Gefitinib for Recurrent Non–Small-Cell Lung Cancer: All Things Are Not Created Equal

Suzanne E. Dahlberg, Robert J. Gray, and Bruce E. Johnson, Dana-Farber Cancer Institute, Boston, MA

Two studies reported in this issue of *Journal of Clinical Oncology* have results in which the distributions of outcomes for patients with non–small-cell lung cancer treated with gefitinib were similar to those for standard chemotherapy treatment control arms using docetaxel and vinorelbine. The important question is to what extent these two studies provide evidence that the efficacy of gefitinib is similar to that of the chemotherapy-treated control arms. The report by Crino et al1 discusses a randomized phase II trial designed to determine superiority of gefitinib compared with vinorelbine for previously untreated elderly advanced non–small-cell lung cancer. They conclude in their discussion, “Although the present study was not designed to show equivalence between the two drugs, there was no statistical difference” between the two treatments. The report by Maruyama et al2 presents results from a randomized phase III trial that compared gefitinib with docetaxel in Japanese patients with relapsed non–small-cell lung cancer. This trial was originally designed as a superiority trial, but relevant data regarding the possible equivalence of gefitinib and docetaxel led the investigators to modify the design to target a noninferiority hypothesis.

These studies prompt questions about the concordance of the clinical and statistical interpretation of the data, namely whether there is enough evidence from these two studies for clinicians to conclude clinical equivalence in outcomes of patients treated with the two agents. This issue is directly related to the design of each trial and whether additional studies are needed to formally address the statistical issue of noninferiority of gefitinib to standard of care, given that a soon to be published subsequent trial was designed to formally address noninferiority of gefitinib versus docetaxel in the same patient population as the Maruyama et al study. Although noninferiority designs are potentially important for regulatory approval, one wonders whether they are an efficient use of clinical resources or necessary for making clinical decisions.

Unlike superiority trials, which aim to demonstrate that a novel treatment outperforms a standard regimen by a particular margin of superiority, noninferiority studies are appropriate for determining whether a novel agent is about as effective as the current standard, especially when the new treatment is less toxic. These studies specify a small noninferiority margin, which is the amount the new treatment can be less effective than the standard and still be considered an acceptable treatment, and test whether the difference between the standard and the new treatment is less than the specified noninferiority margin. In considering the magnitude of an acceptable noninferiority margin, the possible small reduction in efficacy must be weighed against the possible improvements in the safety profile, impact on the patient’s quality of life, cost, and other factors. This margin must also be less than the known magnitude of benefit of the standard treatment control arm because, otherwise, a conclusion of noninferiority would not be sufficient to establish that the new treatment was better than best supportive care (BSC).

Consideration of the precision of estimates is crucial in the interpretation of noninferiority studies, and thus, the results can perhaps best be explained through CIs.3-5 If the two-sided CI around the point estimate of the primary end point includes the noninferiority margin, then the study does not establish noninferiority because the results do not rule out an unacceptably large decrease in efficacy. If the CI excludes the noninferiority margin (in favor of better outcomes with the new treatment), then it can be concluded that the new treatment is within an acceptable margin of the standard. Furthermore, superiority of the new treatment may be concluded if the CI excludes equal efficacy in favor of better outcomes on the new treatment. This approach differs from that used in the more common superiority trial, in which the null hypothesis of treatment equality is formally tested against an alternative hypothesis of superiority of the new treatment.

With regard to the report by Crino et al,1 it is important to highlight that their conclusion of “no statistical difference” in treatments does not consider the precision of their estimates. Failure to reject the null hypothesis in a superiority trial is not the same as accepting the null hypothesis because there is always a range of possible null values that would not be rejected. The gefitinib versus vinorelbine overall survival (OS) hazard ratio was 0.98 (95% CI, 0.66 to 1.47). A noninferiority margin was not prespecified, but in this case, an unacceptably large noninferiority margin corresponding to a hazard ratio of 1.47, the upper bound of the reported CI, would be needed for the results to support a conclusion of noninferiority. This ratio corresponds to a 47% higher event rate with gefitinib. Because noninferiority margins are seldom prespecified in superiority trials, claims of noninferiority after failure to reject the null hypothesis in favor of superiority are problematic.5-7 The results of this study also do not provide strong indirect evidence of benefit from gefitinib relative to BSC because a previous study comparing vinorelbine with BSC8 in a similar population had an OS hazard ratio for vinorelbine versus BSC of 0.65 (95% CI, 0.45 to 0.93). Although the point estimate of 0.65 suggests that a ratio less than 1.53 (1:0.65) in the Crino et al1 study might be sufficient...
to establish that gefitinib is superior to BSC, the prior study only established with confidence that the ratio was less than 0.93, corresponding to a 7% reduction in the event rate, and the variabil-
ity in both estimates as well as possible concerns about comparability of the populations from the two studies need to be taken into account.9 We do not believe that the authors intended to conclude noninferiority of gefitinib compared with vinorelbine; a test for noninferiority was not reported. However, it is important to stress that their comment that there was “no statistical difference” should not imply statistical or clinical equivalence of the regimens.

In the trial by Maruyama et al,2 the original superiority design was modified to specify a noninferiority margin for the OS hazard ratio (gefitinib vs docetaxel) of 1.25 (ie, the event rate was allowed to be up to 25% higher on gefitinib), but the sample size was not adjusted accordingly, and the authors state that their power to demonstrate noninferiority was less than 50%. For a study with similar follow-up to have 80% power for a noninferiority margin of 1.25, at least 1,000 patients would need to be randomly assigned. Therefore, this study was severely underpowered for establishing noninferiority, and its only difference from a superiority trial that failed to reject the null hypothesis of equivalence is prior specification of the noninferiority margin. The OS hazard ratio for gefitinib versus docetaxel is 1.12, and the 95.24% CI is 0.89 to 1.40. The authors correctly state that the results do not establish noninferiority of gefitinib but also note that gefitinib was not significantly worse, suggesting that there could be “clinical equivalence.” As the upper bound of the reported CI shows, though, the results are consistent with up to a 40% higher event rate with gefitinib, which would be an unacceptably large difference from standard treatment after considering the magnitude of the benefit for docetaxel. The US Food and Drug Administration label for docetaxel reports an OS hazard ratio for docetaxel versus BSC of 0.56 (95% CI, 0.35 to 0.88) based on the study of Shepherd et al.10 Considering the variability in both estimates, it seems that the results of Maruyama et al2 are also not sufficient by themselves to establish a survival benefit of gefitinib relative to BSC.

Although the conclusions of the two studies may seem different, because Maruyama et al2 formally reject noninferiority, the observed effects in the two studies are similar and the CI on the OS hazard ratio from Maruyama et al,2 who studied 489 patients, lies entirely within that from Crino` et al,1 who studied 196 patients, despite the differences in populations and in active control arms. The results of these studies are also consistent with the overall results of the Iressa Survival Evaluation in Lung Cancer study,11 which had an OS hazard ratio for gefitinib versus placebo of 0.89 (95% CI, 0.77 to 1.02) in 1,692 pretreated, unsel ected patients, although this study did show a larger effect in 375 never-smokers (hazard ratio, 0.67; 95% CI, 0.49 to 0.92) and in 342 patients of Asian origin (hazard ratio, 0.66; 95% CI, 0.48 to 0.91). The results of the INTEREST trial presented by Douillard et al12 at the Twelfth World Conference on Lung Cancer seem to rigorously establish equivalence of gefitinib compared with docetaxel in 1,466 patients with previously treated NSCLC using a noninferiority margin corresponding to a hazard ratio of 1.154.

Noninferiority designs are appropriate for determining whether a new treatment is nearly as beneficial as a previously established active standard therapy, but they generally require large sample sizes and substantially more resources than superiority studies. This leads to a question of whether the large investment needed to establish noninferiority and the rigor needed for regulatory approval are worthwhile for the overall goal of improving outcomes for cancer patients and advancing the field of oncology. An example of how lung cancer physicians have assumed clinical equivalence without statistical evidence of noninferiority is demonstrated in the design of the Eastern Cooperative Oncology Group–led intergroup E1505 study, a large ongoing adjuvant chemotherapy trial studying the effect of chemotherapy with or without bevacizumab. This trial allows for the choice of three different but commonly administered cisplatin–based chemotherapy regimens under the assumption that the choice of doublet (cisplatin plus vinorelbine, docetaxel, or gemcitabine) is unlikely to impact the outcome of the trial even though no previous trial has demonstrated noninferiority of these regimens to each other. Regardless, one must use caution when stating that regimens are not significantly different or equivalent in terms of efficacy, and studies should be planned so that the clinical impact of showing noninferiority is meaningful to our patients and, therefore, worthy of the allocation of additional patients and clinical research resources.

AUTHORS’ DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST
Although all authors completed the disclosure declaration, the following author(s) indicated a financial or other interest that is relevant to the subject matter under consideration in this article. Certain relationships marked with a “U” are those for which no compensation was received; those relationships marked with a “C” were compensated. For a detailed description of the disclosure categories, or for more information about ASCO’s conflict of interest policy, please refer to the Author Disclosure Declaration and the Disclosures of Potential Conflicts of Interest section in Information for Contributors.

Employment or Leadership Position: None Consultant or Advisory Role: Bruce E. Johnson, Genzyme (C) Stock Ownership: None Honoraria: None Research Funding: Robert J. Gray, Genentech Expert Testimony: None Other Remuneration: Bruce E. Johnson, Genzyme

AUTHOR CONTRIBUTIONS
Conception and design: Suzanne E. Dahlberg, Robert J. Gray, Bruce E. Johnson
Data analysis and interpretation: Suzanne E. Dahlberg, Robert J. Gray, Bruce E. Johnson
Manuscript writing: Suzanne E. Dahlberg, Robert J. Gray, Bruce E. Johnson
Final approval of manuscript: Suzanne E. Dahlberg, Robert J. Gray, Bruce E. Johnson

REFERENCES
3. Makuch R, Simon R: Sample size requirements for evaluating a conserva-