint for TTP in Phase III trials; and (3) in the absence of measurable disease, to assess drug activity (both cytostatic and cytotoxic drugs).

Gordon J.S. Rustin, MD, Mount Vernon Cancer Center, Middlesex, England, presented data on CA-125 as an endpoint in Phase II trials and as an endpoint for TTP in Phase III trials. He also described how to use rising CA-125 levels in the absence of measurable disease to assess drug activity. One of the difficulties in data collection, he said, is that endpoints are measured at different points in time. In an individual patient, a clinical trial might collect CT scan data at pretreatment, 3 months, and 6 months, but collect data on CA-125 levels at every visit. This practice results in poor correlation because different events happen at different points in time. A sensible approach would be to look at trials to find out whether, had CA-125 had been used as the measure, the result would have been any different with regard to the RECIST criteria.

Data were obtained from 25 different treatment groups in 19 clinical trials using 14 drugs. A total of 1,092 patients had both measures of CA-125 and RECIST data. A hypothetical Gehan two-stage Phase II trial was created; target drug efficacy was 20%, and the rejection error was 5%. Precise definitions based on a 50% or 75% fall in CA-125 accurately predicted drug activity. CA-125 and standard response criteria were completely concordant in 20 of 25 groups. They were discordant in five groups: in four, paclitaxel was rejected by standard criteria but not by CA-125; and in one, etoposide was rejected by CA-125 but not by standard criteria (Rustin et al, 2000). Many people said that the definition was overly complicated, however. A simplification of the 50% or 75% CA-125 response criteria to just 50% remains very accurate. In an analysis of 236 patients by GINECO, only three patients responded according to CA-125 criteria yet had progressive disease according to the RECIST criteria.

In the GCIG definition of CA-125 response to therapy for relapsed ovarian cancer, a CA-125 response has occurred if, after two elevated levels prior to therapy, there is at least a 50% decline that is confirmed by a fourth sample. This approach requires two pretreatment samples, both of which are $\geq 2x$ the upper limit of normal, one within 1 week of starting therapy and the other within 3 months. The third sample must be $\leq 50\%$ of the second sample. A confirmatory fourth sample is taken ≥ 21 days after Sample 3 and must be $\leq 50\%$ of Sample 2. The samples are not evaluable if interference with pleura/peritoneum has occurred in the prior 28 days or the patient has received mouse antibodies. CA-125 cannot be used if the peritoneum is drained to relieve ascites.

Recommendations

Dr. Rustin made the following recommendations for the use of CA-125 in Phase II trials:

- Use CA-125 to support “go/no-go” decisions.
- Define the RR below which further development should be halted.
- Define a minimal acceptable RR.
- Define the number of patients required to achieve 90% power to define response by CA-125.
- If the CA-125 RR is greater than the minimal acceptable rate, the trial should continue so that response can be measured with the same power using the RECIST criteria.

Use of CA-125 to Define Progression

Defining progression of ovarian cancer during follow-up according to doubling of CA-125 from the upper limit of normal has a low false-positive rate, Dr. Rustin said. In a trial of 5 vs. 8 courses of carboplatin, analysis after 87 relapses in a group of 131 evaluable patients whose CA-125 had fallen to <30 U/mL yielded a sensitivity of 84% and a false-positive rate of 1.4%, which is very low. The median lead time to clinical progression was 63 days.

In many patients, CA-125 never falls to normal. Confirmed doubling of CA-125 from nadir accurately defines progression (94% sensitivity), as demonstrated in a study of 302 patients receiving first-line chemotherapy. In 88 patients, CA-125 levels were always >23 U/mL, and at least four serial CA-125 levels were measured prior to clinical progression. In that study, 80 patients had true positives, 2 had true negatives, and 5 had false negatives. Just 1 false positive occurred.

GCIG’s definition of progression for patients with CA-125 in the normal range requires CA-125 $\geq 2x$ the upper limit of normal (ULN) documented on two occasions; progressive disease (PD) is defined as the first date of the CA-125 elevation to $\geq 2x$ ULN. For patients who never normalize,

progression is defined as CA-125 >2x nadir value on two occasions; PD is defined as the first date of the CA-125 elevation to >2x nadir value.

Results from a comparison of CA-125 and standard definitions of progression in the Intergroup trial of cisplatin and paclitaxel vs. cisplatin/cyclophosphamide found that CA-125 accurately predicted progression.

Recommendations

Dr. Rustin recommended that CA-125 be incorporated into trial protocols of first-line and relapse therapy of ovarian cancer. However, progression according to RECIST criteria should always take precedence. PFS will appear shorter when measured by CA-125 than by RECIST. It also reduces the number of CT scans required during follow up. The following conditions should apply:

- CA-125 measurements occur at the same time point on all arms of randomized trials.
- If mouse antibodies are used, they do not interfere with the assay.
- If biological/targeted therapy is used, data from Phase II trials show an acceptable number of discordant results.
- If IP therapy is given, CA-125 levels have returned to within normal range and measurement occurs >28 days from removal of the IP catheter.

Using Rising CA-125 Levels in the Absence of Measurable Disease to Assess Drug Activity

Patients with an asymptomatic rise in CA-125 levels are ideal study participants. To assess drug activity, the protocol would use CA-125 to determine response by GCIG criteria and would also determine CA-125 doubling time before and after the test therapy. The next step is to determine the proportion of patients in whom the rate of rise slows after test therapy. This approach is actually quite difficult; the endpoint is easier and quicker if patients are registered after responding to relapse therapy rather than after first-line therapy. Because patients with relapsed disease all know they will relapse again, it is easy to register them.

An ongoing study on the use of changes in CA-125 doubling time to detect tamoxifen activity has enrolled 40 patients. Huge variability in doubling time has been found between patients, but within individuals the rate of doubling is stable. The trial appears to be feasible, and early data are encouraging.

Points for Discussion

Dr. Rustin raised the following points for discussion:

- What else is required to validate CA-125 as an endpoint in Phase II trials? The criteria should be sensible but not too prescriptive.
- What else is required to validate CA-125 as an endpoint for TTP in Phase III trials?
- How can rising CA-125 levels be used in the absence of measurable disease to assess drug activity?
- What proportion of discordant results invalidates CA-125 for specific agents?
- If patients start second-line therapy just because of rising CA-125, what is the date of progression? It will depend upon the type and frequency of monitoring. Using CA-125 to measure progression will shorten PFS. If relapse therapy is delayed until progression has occurred according to the GCIG CA-125 definition, there is a progression date; however, if relapse therapy is started earlier, the progression endpoint will be lost.

The following research is needed:

- A register of trials to prospectively record CA-125 levels
- Analysis of actual CA-125 levels using GCIG criteria
- Validation of CA-125 criteria on the basis of whether its use alters trial results
- Trial designs that study patients with no measurable disease but an asymptomatic rise in CA-125
- Development of statistical methods for change-point analysis

FDA Perspective: Analytical Aspects of CA-125 Tests

Robert L. Becker, Jr., MD, PhD, Director, Division of Immunology & Hematology Devices (DIHD), Center for Devices and Radiological Health (CDRH), provided background on the regulation of devices approved for commercial CA-125 testing.

In 1987, Centocor submitted an application for premarket approval (PMA) for a radioimmunoassay. The device was Class III, meaning that it required a premarket approval study to establish its safety and effectiveness. The intended use was “as an aid in the detection of residual ovarian carcinoma in patients who have undergone first-line therapy and would be considered for a second look.” The data were reviewed by an FDA device advisory panel and the device was approved after about 2 years.

In 1997, the Centocor device and similar devices were downclassified to Class II; new assays would receive approval under the 510(k) process. In 1996, FDA issued a guidance document for tumor-associated antigens, which is still in use. The goal of 510(k) approval is to establish that a new device is substantially equivalent to the “predicate device”; the process is not as in-depth as a PMA and most studies are analytical, not clinical. Intended uses of the 16 devices (produced by eight manufacturers) approved through the 510(k) process are broad and include “aid in management; aid in monitoring response; serial testing; detection of cancer recurrence; and use in conjunction with other clinical methods.”

The technologies associated with these devices are roughly comparable and their report results are in similarly named units. Nevertheless, the linear correlations between pairs of assays yield regression slopes ranging from 0.77 to 1.34. Hence, even with high inter-test correlations (0.95 to 0.99) and good intra-test precision (CVs from 2.7% to 6.9%), the analytical comparability of results from different test is open to question. Dynamic range may be another complicating issue, since the amounts of detectable CA-125 can vary by several orders of magnitude from device to device.

Similarly, correlations with clinical progression across assays are variable. One reason may be that by looking at the same receiver operating characteristic for all the assays but using different cutoffs associated with those assays, one might trade off sensitivity vs. specificity regarding progression. If so, that is not by design, because the assays generally adhere to the same 35 unit/mL standard as the original Centocor assay. The difference in responsivity may impose the effect of a different cutpoint, resulting in skews one way or the other for sensitivity and specificity of an assay compared with its predicate device in signaling tumor progression. Another source of variation in the agreement of CA-125 test results with tumor progression may be the varying criteria used to conclude that CA-125 levels have changed in the course of looking for progression. Furthermore, there is no gold standard for measuring change in the disease. Finally, because 510(k) submissions often involve small patient sets with differing

populations, random noise, and post hoc fitting, there is substantial opportunity for bias in the assessment of concordance between CA-125 test results and the disease state.

With these issues in mind, those who consider using CA-125 as a surrogate endpoint biomarker should select the test with care and know it well. They should read the label and information beyond the label to understand the test's performance characteristics. Consistent use of one test can help to avoid problems caused by inter-test variability. In using the test as an indicator of changing disease, it is important to fully define the criteria for interpreting change or trends in test results. It is also important to characterize as well as possible test interactions with other clinical features in order to know the extent of new information contributed by the test.

Panel Questions

Dr. Pazdur opened the floor to the panel for questions and clarifications from the presenters.

Dr. Vergote said he supported everything Dr. Rustin and Dr. Thigpen had said except as it relates to maintenance therapy and PFS. What is important concerning PFS in maintenance therapy is the clinical benefit to patients. PFS is accepted in assessing breast cancer treatment and can be used in maintenance trials.

Dr. Spriggs asked how useful Dr. Rustin's data set is for patients undergoing IP therapy. He noted that Dr. Rustin had said that removal of the IP catheter is a potentially complicating factor. CA-125 is considered an appropriate measure for patients receiving IP therapy at Dr. Spriggs' center. Dr. Rustin replied that the effects are less than anticipated. If it is accepted that IP surgical intervention invalidates the assay, it is unclear how many exceptions one should make. Also, the response is not usually an endpoint during first-line therapy; the crucial variable is TTP. If the patient receives IP paclitaxel and has IP inflammation, CA-125 levels rise, but this could be a false positive.

Dr. Ozols said that the data on CA-125 are persuasive. He asked whether one could use CA-125 if the patient is receiving biological agents, especially monoclonal antibodies, given that these agents might interfere with the CA-125 assay. In response, Dr. Rustin referred to a letter in the *Journal of the National Cancer Institute*, which described a group of patients in the oregovomab trial. The study used two different assays. If one uses a CA-125 assay in patients receiving an antibody and the first capture antibody is goat or rabbit, then one washes away any circulating anti-mouse antibody before adding the murine tracer antibody. But if the first capture antibody is murine, then there could be interference with the assay. One has to check which assay is being used first. Increasingly, the antibodies used have a lower murine component.

Dr. Nerenstone noted that liposomal doxorubicin had received AA on the basis of RR; post-approval follow-up, however, found that TTP was not significantly prolonged, but OS was. That drug would not have been approved on the basis of a TTP endpoint, but it showed a survival advantage. Dr. Nerenstone asked Dr. Pai-Scherf to comment; Dr. Pai-Scherf replied that she was not part of the review team and could not comment on the specifics.

Dr. Pazdur noted that "the devil is in details" with any endpoint that has a degree of subjectivity. How adequately is TTP defined, and who is looking at it? How confident are panel members in looking at TTP? What magnitude would one need? Theoretically, one could incorporate blinding, but it is difficult to do so in oncology trials.

Dr. Eisenhauer noted that Dr. Rustin had presented definitions of response and progression, but the context in which they were developed was important to consider. CA-125 response definition was for use in a Phase II relapsed-disease setting; it was not intended to be a first-line therapy endpoint. Similarly, the definition of progression was developed for first-line trials and reflected a change in practice to use the observation of an increase in CA-125 as reflective of recurrent disease. It would create confusion if patients in a trial whose CA-125 levels rose were not counted as having had an event and were started on second-line therapy. Thus, should the definition of progression apply only during first-line therapy?

Dr. Rustin replied that one of his “proudest papers” (in part because it demonstrated a way to reduce costs) did use that definition of progression. It looked at patients receiving first-line therapy, of whom about 15% progressed, and asked how much would have been saved by not giving the chemotherapy that had been shown by CA-125 to be useless in those patients. The study found that doing so would have saved money. The ICON5 trial had clear details about not changing chemotherapy based on CA-125 unless the patient had symptoms as well as progression.

In addition, the definition of progression was developed for use after first-line therapy. Dr. Rustin said that, when putting together his presentation for the meeting, he thought about the recent discussions he had had about whether one should use CA-125 for relapse therapy, especially now that the number of RCTs involving relapsed patients has increased. It probably would not make any difference to define progression according to CA-125 after first- or second-relapse therapy. However, as a patient has more treatment, CA-125 tends to not fall to normal. Levels go higher and higher; few are in the normal range. Few patients double from 10,000 to 20,000, so CA-125 will pick up fewer progressions.

Dr. Bast noted that the topic involves the issue of dynamic range and asked Dr. Rustin what the average CA-125 level is. Dr. Rustin replied that many patients double from 30 to 60, and some patients have levels in the thousands. Once the level is around 10,000, few patients double again. Dr. Bast noted that in his clinic, few patients with CA-125 levels of 10,000 are eligible for drug therapy. Dr. Rustin noted that first-line therapy includes surgery and chemotherapy; the two modalities cannot be disentangled when evaluating response.

Dr. Pai-Scherf noted that Dr. Bast had shown that CA-125 shedding is not altered by numerous biological agents and asked about similar data for other agents and antibodies. Dr. Bast replied that he had not studied other antibodies.

Dr. Brady noted that GCIG had accepted the definition of CA-125 response and asked whether it had (a) been accepted as a validated response or (b) been accepted in order to move forward and validate the response. Dr. Rustin responded that it had been accepted as a definition that has now been validated. The GOG-111 trial found that the activity of the platinum/paclitaxel regimen was not well characterized by CA-125 response. He asked Dr. Rustin if that finding gave him pause that at least one class of agents is not classified well by response. Could there be other classes? If one uses the no/no-go approach, one could misclassify active agents too early.

Dr. Rustin said it was strange that the GOG-111 data would contradict identical trial data from other studies with the same combination of drugs. The OV10 data (same trial design) found CA-125 to be a reliable marker and may have had more CA-125 results within it. Dr. Brady replied that CA-125 data from the OV10 trial were based on progression, and he was referring to response. Dr. Rustin said that CA-125 was not a marker for response in GOG-111 because the

patients had had surgery as well. It takes a month for the effects of surgery to be lost on CA-125. When he looked at CA-125 studies that found abnormal results with paclitaxel, he found that CA-125 was measured weekly. It can increase in a responding patient. Reanalysis found that it worked perfectly. Dr. Bast said that he and his research team have looked at this issue in great detail and have found no difference between paclitaxel and platinum patients. The only caveat is that RR is sometimes overestimated by 1% or 2%.

Dr. Spriggs observed that several issues concerning discordance between CA-125 findings and CT scan findings are unresolved. Radiographic tools for assessment have progressed remarkably, increasing the ability to detect abnormalities. As a result, no one has a normal scan anymore. When there is a significant discordance, is a CT scan really more reliable than CA-125 as a marker, he asked.

Dr. Arbuck asked Dr. Brady to comment on GOG-111. He replied that the researchers had conducted a path analysis of the data using several definitions of response and had found that CA-125 was not a good surrogate for determining OS. They were trying to use the path analysis as a surrogate to predict a true clinical outcome rather than use a surrogate to predict another surrogate.

Dr. Freedman noted that a large number of different assays are being used; he asked to what extent in GOG-111 the researchers could determine which assays were being used. Dr. Brady responded that his group did not mandate a particular assay; as for how reliable the assays are across institutions, he deferred to Dr. Bast. Dr. Freedman noted that there is no reference standard for CA-125. Dr. Rustin replied that in European trials, the assay must abide by a recognized quality-control scheme as a prerequisite for its use. Patients must have same type of assay throughout the trial. He added that he was surprised that Dr. Brady was trying to correlate response with survival, because correlation is poor for first-line therapy.

Dr. Bast agreed with Dr. Brady's statement that the nadir of CA-125 should not be viewed as an endpoint for drug approval. In some instances, using CA-125 can accelerate or truncate the execution of trials. If one could make a no-go decision on the basis of consistently rising CA-125 levels, one could truncate Phase II trials that were not evaluating useful drugs. Also, if TTP is a useful endpoint, then CA-125 would truncate TTP by a couple of months on average.

Dr. Brady emphasized that surrogate endpoints are validated against clinical endpoints. No one yet knows how to validate a surrogate endpoint against another surrogate endpoint.

Questions and Discussion

Panel members turned their attention to the general discussion questions posed by FDA.

1. Should CA-125 be used as an endpoint in clinical trials intended to support drug approval?
2. Should CA-125 be used as a marker of response, progression and/or relapse in clinical trials intended to support drug approval? If yes, are the CA-125 defined endpoints validated? If not, what data are needed to validate CA-125 as an endpoint?

Dr. Pazdur said he was taking the liberty of rephrasing some of the questions to eliminate redundancy and because the panel had already covered some of the topics. The question FDA was asking the panel to consider was the following: Should CA-125 be used as a marker of response, progression, or relapse in trials intended to support approval of drugs for ovarian cancer? Is CA-125 validated as a marker in those settings? What further trials need to be done?

Dr. Pazdur noted that in an RCT looking at TTP, one would likely use a combined endpoint consisting of radiographic progression and CA-125 level. Dr. Pazdur asked the panel to discuss progression first, then response. He also asked the panel members whether they were comfortable with a composite endpoint.

Dr. Eisenhauer said she was comfortable with the use of CA-125 in that it is already being used in front-line randomized trials as a definition of progression. However, validation of that use is based on a single retrospective data set. The data sets that are currently being accumulated will show more definitively that the use of a composite endpoint (i.e., progression defined as either CA-125 progression or progression according to the RECIST criteria, whichever occurs first) bears the same relationship to OS as the historical objective endpoint did. In a trial that her group closed about a year ago, there were 400 progression events, just 10% of which were detected only by CA-125. Almost always, when CA-125 rises to level of progression, CT scans reveal something.

Dr. Rose agreed with an earlier comment by Dr. Eisenhauer that CA-125 has to be used in a clinical setting to document progression because it is being used that way in the community. In 1992, when GOG-152 and GOG-162 were being developed, he and his colleagues examined second-look data to see what percentage of patients had positive second looks based on their CA-125 level; at 100, all patients had positive second looks, and at 35, 90% were positive. At that time, Dr. Rose's group developed a definition of doubling of CA-125 value and CA-125 value >100 as evidence of progressive disease. Without such a definition, all patients would be lost to second-line therapy without documented progression.

Ms. Solonche asked if a clinician on the panel could explain whether he or she would start retreating a patient in whom CA-125 doubling was the only sign of recurrent disease.

Dr. Nerenstone disagreed with Dr. Rose and expressed concerns about separating lab results from the clinical analysis. Using CA-125 elevation alone as an indication of recurrence and treatment success is problematic. These patients probably have active disease that is not curable and will have a short symptom-free interval (the median is 63 days). Some patients have many months of undetectable disease, however. Separating the clinical benefit from the lab value will subject the patient needlessly to unnecessary toxic treatment. Small trials, not large ones, are the primary concern. Making decisions based on TTP could be wrong, especially if it is determined solely on the basis of CA-125 elevation. Dr. Nerenstone said she does not give chemotherapy to patients in whom the only indication of recurrent disease is a rise in CA-125, although she sometimes uses nontoxic treatments such as hormonal therapy.

Dr. Vergote noted that the scientific question will be answered by a European study in which >1000 patients were randomized and are being treated based on CA-125 increase alone. The practical question is: What do doctors and patients do? It is better to have a clear endpoint based on CA-125 or RECIST than an artificial one. Dr. Rose clarified that he had been referring to trial endpoints, not treatment decisions, in his earlier comments.

Dr. Thigpen said that the question the panel had been asked is the subject of wide debate, and a variety of approaches exist. People have strong feelings on both sides. Because in some instances, decisions about progression are made on the basis of CA-125, that has to be taken into account in trials. The trial approach must be uniform to get an interpretable result. What is the optimal way to incorporate CA-125 into the trial definition so that results are interpretable? Dr. Pazdur suggesting using CA-125-based decisions as a stratification factor.

Dr. Ozols said the goal is to get drugs evaluated more rapidly. Until a consensus is reached on the need to treat patients who have a rise in CA-125, patients in whom an elevated CA-125 is the only evidence of recurrence should be offered the opportunity to participate in clinical trials. These trials should stratify patients according to whether or not elevated CA-125 is the only evidence of recurrence.

Dr. Eisenhower noted that the focus of the discussion is the use of CA-125 progression as an endpoint in the regulatory sense, not as an eligibility criterion in subsequent trials. If in an RCT, one saw a substantial difference in TTP, defined in a way that captured both CA-125 progression and objective measures, would that support drug approval and is it validated? A number of trials have used such composite endpoints prospectively. With a database like that, how would one go about validating that the use of that endpoint had not over- or understated progression? Dr. Rose reiterated that in his earlier study, which used a composite indicator of clinical progression or marker progression, 90% of patients met the criteria for clinical progression and 10% met the criteria for marker progression.

Dr. Pazdur said that researchers approach FDA proposing the use of composite endpoints in RCTs. FDA is interested in RCTs that reflect the population using the drug. He asked whether the definition of progression proposed by Dr. Rustin was acceptable. Dr. Vergote noted that all the current trials that use PFS as the primary endpoint have used the GCIG definition.

Dr. Freedman noted that some patients have normal CA-125 levels to begin with (patients in the “C” category). How does that subgroup behave in relation to clinical indices? Dr. Rustin replied that on the whole, those patients are a better prognostic group because they never had elevated CA-125 levels and probably had smaller-volume disease. Some patients are CA-125–negative because the volume is too small when first measured; they become CA-125 positive later. The A, B, and C categories are artificial. Dr. Freedman said that was the argument for retaining objective criteria. Dr. Nerenstone said that one must be careful about how one specifies for values that occur in between data collection points; otherwise, one can introduce investigator bias.

Dr. Pazdur reiterated Dr. Nerenstone’s concerns over the timing of data collection and noted that this is always a problem with subjective endpoints like TTP. Summarizing, he stated that the committee seemed to be in agreement that using CA-125 as part of a composite endpoint is acceptable, with the caveats mentioned in the discussion.

CA-125 as a Marker of Response

Turning to the issue of response, Dr. Pazdur asked for the panel’s thoughts on a sponsor asking for AA on the basis of single-arm trial in a refractory disease setting, with an RR of X%. The data would probably consist of both radiographic information and CA-125 responses. If the drug were approved, FDA would request confirmatory data or further RCTs.

Dr. Rustin observed that in the real world, patients are trying to get into Phase II trials, but some do not have fully evaluable disease. If the first 15 patients are CA-125–assessable and the RR is high enough, he would advocate that the trial be continued so that there were sufficient RECIST-evaluable patients as well. He asked whether one could put both groups of patients together in a submission for drug approval rather than just the RECIS-evaluable patients. Dr. Pazdur replied that it could be done, but the question before the committee is what weight to give the patient

population in whom response is measured on the basis of a lab parameter only. They could be entered in the trial, but what consideration would one give that population?

Dr. Eisenhauer said that she would not put a lot of weight on those data. She would want to see evidence of the variable that would track with symptom improvement. It has not been shown that a fall in CA-125 tracks with symptom improvement. If such data could be obtained, it would be compelling support for CA-125's utility as a marker.

Dr. Rustin asked Dr. Eisenhauer what valid symptom score she would accept. He was not aware of any validated score but would love to make that correlation. Dr. Pazdur reiterated that the focus was single-arm trials, making it hard to look at symptoms. Dr. Wenzel added that this topic would be revisited later that afternoon in relation to patient-reported outcomes.

Dr. Freedman suggested that two separate groups would be needed if one were going to use CA-125 alone. In his experience, patients could have both stable disease and normalization of CA-125. A group with no objective measurements such as findings from CT scans will have a different RR from a group that has measurable disease.

Dr. Ozols said he was willing to give CA-125 a little more credibility than Dr. Eisenhauer was. CT scans have the same problems: Is a 50% reduction in a 3 cm lesion clinically meaningful? If one is looking for AA, CA-125 is as good as many of the other imperfect measures in use.

Dr. Rose noted that most patients used to present with symptoms, but cancer is now caught before major symptoms occur. Patients are not all presenting with a large pelvic mass. Dr. Spriggs said that he would put more credibility in CA-125 as a measure of drug activity in a Phase II trial. It is at least as good as a CT scan, and it is an approved assay with good reproducibility. Dr. Vergote agreed with Dr. Spriggs and Dr. Rose and added that the results of CT scans can be unreliable.

Dr. Bast said that in the context of AA, he would give priority to studies that correlate symptom control with changes in CT and CA-125. Dr. Pazdur suggested that one would want objective radiographic data in combination with other responses. FDA has some regulatory discretion in that regard. Dr. Eisenhauer added that CA-125 response itself, in the absence of other objective responses, should not be sufficient for AA for a completely new drug. It may, however, be a useful marker for whether to pursue a drug further.

3. What differences in analytical performance characteristics among CA-125 measurement devices should be considered if the marker is used as a surrogate endpoint?

Dr. Pazdur asked Dr. Becker to elaborate on this question. Dr. Becker responded that the question pertains to the differences among assays and comes back to the point, already covered by the panel, that one should be sure that patients are followed using one assay consistently.

Regulatory Endpoints for First-Line Therapy

Is PFS a Valid Surrogate for OS in Advanced Ovarian Cancer? A Meta-Analysis

Marc Buyse, ScD, Department of Clinical Research & Biostatistics, Institut National du Cancer, Paris, France, presented findings from a meta-analysis of trials using PFS as a surrogate for OS in advanced ovarian cancer. He defined *clinical endpoint* as a characteristic or variable that reflects how a patient feels, functions, or survives. A *surrogate endpoint* is a biomarker or endpoint that is intended to substitute for a clinical endpoint. A good "correlate"

may not make a good surrogate. A surrogate endpoint is expected to predict clinical benefit (or harm) or lack thereof.

In studies using surrogate endpoints, the surrogate must fully capture the effect of treatment on the true endpoint, and the effects of treatment on the surrogate and on the true endpoint must be correlated; the latter can be shown in a single trial. Multiple trials, however, are needed to show the effects of treatment on the surrogate and on the true endpoint.

Four trials compared cyclophosphamide-cisplatin (CP) with cyclophosphamide, adriamycin, and cisplatin (CAP): the Gynecologic Oncology Group (GOG, United States); Gruppo Interegionale Cooperativo Oncologico Ginecologia (GICOG, Italy); the Danish Ovarian Cancer Group (DACOVA, Denmark); and Gruppo Oncologico Nord-Ovest (GONO, Italy). Accrual occurred between 1980 and 1986, and median follow-up was >10 years. Data were collected on 1,194 patients (952 deaths). The trials involved 39 centers with >3 patients per treatment arm. Endpoints were clinical response, PFS, and OS. The study looked for correlation between the PFS hazard ratio (HR) and the OS HR and for association between PFS and OS.

Dr. Buyse presented PFS and OS curves showing that the median PFS was about 16 months in the CP group and 8 months longer in the CAP group. Analysis of the correlations between PFS and OS resulted in an R^2 of 0.88, a remarkable correlation. The findings suggest that PFS is not a subjective measure; if it were, the correlation between PFS and OS would not be so strong.

The group-level association between treatment effects (CAP vs. CP) in centers was modeled through linear regression between $\log \text{HR}_{\text{PFS}}$ and $\log \text{HR}_{\text{OS}}$, yielding $\rho = 0.94$ [0.90, 0.97]. The surrogate threshold effect (STE) is the treatment effect on PFS that predicts a significant treatment effect on OS. $\text{STE} = \text{HR}_{\text{PFS}} = 0.55$; in other words, the treatment must reduce the risk of progression or death by at least 45% for a survival benefit to be expected. If one takes an optimistic view, these data show PFS to be an excellent surrogate for OS.

To summarize, individual-level surrogacy establishes a strong association between PFS and OS; this finding is useful for patient management. Trial-level surrogacy establishes a strong association between the effects of treatment (CAP vs. CP) on PFS and OS; this finding is useful for assessing new treatments. Trial-level surrogacy is needed for the FDA to consider that a surrogate is valid. Ideal requirements for validation include the following:

- Individual patient data from multiple comparative (preferably randomized) trials or other analysis units (e.g., centers or countries)
- A range of treatment effects on S and T (heterogeneity is an asset)
- A range of treatment questions to assess any treatment dependency of surrogacy
- Large numbers of observations and analysis units.

Clinical Trial Endpoints for Regulatory Approval: First-Line Therapy for Advanced Ovarian Cancer

Elizabeth Eisenhauer, MD, NCI-Canada Clinical Trials Group, Cancer Research Institute, Kingston, Canada, reviewed the options for endpoints: OS, which is the gold standard; disease progression (measured objectively, by CA-125, or by means of a composite endpoint); and a “symptom-free” period or other quality-of-life measure.

Two GCIG statements address the question of endpoints in front-line ovarian cancer trials. The first, from *Trial Methodology*, says that for first-line treatment of advanced disease, “Both PFS

and OS are important endpoints to understand the full impact of any new treatment. Thus either may be designated as the primary endpoint. Regardless of which is selected, the study should be powered so both PFS and OS can be appropriately evaluated.” The second, from *Standard Approaches*, says that for post-progression therapy, “Although overall survival is an important endpoint, progression free survival may be the preferred primary endpoint for trials assessing the impact of 1st line therapy because of the confounding effect of the post-recurrence/post-progression therapy on overall survival. When progression free survival is the primary endpoint, measures should be taken to protect the validity of analysis of overall survival.” Dr. Eisenhauer noted that the focus of her presentation was PFS as an endpoint for regulatory approval.

Potential Arguments in Favor of PFS as an Endpoint for Regulatory Approval

The meta-analysis that Dr. Buyse presented, other trial results (HR relationships), and disease-related symptoms (inference) all support the validity of PFS as a surrogate for OS.

Validity of PFS as a Surrogate for OS

Dr. Eisenhauer presented a table listing recent trials of experimental vs. standard drug therapies, the HR observed for PFS, and the HR for OS. If the HR is <1 , outcomes in the experimental-therapy arm are better; if the HR is >1 , outcomes in the standard-therapy arm are better. She presented the data as a graph plotting HRs of PFS vs. OS within trials. The unity line is theoretical; above that line, PFS predicted an OS improvement in the experimental arm of greater magnitude than seen. Below that line, PFS predicted worse OS than that seen. The relation between the HR for progression and the HR for OS was very similar, with two exceptions: GOG-132b (paclitaxel vs. platinum arm) and GOG-47 (ca-platinum vs. ca-adriamycin).

The HRs of PFS and OS are similar within trials, suggesting a strong relationship between the behavior of PFS and OS. The data support the argument that PFS is a good surrogate for OS in first-line ovarian cancer treatment. However, that is not the case in trial(s) in which one arm did not contain platinum. In those cases, the salvage platinum therapy seemed to overcome the PFS disadvantage in the non-platinum arm to render survival more similar, although still not identical.

Impact of Post-Progression Therapy on Survival

One argument is that PFS is useful because post-progression therapy obscures the OS effect. This argument is weakened by the data just presented. Except for administration of second-line platinum (when it was not given as first-line therapy), other therapies do not seem to have obliterated the relationship between PFS and OS. Nonetheless, this is a theoretical possibility if the new treatment in an experimental arm is very active, if therapy after relapse is not balanced, and if a high proportion of standard-arm patients get new therapy at relapse. It should not be an issue if the pattern of second-line care is similar between study arms.

Value as an Indicator of Time Without Disease Symptoms

Another argument can be made that PFS is a meaningful endpoint on its own because it is a marker for time without disease symptoms. In front-line ovarian cancer treatment, most patients respond to therapy. At the end of treatment, 50% to 60% have either continuing no evaluable disease (NED) or CR and are thus clinically and radiologically disease free. The median time between the end of therapy and progression is about 10 to 12 months (calculated by subtracting the median duration of therapy of 5 to 6 months from median PFS of 16 to 18 months). As this period is one without both treatment and evident disease, prolonging it by more effective therapy may be meaningful in its own right if, as is inferred, most patients are without symptoms of disease or treatment for that period. Direct evidence to support this is not available, however.

Data supporting this hypothesis may be found by mining disease symptom content of quality-of-life information in many recently completed trials.

Potential Pitfalls in the Use of PFS

When OS is measured, only one date of the “event” is possible. One problem with the use of PRS is that there are several definitions of progression. GCIG has adopted RECIST and its own CA-125 definition (for front-line therapy). Further, when using PFS, assigning a date for the event may be biased even when objective measures are used because these measures are sensitive to the timing of the investigation. An imbalance in assessment times between study arms can lead to apparent differences in PFS that are not real.

One example of this problem occurred in the Genasense trial for melanoma. Using data presented by the FDA reviewer at the ODAC meeting at which this agent was discussed, Dr. Eisenhower argued that the problem of assessment timing is dependent on the size of the absolute improvement in PFS that is being targeted in the trial. This issue is most problematic when a small absolute improvement in PFS is being sought because, in that case, the postulated difference in PFS may be similar to the interval of disease assessment. For example, a 33% improvement in PFS translates into only a 2-month difference when median PFS in the standard arm is 6 months. This issue is not likely to be relevant in first-line trials in which PFS is longer and so improvements of the order of 33% represent much larger absolute differences.

Summary

OS is the gold standard, and trials should be powered to assess it. Nevertheless, PFS is also an appropriate endpoint for regulatory approval in the front-line setting. Good evidence supports its use as a surrogate for OS. Second-line therapy appears to have little impact on the effect of PFS on OS, but it could have an impact if administration of highly active post-progression treatment is substantially imbalanced in randomized arms. PFS may correlate with freedom from disease-related symptoms; additional data are needed to support this, however.

One potential pitfall of PFS is that it is sensitive to the timing of the investigation; this is unlikely to be relevant in first-line trials but may be more relevant in second-line trials where smaller absolute differences are targeted. In addition, PFS definitions are shifting to incorporate CA-125 measures.

Questions to be addressed are

- Does PFS correlate with freedom from symptoms?
- Does the use of CA-125 to define PD change the relationship of PFS to OS?

Dr. Pazdur commented that this endpoint highlights the need to strengthen regulatory standards, not weaken them. Methodological issues are present in discussion of endpoints for many other cancers. Meticulous attention must be given to this endpoint.

First-Line Therapy in Ovarian Cancer: Surrogate Endpoints for Accelerated Approval

Mark F. Brady, PhD, Gynecologic Oncology Group, Statistical & Data Center, Buffalo, NY, began by noting that the results described in his presentation are from work in progress and are not to be used for publication or reference.

To validate a surrogate endpoint, one must show that the treatment's effect on the true endpoint is accounted for by its effect on the surrogate endpoint. Often, the surrogate is shown to be predictive of a good outcome; this is not sufficient evidence to validate a surrogate.

For example, it can be shown that when ovarian cancer patients' CA-125 levels return to normal, the difference in survival is huge when compared with patients whose values do not return to normal. Some treatments can cause the CA-125 values of a greater proportion of women to return to normal levels. But is it the treatment's ability to drive CA-125 down that accounts for increased survival? One needs to assess the evidence at both the individual patient level and the trial level.

Dr. Brady described an analysis of six GOG randomized trials involving patients with recently diagnosed, optimally debulked, advanced epithelial ovarian cancer and eight GOG randomized trials involving patients with recently diagnosed, suboptimally debulked, advanced epithelial ovarian cancer. The analysis involved a total of 5,826 patients, 14 trials, and 30 first-line treatments.

Dr. Brady presented graphs showing the PFS and OS treatment HRs from these 14 GOG trials. This trial-level evidence indicated that the correlation between these observed HRs was 0.84. Adjusting this correlation for the fact that these are estimated HRs, the estimated correlation of the true HRs is 0.93, which is very consistent with Dr. Buyse's results. The GOG data assessed at the patient level with Kendall's tau and median concordance also supports the validity of PFS as a surrogate for OS. These analyses did not include trials where patients could not receive platinum as part of their first-line therapy but could receive it as part of their subsequent therapy.

Summary

Use of PFS as a surrogate is justified for the following reasons:

- Increasing disease burden is in the etiologic pathway to death.
- Clinical symptoms sometimes accompany progression.
- PFS duration is usually unperturbed by salvage therapies.
- PFS comparisons mature more quickly than survival comparisons.

Drawbacks to using PFS as a surrogate are that the onset of clinical progression depends on assessment timing, PFS is susceptible to bias due to differential timing of assessments, and PFS may not capture all direct effects of treatment.

When the PFS data from a treatment comparison indicates that a new treatment is active, but this is not supported by the survival data, it can be difficult to determine if this is due to bias (e.g., the Genasense trial) or to active second-line treatments (e.g., the first-line ovarian cancer treatment trials comparing platinum and non-platinum regimens).

A good Phase III trial design in advanced ovarian cancer implements procedures that protect the validity of both the PFS and OS endpoints. To protect the validity of PFS and OS in Phase III trials if PFS is the primary endpoint, double-blinding of treatments and standardizing of the schedule and procedures for disease assessments should be considered. Observing a small but statistically significant difference may not be enough. Consider the direct clinical relevance of the PFS effect size and the predicted benefit in the true clinical endpoint. Interim analyses that are based on PFS also should consider the interpretability of secondary endpoints (i.e., survival). If survival is the primary endpoint, evaluate the potential effects of subsequent therapies. Is there evidence that post-study treatments were available that could have differentially altered survival?

Questions and Discussion

Dr. Pazdur opened the floor to the panel for questions and clarifications from the presenters.

Dr. Pazdur noted that in other disease settings, surrogate endpoints face difficult criteria. Few traditional endpoints in oncology truly meet the criteria. Correlation should not be confused with surrogacy. Many treatments are marginal in their impact on both survival and TTP. When discussing surrogate endpoints and trying to establish a link to ultimate clinical endpoints, the magnitude of the impact is often not discussed, for several reasons. For example, if a trial is not sized to measure survival, a survival endpoint is short-circuited. If only PFS is assessed, the result will be progressively smaller, more underpowered trials. The magnitude of benefit for PFS must be examined. How do we characterize and discuss PFS? Traditionally, we use means. Should we use HRs? PFS is a difficult endpoint, he concluded.

Dr. Weiss said that she struggles with this issue: If a drug has an effect on PFS, and there is large crossover to the treatment arm, how can one protect the validity of OS? OS will not be evaluable. Dr. Eisenhauer replied that the issue relates to the power of the study, not to the analysis. One must have adequate power to observe survival. Dr. Thigpen noted that one could power for survival by disregarding the impact of post-progression therapy.

Dr. Arbuck noted that ovarian cancer generally is not the first disease for which a drug is approved. Dr. Nerenstone added that a clinician needs to know whether using two drugs does not provide a survival advantage. Dr. Brady said OS and PFS both provide important information. For example, if a treatment demonstrates a PFS advantage but no survival advantage and the explanation is that subsequent therapies eliminated the survival advantage, this can be clinically informative. It may mean that the study treatment did not need to be initiated when it was, since the subsequent interventions were just as good. In cancer studies, the only way one can evaluate new treatments is in a milieu where subsequent therapies are available.

Dr. Pazdur asked the panel to discuss whether PFS itself should be considered a clinical benefit. Doing so would obviate the whole discussion of surrogacy. Dr. Ozols said improvement in PFS is intrinsically of benefit to patients.

Dr. Pazdur noted that as trials are powered for survival, an overpowering of TTP will occur. A statistically significant difference does not indicate a clinically significant difference. How, therefore, can we characterize progression?

Dr. Eisenhauer said that in first-line therapy, it is reasonable to infer that when most patients end treatment without disease or symptoms, there is a window of time that, if prolonged, is meaningful to the patient. Existing databases could be mined to confirm that. Then the question becomes: What magnitude of increase in time without symptoms is important? Dr. Pazdur noted that this question raised the risk-benefit issue.

Dr. Buyse said it is important to continue to look at the relation between the endpoints. Dr. Pazdur replied that smaller trials are not desirable. People look at the number of patients they think they can accrue to a trial, then start estimating effect sizes.

Dr. Wenzel said that one must examine symptom-free intervals and their importance to the patient. Studies are not designed with that in mind. Dr. Pazdur said FDA suggests that manufacturers measure time to symptomatic progression, but that approach is not commonly

used. Even if a therapy were interrupted, FDA would consider that measurement approach appropriate.

Ms. Solonche noted that there will always be patients who will want to be on treatment the minute a symptom appears; others will want a break. It is a difficult situation for most women, with no definitive answer. Dr. Arbuck replied that this was an important point. How important is it to patients to have a new drug available based on time to occurrence of actual symptoms? Is it a benefit if the occurrence of symptoms is delayed? Dr. Pai-Scherf replied that a long period without symptoms is an improvement, but how realistic is it to be able to capture that? What would be the best timing for data collection? Dr. Wenzel said that it would be optimal to design the study prospectively, not retrospectively. It is possible to capture data in a meaningful manner.

Dr. Vergote noted that in GOG-111, the difference in PFS was only a few months, but that was an exception. In other Canadian and European trials, the difference in PFS was 12 months.

Dr. Buyse raised the issue of how to quantify the impact of treatment. A 3- to 4-month improvement in median PFS is not meaningful; HR is a much more meaningful measure and is more commonly used. An 0.75 HR equals a 25% reduction in the risk of progression, which is meaningful. HR is also more robust and less sensitive to baseline risk than median PFS.

Dr. Basen-Engquist said that one disadvantage of retrospective studies is that some of the measures are not enthusiastically endorsed by FDA for regulatory purposes. Dr. Pazdur said the major issue is that symptom analyses need to be prospectively defined in the study protocol. Frequently, quality-of-life instruments are tacked on at the end with little thought. Another problem is lack of blinding. With some of the newer drugs, which are oral and less toxic, an increased number of blinded trials have taken place. These trials involve asymptomatic populations, so the drugs cannot show improvement of symptoms. Dr. Basen-Engquist noted that many patients who present with progression may not have symptoms.

Dr. Pazdur said FDA could accept PFS as an endpoint in a first-line study. If the drug produces a clinical benefit in first-line therapy, it is hard to say, from a regulatory perspective, that the drug is not of clinical benefit in a refractory disease setting.

Dr. Brady said one approach would be to try to summarize the data available from all of the randomized trials of treatment for refractory disease conducted by the cooperative groups participating in the GCIG.

Dr. Ozols agreed that if PFS is a good endpoint for first-line treatment, it is good for refractory disease. Patients understand what progression means. Anything that delays progression is important.

Dr. Eisenhauer said that in first-line therapy, one does not know who might be alive 10 years from now. Goals for treatment are to increase the proportion of long-term survivors. If a patient has no recurrence after CR, the longer that recurrence-free period is, the more likely it is that the goal (long-term survival) will be achieved. In second-line therapy, when patients have disease with an inevitable end, it seems more important to show that treatment is doing more than changing the time at which there is a 20% increase in disease progression on CT scan. If the patient is asymptomatic and stays asymptomatic when she has that increase, then from the patient's point of view, nothing is different. The use of progression as an endpoint in second-line therapy may mean something different to the patient than it means after first-line therapy. Dr.

Spriggs added that in a first-line setting, the real value of PFS is as a surrogate. In second-line treatment, PFS has intrinsic value independent of its ability to predict survival.

Dr. Nerenstone noted that another complicating factor is third-, fourth-, or fifth-line therapy. Stable disease means something, but it is difficult to catch or quantify. Adding CA-125 to the mix and not going by clinical parameters could result in a “mishmash” of data.

Dr. Thigpen noted that everyone thinks PFS is important in second-line therapy, so why the hesitance to use it as an endpoint in first-line treatment?

Dr. Freedman said that as disease moves toward refractoriness and as the patient’s motivation and energy with regard to treatment change, QoL becomes more important. Dr. Wenzel noted that in second- and third-line therapy, clinicians are in a good position to collect data on whether patients are benefiting from palliative chemotherapy.

Dr. Pazdur raised the issue of difficulty in measuring PFS. Protocols require meticulous detail, and the timing and symmetry of scans are important. He asked the group for their thoughts on collecting measurements at a single time point. Would doing so obviate some of the need for multiple scans and reviews? In response, Dr. Rose suggested a 6-month time point.

Dr. Pazdur asked how a pharmaceutical company approaches the need for x-ray analyses. Dr. Arbuck noted that these analyses are not just a burden on the company and on FDA but also impose a burden at the clinical site. A single time point might be appropriate in a setting in which enough was understood about a drug and its effects that a good endpoint could be selected.

Dr. Bast noted that heterogeneity is a problem with some second-line treatments. CT at 6-month intervals sounds good but is hard to achieve. Dr. Eisenhauer’s earlier comment—that only about 10% of patients in first-line therapy were picked up by CA-125 alone—is intriguing. CT scans cannot be done every 3 months, but CA-125 or other biomarkers can be assessed. Is it possible to time CTs more intelligently to minimize costs, both in front-line therapy and in Phase II studies in patients with recurrent disease?

Dr. Arbuck said that Dr. Bast raised a critical point about which she would like more discussion. PFS was a good endpoint for some studies discussed this morning. The reality is that when trials are designed, the cost of frequent scans and the independent review that is sometimes required of those scans is a huge issue. Since ovarian cancer is not the first indication for which a company will generally seek a drug’s approval, the study drug is often available to patients on the control arm of the study when their disease progresses. Costs limit the ability to conduct research. How can we incorporate CA-125?

Dr. Rose noted that part of the issue is determining an acceptable endpoint. Should it be PFS? RR? If one is willing to divorce from RR (which is fine because it is so difficult to measure), it will be easier to use less frequent measures. Dr. Pazdur noted that panel members had expressed support for incorporating CA-125 as an endpoint. Dr. Eisenhauer said that in the relapsed-disease setting, CA-125 seems more exploratory for regulatory approval. Everyone is using CA-125 as part of the composite definition of progression in front-line therapy. Dr. Pazdur said that in a randomized setting, one would be looking at a time-to-event endpoint rather than an RR endpoint because doing so captures an effect on an entire population. Progression and time points need to be defined in the protocol; it is up to the sponsor of a drug approval application to do that.

Dr. Vergote suggested that FDA use the GCIIG consensus to determine the timing of CTs and CA-125 rather than leaving it up to the sponsor.

Dr. Buyse noted that the issue is not so much frequency as the possibility of bias. In double-blinded trials, it does not matter how frequently one measures (CT scans or CA-125). In open-label trials, PFS assessment might be biased by knowledge of treatment. In that case, he would support using a single time point for open-label trials. Dr. Pazdur noted that many FDA statisticians do not support this approach because it results in a loss of statistical power in studies. Dr. Buyse said that was a sacrifice worth making.

Dr. Bast said that, with regard to AA endpoints, taking the emphasis off RR and looking at PFS for refractory disease requires a composite index. It begins to make sense with targeted therapies. Dr. Pazdur said that raises the issue of what happens if a trial shows an increase in PFS without a corresponding increase in OS.

1. Is PFS a reliable surrogate for OS in randomized front-line ovarian cancer trials? For second-line (or third-line) trials?

Dr. Pazdur summarized for the panel: For front-line trials, the panel agreed that PFS is a reliable surrogate for OS.

2. If yes to either, how should progression be documented: objective measures and marker change using definitions from GCIIG?

The panel agreed that progression should be documented using objective measurements and marker changes using definitions from GCIIG: a composite endpoint comprising x-rays, CA-125, clinical markers, and patient-reported outcomes.

3. If PFS is *not* a valid surrogate for OS in first- or second-line treatment, can it stand alone as a reasonable endpoint on which to approve new agents? Is the answer to this dependent on the absolute gain in PFS? On changes in disease-related symptoms?

N/A because of the answer to Question 1.

4. Should an improvement in DFS without an improvement in OS support approval of a new drug or indication in advanced ovarian cancer? If so, in what patient populations? First-line treatment? Second-line platinum sensitive? Second-line platinum refractory? Third-line and beyond?

Dr. Brady noted that in past first-line ovarian cancer trials PFS and OS have consistently tracked each other, except where known active treatments (platinum or paclitaxel) were not included in the first-line regimen but were available for subsequent treatment. If a big difference is seen in PFS but not in OS, the study has a methodological problem. More questions have to be asked.

Dr. Pazdur responded that some of those questions would involve matters such as powering of a trial; patient crossover; or subsequent therapies, which often are poorly captured in oncology trials. If one sees an effect on PFS but not on OS, could that be a confounding effect of subsequent therapy? Dr. Brady said he does not have a problem with recording the subsequent therapies, but it is unclear what one would do with this data. There is no good way to adjust for these subsequent nonrandomized therapies to “clean up” the data analysis. Dr. Pazdur replied that descriptive analysis is more likely.

Dr. Ozols said that improvement in PFS was an absolute indication for approval. Dr. Freedman added that he would like to see approval of post-treatment PFS in combination with symptom approval as an endpoint in studies of refractory disease. Dr. Thigpen concurred with Dr. Ozols.

Ms. Solonche said she did not think a first-line treatment drug with good DFS but no improvement in OS would be a benefit. Dr. Arbutk replied that it is important to make sure that the drug is not causing a decrement in OS. Dr. Pazdur said that FDA always looks at OS. Dr. Brady said evidence that the drug is the active agent would be needed.

Dr. Spriggs agreed with Dr. Ozols that PFS is absolutely an appropriate endpoint to demonstrate efficacy in front-line ovarian cancer therapy. In maintenance and second-line therapy, the circumstances are less clear and PFS might be an appropriate early-approval endpoint to be confirmed by subsequent follow-up if the toxicity profile was modest. In Dr. Brady's analysis, the studies that do not track OS/PFS have important crossover effects. Unless one is going to delay the general use of a drug until all data are available, it is necessary to accept PFS despite the risk. Dr. Vergote concurred.

Dr. Eisenhower noted that the panel had seen some compelling data showing that PFS differences and OS differences move in the same direction. The scenario of concern to the panel members has not happened yet, and she wondered if it would.

Audience Questions and Comments

Paula Kim, Translating Research Across Communities, an organization representing the patient perspective, noted that to the ovarian cancer community, the important aspects of using PFS as an endpoint are: What does it mean to the patient? Is there a benefit? Are we extending the time between treatments, the number of days patients feel good? That is an important clinical benefit. Benefit to the patient must be the central concern. In lethal cancers, most patients go on first-line treatment, do not respond, and go on second-line treatment. In that setting, PFS must be given stronger consideration as an endpoint.

A member of the audience who identified himself only as Adrian commented on a trial that saw an increase in disease progression and no change in OS. This trial involved a second-line ovarian cancer treatment, so the scenario about which the panel had expressed concern has happened. If clinically meaningful data consists of OS or improvement in symptoms, then if there is no increase in OS a trial should at least show improvement in symptoms.

Regulatory Endpoints for Maintenance Therapy

Maintenance Therapy in Ovarian Cancer: PFS and OS as Endpoints of Therapeutic Clinical Trials

David R. Spriggs, MD, Memorial Sloan-Kettering Cancer Center, described endpoints for clinical trials for recurrent disease. Between 20% and 25% of all ovarian cancer patients are diagnosed with FIGO stage I and II (early stage) disease. Essentially all early-stage patients will be in a clinical CR after surgery and chemotherapy; 25% of those patients will relapse and could benefit from maintenance or consolidation therapy.

Between 75% and 80% of patients are diagnosed with FIGO stage III and IV disease; 75% of those patients will achieve clinical CR after cytoreductive surgery and carboplatin/paclitaxel therapy. Approximately 75% of patients in CR will relapse. Overall, 60% to 65% of all patients with ovarian cancer could potentially benefit from effective maintenance therapy.

By convention, consolidation therapy is of relative short duration, such as high-dose chemotherapy with a transplant, IP ^{32}P , or whole-abdominal radiation (WAR). Maintenance therapy is of extended duration (6 or more months) or continuous treatment until disease progression. Maintenance therapy for patients who respond to initial treatment includes chemotherapy or biological agents. Consolidation therapy includes IP ^{32}P or radioimmunoconjugates, WAR, high-dose chemotherapy with bone marrow transplant, and IP chemotherapy.

A small number of RCTs for consolidation therapy have been reported and were recently summarized (Sabbatini and Spriggs, 2006). Trials examining high-dose chemotherapy with peripheral stem cell rescue; IP ^{32}P for negative second-look patients; and use of an yttrium-labeled beta emitter conjugated to a monoclonal antibody targeting HMFG1 showed no substantial benefit for survival or recurrence (Verheijen et al, 2006). Piccart et al (2003) compared four cycles of IP cisplatin versus observation in patients with pathologic CR; the trial did not meet its accrual goals and is noninterpretable.

Studies comparing WAR and maintenance chemotherapy (Sorbe, 2003a, 2003b) found a slight advantage for WAR, but the numbers were small. Randomized trials of extended initial chemotherapy found no significant change in PFS or OS. Results from adding a third drug (epirubicin or topotecan) for patients who have pathological CR or for all patients have been disappointing. Finally, GOG-178 compared patients in clinical CR randomized to maintenance paclitaxel for 3 or 12 cycles. Like the IP chemotherapy trial, it did not meet its accrual goal, but the study was stopped early because of the clear PFS advantage of 12 cycles (Markman et al, 2003).

A number of biologic therapies have been studied, including interferon alpha, IP interferon, and oregovomab; none of these studies showed a significant advantage for maintenance therapy. The results suggest that although there is a good rationale for maintenance therapy, it has been difficult to translate that rationale into significant patient benefit.

In conclusion, neither maintenance nor consolidation therapy has been shown to improve survival. One trial of WAR and one trial of IV paclitaxel demonstrated improvement in PFS. The toxicity of WAR and paclitaxel is substantial, however.

Robert F. Ozols, MD, PhD, Fox Chase Cancer Center, reviewed the GCIG OCCC 2004 consensus statement. The recommended primary endpoint for future clinical trials in ovarian cancer (for maintenance following first-line therapy) is OS; a minority voted that in certain situations PFS could be a primary endpoint. Those “certain situations” are nontoxic therapy, biological therapy that will not affect subsequent chemotherapy, and clinically significant improvement in PFS. Many considerations determine which situations are appropriate for using PFS as an endpoint in maintenance trials.

Reasons to use PFS as an endpoint in maintenance therapy include its intrinsic clinical benefit. Patients understand that relapse is linked with death, so patients who have finished primary therapy and are put on a maintenance therapy that delays recurrence and subsequent—and highly toxic—therapy may enjoy improved QoL. Treatment effects are not confounded by second- and third-line treatments. Using PFS an endpoint enables quicker evaluation of the benefit of new therapies.

One reason not to use PFS as an endpoint for maintenance therapy is that maintenance therapy is toxic and affects QoL. Maintenance therapy may make treatment at clinical progression more difficult if the patient has used up her “allotment” of tolerable chemotherapy. About 50% of patients have macroscopic/microscopic (residual) disease detectable only by second-look laparotomy. In other words, the treatment is not maintaining CR but addressing residual disease.

Dr. Ozols listed several ongoing RCTs of maintenance therapy, two of which are using PFS as a secondary endpoint and one of which is using it as a primary endpoint. The results of these trials will not be available for some time.

In reviewing the strategy for and regulatory approval of drugs for maintenance therapy, the issue is whether they are going to be greatly related to PFS with regard to their relationship to toxicity.

Discussion

Dr. Thigpen asked whether Dr. Ozols was recommending that OS be the only endpoint or whether other endpoints should be considered. Dr. Ozols replied that other endpoints should also be considered. Toxicity is an important consideration. If a treatment is nontoxic, the acceptable magnitude of OS could be less; if the treatment is highly toxic, the bar would need to be higher. PFS should be considered a regulatory endpoint for patients on maintenance trials.

Dr. Pazdur noted that one point brought up by FDA in preparing the questions for the panel was whether PFS alone was an acceptable endpoint to support regular approval in studies of maintenance therapy. If a drug is given AA, and if a significant number of patients have not received the drug that is now commercially available, the approval would create the quandary of an approved drug for which a trial could never really be completed.

Dr. Ozols noted that the plan with the GOG trial of bevacizumab is to seek AA if the trial meets the endpoint of PFS, then to continue with OS as the endpoint for final approval. It is difficult to do an RCT looking for OS. Given a trial with a biological agent, such as the antibody against CA-125, that shows a significant improvement in PFS with essentially no toxicity, does one do a repeat study looking for OS? It would be difficult to recruit patients for such a trial.

Dr. Nerenstone said that in some cases clinical trials had shown that new biological or targeted therapies (e.g., trastuzumab) were effective in the adjuvant setting after those agents were already available for metastatic disease. Also, because some therapies are very expensive, it is important to know if OS is not affected even if PFS is improved. Dr. Pazdur noted that for regulatory purposes, FDA does not assess financial considerations.

Dr. Eisenhauer noted that any substantial gain in PFS from first-line therapy is normally associated with a gain in OS. The problem in the maintenance setting is that no maintenance trials have been positive for survival, so no determination can be made as to whether a similar relationship between gains in PFS and OS exists in this setting. In the maintenance setting, if PFS were the primary endpoint, one would need to do everything possible to protect the ability to interpret OS; that may mean not allowing patients in the nontreatment arm to get the experimental treatment. This is fairly easy to do when a treatment is not marketed.

The other reason favoring PFS as an endpoint for first-line combination regimens is the period without treatment effects or disease symptoms that prolonged PFS may represent. One of those variables may not be relevant in a maintenance trial. If the maintenance treatment produces

symptoms in 90% of patients, it is important to know that, for example, giving maintenance therapy for a year produces a gain in OS.

Dr. Ozols noted that in two of the trials he referred to, PFS is the endpoint. The use of a taxane in maintenance trials may compromise patients' ability to receive taxanes when their disease recurs. If biological agents lead to an improvement in PFS, they will not alter the patient's ability to receive subsequent effective second-line treatments. Theoretically, if one postpones progression for 3 or 4 months using a biological agent, one could extend the disease-free interval with platinum therapy and get a better second-line effect on survival. It may be that the subsequent treatment, not the initial treatment, affects survival.

Dr. Spriggs said that the line between consolidation and maintenance treatment and the end of primary treatment is arbitrary. If PFS is an acceptable endpoint for first-line therapy, it strains credibility to say that consolidation therapy is different from first-line therapy. In some cases, the drug alone is inactive and is only active in the context of small-volume residual disease. For example, erlotinib alone does not have a huge impact, which would support the argument that crossover is not going to have a large effect on overall outcomes. Agents like bevacizumab, which clearly have single-agent activity in the treatment of established disease, will be a "hard sell," and researchers will be left with PFS and no survival data in that setting. GOG-178, which had an early-stopping rule based on marked PFS effect, exemplifies this dilemma.

Regulatory Endpoints for Subsequent Therapies

Endpoints in Platinum-Sensitive Ovarian Cancer

Dr. Spriggs presented a diagram of a disease-states model to describe potential homogenous groups for clinical trials. His focus was the potentially platinum-sensitive group. The CA-125 nadir is highly predictive in the first CR group and represents an important stratification factor when deciding on maintenance versus consolidation therapy for that group of patients (Crawford and Peace, 2005; Markman et al, 2006).

The potentially platinum-sensitive group has recurrence >6 months after the last dose of platinum therapy. One can stratify TTP according to the time since the last platinum dose. Dr. Spriggs listed three trials for treatment of platinum-sensitive disease (ICON4, AGO, and the subset analysis of the trial that led to the approval of pegylated liposomal doxorubicin [PLD]). CA-125 was not listed as a progression endpoint in any of these trials. The two trials that reported both endpoints (ICON4 platinum/paclitaxel and PLD) favored both PFS and OS.

Two of three trials showed that PFS and OS are correlated, but the data set is limited. Dr. Spriggs noted that the panel had discussed CA-125 response as a yes/no question, but in the context of certain chemotherapy agents, the response does not happen instantly. If one looks at the three common drugs—carboplatin, liposomal doxorubicin, and topotecan—liposomal doxorubicin stands out as being an agent for which the time to CA-125 response is substantially delayed. Only about half of patients will have a response by CA-125 after the first cycle of therapy. Many clinicians know this, but it is hard to find data; properties like half-life may have a profound effect on the ability to identify progression or response, particularly early in the course of treatment. Timing can matter a great deal.

Commentary

All the trials used classic chemotherapy. Only ICON4 was a prospective trial with OS as the primary endpoint. The AGO trial was sized for PFS only, and the PLD trial is a subset analysis.

None of the trials used a consistent definition of CA-125 as an endpoint and the relationship between PFS and OS is not proven in this setting. Not all agents have the same kinetics of response.

Questions for Discussion

- What constitutes a “sufficient” data base for the use of CA-125 response as an endpoint in studies of a new agent?
- Are tertiary treatments sufficiently effective to move the OS endpoint in a second-line therapy?
- Do disease bulk, surgical status, or other factors significantly alter outcomes?

Second Complete Remission

Patients in their second CR are likely to have a relatively short CR but may be relatively successfully entered into trials as a group with residual disease. PFS may be a useful endpoint with this group. The clinical CR group is defined as normal CA-125 with no lesion >1 cm and a negative physical examination.

Fewer than half of patients re-treated with platinum achieve a second CR; the duration is around 13 months and may be extended by second debulking surgery. The duration depends in part on the CA-125 nadir and is usually shorter than the first remission; it is longer than first remission in perhaps 3% to 10% of patients.

In the second-remission setting, Dr. Spriggs’ group, under the direction of Paul Sabbatini, has conducted several trials examining uniform trial eligibility criteria: entry within 3 months of end of therapy and NED. Follow-up consists of CT every 3 months and monthly assessment of CA-125. Progression is defined as either a new lesion on CT scan or a confirmed CA-125 >70 U/mL. The population can be identified as relatively homogenous and, with appropriate stratification, can be studied in the second-remission setting. This is a population that offers a clear opportunity. The OS endpoint is probably too remote; PFS is probably still appropriate for these patients. CA-125 stability may be an issue in treatment with biologic agents. Also, much is unknown about the biology and function of MUC16. Two other mucins have a similar biology. MUC4 is present in several gastrointestinal cancers; MUC1 is present in a variety of cancers. Dr. Spriggs presented a graph illustrating the effects of various biologics on CA-125 release.

The small numbers of RCTs make it difficult to make endpoint recommendations in this setting. CA-125 appears to be a powerful stratification tool and is generally predictive of progression in recurrent ovarian cancer treated with classic chemotherapy agents. The reliability of CA-125 as a primary endpoint in this setting may depend on improved knowledge about the biology of MUC16.

Evaluation of Chemotherapy Efficacy in Platinum-Resistant Ovarian Cancer

Peter G. Rose, MD, Cleveland Clinic Foundation, began by stating that second-line therapy goals are *primum non nocere* to improve symptoms, prolong symptom-free survival, optimize QoL, and prolong OS. The significance of tumor shrinkage with regard to outcome is not known. For recurrent ovarian cancer, the best measures of treatment efficacy could be RR, median PFS, median OS, or the percentage of patients who are progression-free at 6 and at 12 months.

Potential platinum sensitivity can be defined as at least a partial response to primary therapy with an organoplatinum compound; it can be further subdivided into platinum-free intervals of <6

months, 6 to 12 months, and >12 months. Platinum resistance can be defined as progression on primary therapy, demonstration of less than a partial response on primary therapy, and recurrence of disease within 6 months of completing a platinum-based regimen. Even for platinum-sensitive recurrent ovarian cancer, platinum resistance eventually develops. The RR to second-line agents for platinum-resistant disease is low and survival is limited.

Dr. Rose presented a schematic diagram of the course and treatment of recurrent disease and noted that patients with platinum-resistant disease are a more homogenous population with respect to RR, PFS, and OS. RRs are low (10% to 15%), CRs are seen in only 1% to 2% of patients and stable disease is seen roughly twice as often as an objective response (20% to 30%).

Measurable disease has been required in Phase II trials, but measurement of response in platinum-resistant ovarian cancer is challenging. In physical exams, large variation exists among different observers for the same object. On CT, 30% to 40% of responses are not confirmed by independent radiology review. In one study, the RR to topotecan decreased from 25.8% to 15.2%. In another study, 10% of patients of patients retrospectively evaluated from three prospective trials with docetaxel could not be evaluated because of equivocal CT findings.

Stable disease may be the best response in platinum-resistant ovarian cancer, but the benefit of stable disease is controversial. Some feel it is not of benefit and might only reflect the natural history of the disease. Others have stated that “the attainment of stable disease with acceptable levels of toxicity is a valid clinical endpoint” (Markman and Bookman, 2000).

An analysis of data from the Copenhagen Database for Ovarian Cancer compared patients who had complete response, partial response, or stable disease following second-line therapy. Each category was significantly different from patients with progression-free disease. Patients with partial response and stable disease had no significant difference in outcome. Cesano et al (1999), in a study comparing regimens of paclitaxel and topotecan, found that stable disease is equivalent to partial response as a surrogate for survival.

Eleven GOG trials in platinum-resistant ovarian cancer (GOG-126B-E, 126G-L, and 126N) examined patients treated with a variety of drug regimens who experienced a first recurrence within 6 months of therapy. Outcomes for patients in these 11 studies were compared with those for patients in two different trials for two approved second-line agents: liposomal doxorubicin and topotecan. Median PFS, median OS, PFS at 6 months, and RR correlated well across the trials. A comparison of RR in the GOG-126, liposomal doxorubicin, and topotecan trials found no statistically significant difference across trials.

One problem with looking at median PFS is that if fewer than 50% of patients in the study group respond or have stable disease there may be no effect on median PFS, although a subset of less than 50% may benefit. A comparison across the same trials (the 11 GOG trials and the PLD and topotecan studies) found statistically significant differences in median PFS and in the proportion of patients who were progression free at 6 months and 12 months. Statistically significant differences also were found in OS across the trials.

Graphing the different endpoints revealed that certain agents seem to be correlated with significant improvement in median PFS, improvement in PFS at 6 months, and reasonable improvement in OS and RR. Dr. Rose presented graphs showing examples of surrogates predicting longer median survival, average median survival, and shorter median survival. Some agents had a high RR but only average cytostatic qualities and effect on median PFS or PFS at 6

months. Some agents had high cytostatic qualities but only average cytotoxic activity and modest RR. Some had a low RR but high survival.

The GOG-126 data showed that all the measures of clinical efficacy studied (RR, median PFS, proportion of patients progression-free at 6 months, median PFS, and OS) are correlated. Although response is a surrogate for clinical benefit, RR was unable to reliably predict large variations in PFS and OS in these studies of platinum-resistant ovarian cancer. Some regimens appear to have different degrees of cytotoxic and cytostatic activity—that is, some agents have high RRs with minimal effect on PFS or OS. PFS is important because it takes into account both responding patients and patients with stable disease. PFS is a “cleaner” endpoint than OS because of the effects on OS of subsequent therapies that patients may receive. RRs are only one measure of chemotherapy activity that should be considered in evaluating chemotherapy efficacy in this patient population.

Questions and Discussion

Dr. Pazdur opened the floor to the panel for questions and points of clarification.

Dr. Ozols asked whether patients with platinum-sensitive disease who were randomized to PLD had better survival because of PLD or because they could tolerate subsequent therapies better.

Dr. Spriggs replied that it was hard to tell; it is important to recognize that if one does not separate platinum-sensitive and platinum-resistant patients in recurrent-disease trials, one can end up “with a muddle.” Some post-treatment effects are probably unavailable and unmeasured.

Dr. Pazdur noted that the panel was emphasizing subsequent therapies, but when FDA sees an application that shows improvement in OS and a problem with TTP, the agency wonders whether TTP was measured appropriately. The agency places more emphasis on improvement in OS because of the relative lack of bias in its ascertainment.

Dr. Ozols noted that platinum therapy was clearly an issue with regard to the outliers in the data Dr. Rose presented. Dr. Thigpen noted that as a result of the plethora of active drugs, survival has become an endpoint that is probably equally biased in terms of what one chooses as subsequent therapy. Dr. Brady said that if a trial finds a difference in PFS but not in survival, the researcher must look elsewhere for additional information. The Mayo trial of platinum versus no platinum did show a difference in PFS and OS. That trial was not affected by crossover, but other trials have been. It was a small trial, but the effects were so large that it was terminated.

FDA Questions

Dr. Pazdur noted that the panel had covered some of the question topics in its discussion and said that he would paraphrase some of the questions.

1. Is PFS alone an acceptable endpoint to support regular approval in studies investigating the role of maintenance therapy following first-line therapy? Accelerated approval?

Dr. Pazdur noted that in the morning session, the panel members expressed some support for using TTP and PFS with a radiographic and CA-125 composite endpoint for approval of drugs in first-line treatment. In second-line and subsequent therapy, if one considers PFS as previously defined to be of clinical benefit in itself, how should it be viewed in more refractory disease settings? The context is regular approval, not AA.

Dr. Spriggs replied that it comes down to the quality-of-life issue; a formal quality-of-life assessment should be conducted, if possible. The more formal the QoL assessment, the more comfortable he would be with a PFS endpoint.

Dr. Rose stated that if the patient is platinum resistant, she is unlikely to achieve CR. Toxicity is an important consideration.

Dr. Bast said that one presumably would not need a controlled study. He asked: Is the goal to show no decrease in QoL over baseline or to show improvement in QoL? Dr. Rose responded that a German trial showed improved PFS and OS and is often cited as almost a no-treatment control arm.

Dr. Spriggs replied that the problem with the framing of Dr. Bast's question is that the natural history of ovarian cancer involves a decrement in QoL. Most patients who are withdrawn from a study because of disease progression are probably going to have reduced QoL compared with baseline.

Dr. Basen-Engquist said that patient-reported outcomes (PROs) in oncology have focused on toxicity. Toxicity criteria are rated by clinicians, but quality-of-life assessments often look for negative as well as positive outcomes of treatment. When trying to prolong survival for a few months, one wants quality survival, not severely compromised survival.

Dr. Ozols said that in recurrent disease, platinum-sensitive patients will have symptoms when they have progressed. PFS will be a beneficial measure in that case. Progression is usually easy to determine in platinum-sensitive patients. Progression should be an endpoint for approval.

Dr. Nerenstone said that the PFS issue is very difficult when talking about second-line therapy. Those patients can be asymptomatic for a long time without treatment. They differ from refractory, sick patients. For refractory patients, RR and clinical improvement are important. They are very different patient populations.

Dr. Pazdur, summarizing for the group, said that given the right population, QoL, and the adverse event profile of a drug, PFS itself constitutes a clinical benefit in the maintenance-therapy setting.

Second-Line and Subsequent Therapy Setting

Dr. Pazdur described the AA process and reiterated the requirements for the granting of AA. A new drug can be granted AA on the basis of a surrogate endpoint that is "reasonably likely to predict clinical benefit." The new drug must provide a benefit over available therapy, which may consist of approved drugs or therapy well recognized by the oncology community. Many trials aimed at AA are conducted in the setting of refractory disease for which there is no available therapy and use single-arm trials and RRs. Trials for regular approval must be RCTs.

2. Could RR with adequate duration of response in a single-arm study support AA in a second-line, platinum-refractory setting?

Dr. Pazdur asked panel members to comment on whether RR would be reasonably likely to predict clinical benefit. Also, would there have to be no available therapy for this patient population?

Dr. Ozols replied that in practice, patients now receive multiple lines of treatment. In a defined group of patients with platinum-resistant disease, a 10% RR would indicate potential clinical benefit and could be considered for AA. Toxicity is a key factor.

Dr. Rose said that TTP is also important. Patients with stable disease do quite well and they need to be incorporated into this process. Dr. Pazdur concurred but emphasized that analysis of TTP and stable disease would have to take place in a randomized setting, not a single-arm trial. Dr. Freedman added that he would not want to look at RR alone in the second-line setting because refractory patients have a short duration of response; 2 months might not be significant because of ongoing disease symptoms.

Dr. Pazdur noted that if one were looking only at RR in a single-arm trial, one could discard a potentially active drug. Dr. Rose said that he and Dr. Spriggs had demonstrated that fact during the meeting: Although RRs for the drugs they compared were similar, there were significant differences in PFS and OS. Dr. Eisenhauer said that presumably one would need to have a high RR to obtain AA. Nonrandomized data are likely to indicate that the drug may have an impact on other endpoints; historically, the ovarian cancer drugs that produce $\geq 15\%$ RR have in subsequent trials shown that they promote PFS and OS. That is different from saying that a drug with a low RR might affect those endpoints, but such a drug would not be eligible for AA and would have to be approved with an RCT. Dr. Pazdur replied that one advantage of identifying RR in a refractory population is that doing so could identify drugs with a unique mechanism of action.

Dr. Bast noted that one would have to define refractory quite precisely. A 10% RR is “pushing it.” It would help if there were a way to identify that 10%.

Dr. Brady said that some people may be willing to extrapolate from a first-line setting to a second-line setting. Torri et al (1992) reported a meta-analysis that was conducted to assess whether RR predicted OS and concluded that RR is predictive of OS in the first-line setting. Accepting this study as evidence, however, requires some extrapolation because no similar analysis has been done in the second- or third-line setting.

Dr. Spriggs said that platinum-resistant patients are a heterogeneous group. Through patient selection, one can significantly alter the result of a trial. Dr. Pazdur replied that when companies want to conduct a single-arm trial, they have to narrow the patient population.

Dr. Rose noted that if using PFS in second-line settings would offer the advantage of not having to distinguish whether the drug were cytotoxic or cytostatic. If the patient does not progress, she does not progress, by whatever mechanism. One of the benefits of focusing on PFS is that the RR variability is very close—statistically identical.

Dr. Pazdur noted that to adequately evaluate PFS, one really needs an RCT. He reviewed the benefits of RCTs. Dr. Rose asked why one could not use PFS in a single-arm trial to identify which agents one might bring forward to an RCT. Dr. Pazdur said that was a different issue from whether FDA should approve a drug.

Dr. Vergote asked why RR would be so much better than PFS. Dr. Pazdur replied that if a trial generates a 10% RR, that effect is attributable to the drug, not to the natural history of the disease. A reduction in tumor size is attributable to the drug. When evaluating improvement in a time-to-event endpoint, however, it is hard to factor out the natural history of the disease. Evaluating TTP in a single-arm trial presents the possibility of approving a placebo.

Dr. Buyse asked whether FDA included in single-arm trials randomized Phase II trials in which there are multiple arms but no intention to compare them. Dr. Pazdur replied that the agency was referring to ordinary single-arm trials.

Dr. Rose asked whether, given an average 6-month survival of 30% for 11 or 12 different regimens, the ability to show a 40% or 50% benefit over that outcome would be one way to demonstrate that a new drug is better than previous drugs and eligible for AA pending the outcome of RCTs looking at PFS and OS.

Dr. Brady noted that in the Phase II setting the goal is to screen agents to identify those in which it is worthwhile to invest further effort; agents that pass that threshold can be studied further. Dr. Pazdur noted that screening a drug is different from bringing it to FDA for approval. In the regulatory setting, the trial endpoint must be reasonably likely to predict clinical benefit.

Dr. Freedman asked about the impact on accrual if patients realize that 90% of them will see no benefit and may face substantial toxicity risks. Dr. Pazdur replied that many drugs had an RR of 12% to 15%. In a single-arm trial, that's all one can say. One of the benefits of randomized controlled trials is that they can demonstrate a survival advantage.

Dr. Pazdur said there are many ways to obtain AA other than single-arm trials; he described an approach used in the HIV/AIDS research community. He noted that sponsors like single-arm trials because they carry little risk. Dr. Weiss noted that one risk is that a drug with a low RR could still have a positive effect on PFS and OS.

3. Could prolongation of TTP in a randomized study be sufficient for AA in second-line setting? Or regulatory approval?
4. What is the role of CA-125 in clinical trials intended for licensure in second-line and beyond – setting in ovarian cancer?

These questions had been addressed in the earlier discussion and the panel did not discuss them further.

Audience Questions and Comments

A woman in the audience who did not identify herself took issue with the use of the term “salvage therapy.” Cancer survivors cringe when they hear this term used, she said. She suggested that thinking of the treatment of patients with refractory disease as a salvage operation might influence the way clinicians think about procedures and treatment. This way of thinking devalues a person's life, she said. In this context, PFS might not be as important to clinicians as it is to patients.

Patient-Reported Outcomes

Patient-Reported Outcomes: Endpoints for Ovarian Cancer

Lari Wenzel, PhD, Center for Health Policy & Research, University of California—Irvine, said she was heartened to hear the amount of discussion related to PROs. The goals of her and Dr. Basen-Engquist's presentation were to illustrate significant PRO endpoints in ovarian cancer trials, identify PRO measures that are ready for incorporation into registration trials, and highlight PRO issues for further study. PROs are important in ovarian cancer research because ovarian cancer treatments should be evaluated for their ability to improve patient functioning and

reduce symptoms. Treatment-related side effects also must be assessed. Symptoms and function are best measured by asking patients directly.

GOG-172 compared IP and IV paclitaxel/cisplatin regimens. The IP regimen used higher and more-frequent dosing than the IV regimen. Toxicities were greater on the IP arm and fewer patients on the IP arm were able to complete the scheduled six cycles of therapy. However, a statistically significant improvement in PFS and OS was found for patients in the IP arm. The 65.6-month median survival on the IP arm is the longest median survival reported to date in an RCT in advanced ovarian cancer. The 2005 consensus is that the toxicities, inconvenience, and cost of IP therapy are justified by the improved survival seen with this treatment. New, targeted therapies are likely to be more effective in patients who have an excellent response to chemotherapy. While researchers work to improve the tolerability and toxicities of IP therapy, it remains the most effective means of treating ovarian cancer today.

The quality-of-life differences between IV and IP study arms were assessed using several scales:

- FACT-O (FACT-G: 27 items; ovarian subscale: 12 items)
- FACT-Trial Outcome Index (physical well-being [PWB], functional well-being [FWB], ovarian subscale)
- FACT-GOG/NTX (neurotoxicity): 11 items
- FACT-GOG/AD (abdominal discomfort): 4 items

QoL was significantly worse in the IP group before Cycle 4 and 3 to 6 weeks after treatment. No significant quality-of-life differences were found at 1 year except for Neurotoxicity. Neurotoxicity was significantly worse in the IP arm 3 to 6 weeks after completing chemotherapy ($P=0.0004$) and remained significantly worse in that arm 1 year later.

A self-assessment tool enables patients to score their experience with platinum/paclitaxel-related peripheral neuropathy. This 11-item scale assesses sensory, motor, and hearing dysfunction for a full range of sensory and functional concerns. Four items assess sensory neuropathy for efficient separation of groups in relation to chemo-induced neurotoxicity. Four sensory items accounted for 80% of treatment differences and 63% of changes in neurotoxicity scores. The 11-item scale has good internal reliability, construct validity, criterion validity, sensitivity to treatment differences, and responsiveness to treatment cycles. The 4-item scale is an efficient way to differentiate groups, but it misses motor or functional problems; it requires further validation.

GOG-172 also looked at patient-reported abdominal discomfort (AD), which was worse in the IP arm prior to Cycle 4. A 4-item subscale was used. Among 205 women on the IP therapy arm, 138 completed the AD subscale prior to Cycle 4. Both treatments improved AD from baseline to pre-Cycle 4. The difference in improvement between study arms favored IV therapy by a margin of 0.9 units.

The FACT/GOG-AD subscale is a valid and reliable instrument to measure AD. The AD subscale is responsive to change over time and is a useful tool to document the short- and long-term effects of AD on QoL. The AD subscale is worthy of use in future studies.

Compared with patients who received conventional-dose IV therapy, patients who received higher-dose IP therapy experienced more quality-of-life disruption, more abdominal discomfort during active treatment, and more neurotoxicity. However, they lived significantly longer than patients randomized to the IV study arm. From baseline to 12 months after treatment, QoL improved in both groups. Neurotoxicity was worse over time in both groups, but was worse in IP

patients than in IV patients. AD improved in both groups from pre-randomization to pre-fourth cycle.

Implications

Health-related quality of life (HRQL) PROs are useful in interpreting treatment implications. Global HRQL and symptom-specific indices add critical information to IV–IP comparisons. Continued quality-of-life evaluation is critical to weigh the considerable treatment benefits and toxicities and assist in establishing guidelines and safety standards to buffer untoward treatment effects.

Karen Basen-Engquist, PhD, MD Anderson Cancer Center, noted that PROs had not been used to date in any approvals of ovarian cancer drugs. PROs are most likely to be used for patients with recurrent or refractory disease who have more symptoms. Well-validated HRQL questionnaires are available, but FDA has been reluctant to include these broad concepts in labeling.

No validated symptom index exists for ovarian cancer. Issues in developing such an index include relevant symptom targets (single index combining multiple symptoms vs. multiple measures; disease symptoms vs. treatment side effects), reliability and validity, and defining meaningful change.

Dr. Basen-Engquist presented a diagram illustrating FDA’s guidance on PRO measures.

In recurrent ovarian cancer, no single cardinal symptom is expected to improve with treatment. Measures need to target multiple symptoms. When possible, symptoms of disease (which are expected to improve) should be measured separately from treatment side effects (which are expected to worsen). Ideally, the questionnaire should measure only disease-related symptoms, but this is difficult because one cannot always distinguish symptoms from treatment side effects.

See et al (2004), in a small pilot study, interviewed 50 women with platinum-resistant recurrent ovarian cancer. The women were about to receive either chemotherapy or hormonal therapy. They completed the FACT-O; a symptom checklist with symptoms from Memorial Symptom Assessment Scale; and the European Organization for Research and Treatment of Cancer (EORTC) Core Quality of Life Questionnaire, ovarian cancer module (QLQ-OV28). For the next stage of index development, the study participants were asked to identify the most frequent, severe, and distressing symptoms. The researchers added symptoms from patients below the 25th percentile in QoL and evaluated overlap between items.

Dr. Basen-Engquist presented a list of the symptoms identified by the patients (See et al, 2004) and a list of symptoms identified by health care providers (Cella et al, 2003). The most important symptom targets identified by patients also were identified by the providers.

Issues involved in developing a symptom index include whether to create one composite index (can “symptoms” be considered a single domain?) or multiple measures of individual symptoms (e.g., fatigue, pain, GI symptoms, psychological symptoms). Measures exist for fatigue, pain, and psychological symptoms, but not for GI symptoms. Finally, how should mood be dealt with? Does mood disturbance occur as a result of having cancer or can the tumor and treatment exert a direct effect on mood? Does effect on mood ever belong in the labeling for a drug?

The next steps are to further examine reliability and validity issues, responsiveness, and meaningful differences. A minimum important difference can be established using either anchor-based (e.g., anchoring to another PRO or to clinical change) or statistical methods (e.g., using the standard error of measurement). Statistical criteria are easy to apply but do not relate to patient experience. Dr. Basen-Engquist presented data from research that had attempted to anchor patient symptoms to other outcomes: FACT-O scores by extent of return to usual activities, change in ovarian cancer patients on second- or third-line chemotherapy (Doyle et al. 2001), and a correlation of findings on different indexes with changes in hemoglobin (Cella et al. 2002). Comparing changes in fact-based measures to statistical measures yields similar results.

A new direction in PROs is real-time assessment using handheld computers and other devices. Advantages of this approach include better recall, ecological validity, ability to identify short-term pattern in PROs (e.g., diurnal patterns), and better understanding of within-subject variability. A small feasibility study for recording fatigue asked participants to use handheld computers to complete an assessment six times per day; patients who used the computers had high rates of completing assessments. The data collected reveal patterns of fatigue and other symptoms over the chemotherapy cycle.

Another innovation is a computer algorithm that uses the patient's previous response to select the next question; it chooses the question that will provide the maximum information. For example, if depression is being assessed, the first question might ask the patient to rate how often she is sad; if she answers "rarely," the next question will not be about suicide but will be designed to gauge low-level vs. major depression. The approach results in briefer, more precise measures.

The Patient-Reported Outcomes Measurement Information System Initiative is an evidence-based conceptual framework that includes common patient endpoints; it has a large and well-tested repository of questions to measure most common and important symptoms and functional concepts.

PRO issues for further study include using outcomes to assist in documentation of symptoms. What symptoms are most important to patients? How are PROs related to clinical outcome?

Laurie B. Burke, RPh, MPH, Director, Study Endpoints and Label Development Team (SEALD), Office of New Drugs (OND), CDER, noted that until now much of the discussion had focused on TTP and on the measurement of decrements in QoL rather than on mitigation of symptoms. This approach is not unlike noninferiority study designs. Because researchers and clinicians cannot tolerate the possibility of missing meaningful progression of symptoms, much remains to be done in terms of instrument development. It is important to catch decrements in patients (from disease progression or side effects) who may be regressing. Ms. Burke asked if the presenters had thought about that setting.

Dr. Basen-Engquist noted that this problem complicates the issue of looking at minimally important changes. One cannot always assume that one will see symptom improvements, but one might be preventing a symptom from getting worse. A measure has to be equally sensitive to catching both improvements and decrements. Most quality-of-life scales have more "ceiling" effects than "basement" effects.

Ms. Burke replied that when FDA has reviewed quality-of-life instruments, it has observed that developers often do not think about picking up important symptoms of progression. Dr. Basen-Engquist had presented much preliminary data that could help to define exactly what needs to be

incorporated into these instruments. One has to characterize the burden of the disease in terms of patients' symptoms and ensure a sufficiently broad patient population to adequately characterize what patients are feeling and how their functioning has been affected. After that information has been collected, it would be a good idea to move forward with developing a composite index.

Dr. Wenzel said that in a GOG-152 study in which she participated (Dr. Rose was the principal investigator), they did not publish data on decrements per se but could predict survival based on quality-of-life scores. The lowest quartile of patients had the worst outcomes. They had progressing decrements in QoL due to disease.

Ms. Burke said it is important to distinguish between correlations and meaningful outcomes in the context of clinical trials. FDA is reluctant to use some general HRQL instruments because the agency is unable to apply the regulatory requirement that measures must be well developed and reliable to measure a definable concept. Their ability to predict survival is not adequate. Multidimensional concepts are difficult for FDA to interpret, especially if component domains are not adequately defined. Both component domains and items within those domains must be meaningful.

Dr. Basen-Engquist replied that the questionnaires being discussed do conform with generally accepted scientific principles for questionnaire development. It was not clear to her why the measure itself was considered to be lacking.

Ms. Burke said that the content validity aspect is most problematic for FDA. How do the items support the delineation of the concept that is being measured by the scores that are eventually defined as an endpoint and as having meaningful clinical benefit? Questions for the panel are: What is the value of PROs? What concept would be most useful?

Dr. Spriggs said that he was "completely at a loss" as to what Dr. Burke was talking about and asked for some concrete examples of things that FDA considers to be valid. Ms. Burke replied that Dr. Spriggs' question was a good one; it is why FDA wrote a draft guidance. The question is: How do we know that QoL has been adequately captured? It is much easier for the agency to look at things like abdominal discomfort that can be adequately captured.

Dr. Eisenhower said that QoL becomes most important in the setting of relapsed disease. The panel had been struggling with the meaning of prolonging TTP by 2 or 3 months. Such a prolongation might be meaningful for the patient if it meant reduced or delayed disease symptoms. In patients who present with disease-related symptoms, the discussion might be: "We hope this treatment will improve your symptoms for a time, might shrink your tumor." There is a reason to treat the patient for symptom management and control. In patients with small-volume disease; the discussion might be: "We are hoping to prolong the time until the disease begins to cause symptoms."

For example, if one focuses on the symptoms that are problematic for patients and defines what a meaningful improvement of that symptom is—many of the scales are categorical—one could describe the proportion of symptomatic patients getting treatment A versus treatment B who had an improvement of a minimum amount for a minimum amount of time. If the improvement is 80% on one treatment arm and 20% on the other, most people would be happy to receive the treatment that results in an 80% improvement in symptoms.

Dr. Rose noted that these measures are relatively new compared with the conventional practice of measuring tumor changes, but they have validation. He asked whether QoL could be used to predict outcome and, if so, whether this has been done. If one could treat patients in a second-line setting and see that the patients whose disease regresses by classic criteria also have improvement in their quality-of-life scores, it would be a validation of quality-of-life tools and would justify their use.

Dr. Wenzel emphasized that all measurement is an imperfect science. Although there is no existing gold standard for relating PRO and quality-of-life data to clinical measures or tumor size, considerable compelling work in ovarian cancer trials has made it possible to learn from existing data and existing validated measures, thus advancing the field. Those concerned with QoL are asking for an even playing field in terms of opportunities to advance those measures.

Dr. Basen-Engquist said there are advantages and disadvantages to symptom measures vs. broad quality-of-life measures. Symptom measures are useful in the regulatory context: Drug X decreased pain/fatigue/gastrointestinal symptoms. But patients are not symptoms. Symptom measures are missing the effects of disease on overall QoL, such as the impact on family.

Dr. Bast noted the opportunity to use patients' own experience with their previous course of chemotherapy as a baseline and to individualize what is important to the patient in terms of symptoms and life opportunities.

Ms. Burke said that patients and clinicians will have different interpretations of a statement that Drug X improves QoL. The goal is to define what outcomes are important to patients. All symptom and function improvement is important to a patient's QoL. When trying to determine the impact of treatment on a patient, one should be specific. FDA had seen many unsuccessful attempts to add QoL to disease progression and other clinical variables. The agency is advising sponsors to be more focused and specific as to the goals of measurement in order to better detect treatment benefit.

Dr. Wenzel said that perhaps the most conservative approach would be to start with a symptom index or symptom-specific measurement. Dr. Basen-Engquist noted that her group puts procedures into place to ensure that patients complete the assessments before they hear the results of tests.

Dr. Pazdur said medical oncology has done a "miserable job" of symptom assessment. Most therapies are allegedly given for palliation of disease, but it is not known how well they do this. The field of radiation oncology has done a better job; in that field specific information is available to correlate rads given with symptom improvement. This area needs further development.

Submissions of QoL data to FDA have major deficiencies, Dr. Pazdur added. In many cases, more than 50% of the data are missing. In therapeutic areas that rely on PROs, such as in psychiatric diseases, trials are blinded, which is not the case in oncology. If one looks at drugs that have symptom statements in the labeling, they are for diseases that have a cardinal symptom (e.g., dysphagia in esophageal cancer). The patient population has to be well defined. QoL endpoints should have the same validity as survival endpoints. The statistical validity of these instruments and the clinical implications for marketing are important. QoL measurement cannot be approached as an "add-on" to a clinical trial. Sponsors must take QoL measurement seriously

if they want to make QoL part of the process of drug evaluation. FDA is happy to work with sponsors who want to do this.

Biomarker and Endpoint Research Priorities

Edward L. Trimble, MD, MPH, Head, Gynecologic Cancer Therapeutics & Quality of Cancer Care Therapeutics, National Cancer Institute (NCI), described the composition of the GCIG and noted that it meets every 6 months. The GCIG has excelled at setting standards and promoting collaboration. He listed the group members' many closed and open studies along with several planned studies.

Potential biomarker and endpoint research priorities include correlation between PFS and OS in completed first-line studies; correlation between RECIST and CA-125 as markers of progression in completed first-line studies; and correlation between RECIST and GCIG CA-125 in Phase II trials of novel agents. Other potential priorities include (1) evaluation of radiologic intermediate endpoints as markers of refractory disease, persistent disease, or recurrent disease and those endpoints' correlation with PFS and OS and (2) evaluation of symptom benefit for ovarian-cancer-related symptoms in second-line trials.

Dr. Pazdur asked panel members whether they had other suggestions for potential trials.

Dr. Arbuck said that because of the difficulties associated with the use of PFS in various settings, researchers should work on defining time points and appropriate expectations.

Dr. Rose noted, with regard to the QoL/symptom-scale issue, that it is important to see how reliable and important those measures will be for patients. QoL, using whatever scale the experts select, should be measured with the same frequency as any other important measure. Investigators need to look at symptom scales regularly to see how reliably they can predict patients' disease status. They may turn out to be more important than CT scans.

Dr. Freedman said that new drugs such as, for example, trabectedin need to be carefully examined in relation to CA-125 modulation because they can modulate the inflammatory component. Macrophage function is the most common inflammatory component associated with ovarian cancer in the peritoneum. What is the relationship of the inflammatory component to CA-125? With the new generation of drugs, there is an opportunity to see the contribution of the inflammatory component to biomarker change. The issue of discordance between CA-125 and objective response needs to be addressed.

Dr. Buyse asked Dr. Arbuck if she had any experience with multiple time points (e.g., 6- as well as 12-month follow-up). Dr. Pazdur said FDA would like to test this area to see what is found if conventionally timed progression is measured using a log rank and then various endpoints are assessed along the way. FDA was recently asked about a fixed-endpoint TTP (in studies of a different disease). The agency was interested, but there is no regulatory history in this area.

Dr. Eisenhauer asked whether it was possible to review previous data sets to test some hypotheses and start to answer some of those questions. Dr. Trimble indicated that NCI funding might be available for such a study.

Dr. Weiss noted that Dr. Trimble had mentioned fluorine-18 fluorodeoxyglucose (FDG) positron emission tomography (PET) in his list of topics; there is much interest in this imaging modality

in other disease settings, where its use is further advanced. She asked what information is available on the potential uses of that type of assessment tool. Dr. Trimble replied that clinically, FDG-PET can help to localize disease, especially for patients with recurrence. FDA has had some discussions as to whether the agency should consider an FDG-PET scan after four to six courses of chemotherapy as a marker of someone who has persistent disease versus no disease or to help identify patients who are not responding to chemotherapy.

Audience Questions and Comments

Sherry Salway Black, Ovarian Cancer National Alliance (OCNA), referred to Paula Kim's statement in the morning session. Although survival is the ideal endpoint, OCNA supports the use of PFS, given what is known about the devastating nature of ovarian cancer, and is interested in working with FDA to support the use of PFS as an endpoint.

Edwin Rock, MD, Division of Drug Oncology Products and a member of the CDER Patient-Reported Outcome Working Group, addressing Dr. Burke, said it was not clear what the QoL questionnaires were missing. Generic scales of QLQ30 were compared with FACT in four populations with cancer. In each population, statistically significant differences were observed in subscales. If those two generic instruments, which are purported to measure the same domains of QoL, do not give reliable and consistent results, then what are they measuring?

Addressing Dr. Wenzel, Dr. Rock noted that on the neurotoxicity subscale, four sensory items accounted for 80% of treatment differences. If that is the case, he asked, why is that not a sufficient measure of how the patient is feeling with respect to neurotoxicity?

Concerning the Austrian study, Dr. Basen-Engquist acknowledged that QoL measures are far from perfect. However, in other studies there is remarkable consistency with what is expected. They show what patients have been telling clinicians for years, and many studies predict outcomes in expected ways. They are picking up meaningful differences. It would be a mistake to throw them out completely.

Dr. Wenzel said the way she had described the study involving the neurotoxicity subscale was "giving the short story." The instrument was an 11-item index that was launched in GOG-172. It also was used in a study of advanced endometrial cancer. A validation paper now in review for a 4-item subscale of the 11-item scale is based on the endometrial cancer study. The 4-item scale will detect sensory neurotoxicity. This measure was not used because it had not been validated.

Workshop Summary/Conclusions

Stacy R. Nerenstone, MD, Oncology Assoc, PC, Gray Cancer Center, noted that her summary was a joint effort with Dr. Freedman. She identified the following key points that had been made during the workshop:

- Adjuvant intraperitoneal chemotherapy is now the standard of care for first-line therapy of stage III ovarian cancer; that standard must be the comparative basis for future trials. Increase in PFS is a valid endpoint, but studies still need to be sized to measure OS; OS remains an important endpoint, even if crossover interferes.
- Maintenance chemotherapy is similar to first-line therapy; as Dr. Spriggs pointed out, the separation is somewhat artificial. In maintenance therapy, OS is the significant endpoint, especially because treatment entails more toxicity. Improvement in PFS might be acceptable

if the treatment is nontoxic, especially for AA, because it allows conclusions to be reached more quickly. Biological agents might be able to be evaluated on the basis of PFS if they are not going to affect subsequent cytotoxic treatment. Such studies should still be sized to measure OS, however. The use of PFS as an endpoint means that the timing of treatment evaluations must be carefully monitored to avoid bias.

- Second- and third-line therapy, especially in platinum-sensitive patients, is thought to be similar to first-line therapy. An issue that has not been discussed is what increase in PFS would be meaningful. Is a 1-, 2-, or 3-month increase in PFS sufficient for approval? Dr. Nerenstone opined that it was not sufficient.
- In treating refractory disease, RR might be acceptable for AA if it is high enough—perhaps if there are a certain number of CRs and the duration of the response is adequate. Additional supportive studies would be required.
- With regard to using CA-125 as an endpoint, the GCIG usage as part of a composite (radiographic, clinical, and CA-125) endpoint is acceptable. If PFS is being used as an endpoint, it must be remembered that a decrease in PFS will be seen because disease progression will be identified at an earlier stage when measured by means of CA-125.
- The same is true for second- and third-line treatment, with caveats. One must be careful of the timing of the evaluation in RCTs to make sure that both arms are balanced. Bias can be a problem, especially if the study is not blinded.
- The strength of PFS as a surrogate for OS is especially clear in first-line treatment but more validation is needed, particularly in second- and third-line treatment, because there are fewer data and in refractory disease the data are less compelling. There are exceptions: GOG-111 found no correlation. This subject requires further evaluation.
- PFS seems to correlate with OS, especially when there is a large effect on PFS. A smaller increase in PFS may not correlate with OS. Crossover confounds the measurement of OS. As treatment improves, it may be more difficult to increase OS significantly.
- FDA needs to ensure that it is approving truly active drugs. Many issues are outstanding, particularly with regard to the use of CA-125. Further prospective data are needed to validate CA-125 as a correlate for progression. Other concerns are that there is no reference for CA-125; some patients are CA-125–negative; timing of measurements could be important; and new biologics may affect CA-125 expression. Protocols must define TTP precisely and prospectively.
- PROs could be important for drug approvals. Instruments need to be developed and content validity is needed to support approval. The outcome must be important from the patient perspective. Many questions remain about whether current instruments are capturing the most important data. Missing data remains a huge problem in QoL analyses.

Dr. Bast thanked Drs. Nerenstone and Freedman for their summary. He noted that the panel had discussed using CA-125 as an endpoint for response in Phase II trials. He asked whether it would be reasonable to permit patients with either CA-125 elevation or progression according to standard RECIST criteria to enter Phase II trials and to abandon drugs from further consideration

in the absence of a CA-125 response, thus increasing accrual. The group concurred that this would be a reasonable approach.

Dr. Ozols said the only issue that needed to be decided was what level of PFS improvement is clinically useful. In refractory disease, median survival after relapse is 2 years. There is no plateau in that group. Everyone wishes PFS were longer than about 3 months, but in the context of how lethal the disease is after relapse, that is the level of clinical benefit that can be expected.

Dr. Pazdur thanked the participants and the audience, and he thanked ASCO—particularly the work of ASCO staff—for its support of the meeting. He expressed appreciation to FDA staff Linda Burbank and Dianne Spillman.

Dr. Pazdur adjourned the meeting at 4:15 p.m.

References

- Cella D, Paul D, Yount S, et al. What are the most important symptom targets when treating advanced cancer? A survey of providers in the National Comprehensive Cancer Network (NCCN). *Cancer Invest* 2003; 21(4): 526-35.
- Cella D, Eton DT, Lai JS, et al. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) Anemia and Fatigue scales. *J Pain Symptom Manage* 2002; 24(6): 547-561.
- Cesano A, Lane SR, Poulin R, et al. Stabilization of disease as a useful predictor of survival following second-line chemotherapy in small cell lung cancer and ovarian cancer patients. *Int J Oncol*. 1999; 15(6): 1233-1238.
- Crawford SM, Peace J. Does the nadir CA125 concentration predict a long-term outcome after chemotherapy for carcinoma of the ovary? *Ann Oncol* 2005; 16(1): 47-50.
- Doyle C, Crump M, Pintillie M, Oza AM. Does palliative chemotherapy palliate? Evaluation of expectations, outcomes, and costs in women receiving chemotherapy for advanced ovarian cancer. *J Clin Oncol* 2001; 19(5): 1266-1274.
- Markman M, Liu PY, Wilczynski S, et al. Phase III randomized trial of 12 versus 3 months of maintenance paclitaxel in patients with advanced ovarian cancer after complete response to platinum and paclitaxel-based chemotherapy: A Southwest Oncology Group and Gynecologic Oncology Group trial. *J Clin Oncol* 2003; 21(13): 2460-2465.
- Markman M, Liu PY, Rothenberg ML, et al. Pretreatment CA-125 and risk of relapse in advanced ovarian cancer. *J Clin Oncol* 2006; 24(9): 1454-1458.
- Markman M, Bookman MA. Second-line treatment of ovarian cancer. *Oncologist*. 2000;5(1):26-35.
- Piccart MJ, Floquet A, Scarfone G, et al. Intraperitoneal cisplatin versus no further treatment: 8-year results of EORTC 55875, a randomized phase III study in ovarian cancer patients with a pathologically complete remission after platinum-based intravenous chemotherapy. *Int J Gynecol Cancer* 2003; 13 Suppl 2: 196-203.
- Rustin GJ, Nelstrop AE, Bentzen SM, et al. Selection of active drugs for ovarian cancer based on CA-125 and standard response rates in Phase II trials. *J Clin Oncol* 2000; 18: 1733-1739.
- Sabbatini P, Spriggs DR. Consolidation for ovarian cancer in remission. *J Clin Oncol* 2006; 24 (4): 537-539.
- See HT, Walker PW, Palmer JL, et al. Palliative care index for ovarian cancer. Abstract #544, presented at International Gynecologic Cancer Society biennial scientific meeting, Edinburgh, Scotland, 2004.
- Sorbe B. Consolidation treatment of advanced ovarian carcinoma with radiotherapy after induction chemotherapy. *Int J Gynecol Cancer* 2003; 13 Suppl 2: 192-195.
- Sorbe B. Consolidation treatment of advanced (FIGO stage III) ovarian carcinoma in complete surgical remission after induction chemotherapy: a randomized, controlled, clinical trial comparing whole abdominal radiotherapy, chemotherapy, and no further treatment. *Int J Gynecol Cancer* 2003. 13(3): 278-286.
- Torri V, Simon R, Russek-Cohen E, et al. Statistical model to determine the relationship of response and survival in patients with advanced ovarian cancer treated with chemotherapy. *J Natl Cancer Inst*. 1992 Mar 18;84(6):407-14.
- Verheijen RH, Massuger LF, Benigno BB, et al. Phase III trial of intraperitoneal therapy with yttrium-90-labeled HMFG1 murine monoclonal antibody in patients with epithelial ovarian cancer after a surgically defined complete remission. *J Clin Oncol* 2006; 24(4): 571-578.