The hazards of randomized phase II trials

Randomized phase III trials in non-small-cell lung cancer (NSCLC)—particularly in the frontline metastatic setting—have been a high-risk low-yield strategy. In past two decades, many such trials targeting the broad population of advanced NSCLC patients without the benefit of molecular selection have been conducted. In most of these studies, the experimental design has been that of a platinum-based doublet therapy with or without a new or investigational agent. Unfortunately, the vast majority of these large phase III studies has been disappointingly negative, with no advantage seen to the addition of a novel drug to a standard cytotoxic chemotherapy backbone.

In fact, only two phase III trials employing this paradigm have reported an overall survival benefit for the experimental arm. The first of these was Eastern Cooperative Group Trial 4599, which randomized NSCLC patients with nonsquamous tumor histology to carboplatin/paclitaxel with or without the angiogenesis inhibitor bevacizumab. The second was a phase III study (also called the FLEX trial) of cisplatin/vinorelbine and cetuximab, an antibody against the epidermal growth factor receptor (EGFR), in patients with advanced NSCLC whose tumors had at least one cancer cell positive for EGFR by immunohistochemistry. In the FLEX trial, the survival benefit for the triplet arm was quite modest and the experimental treatment was associated with enhanced toxic effects, such as a higher rate of febrile neutropenia. It is notable that of these two trials, only the former has led to a US Food and Drug Administration approval for the use of the experimental agent (i.e. bevacizumab) in the commercial space.

It is also interesting to note that for every 20 or so phase III trials conducted in this clinical NSCLC setting, only 1 or 2 positive trials emerged, mirroring that magical ‘$P = 0.05$’ false-positive rate that clinical trialists have long used to define statistical significance. It is abundantly clear that many more novel agents fail to improve treatment efficacy when combined with frontline platinum-based chemotherapy: these agents include gefitinib, erlotinib, sorafenib, figitumumab, vadimezan, and celecoxib among many others.

In this issue of the journal, yet another promising agent tested in the phase III metastatic NSCLC setting joins this long unenviable list of failures. Here, Manegold et al. (2) report the results of a large phase III study ($n = 839$) that tested the addition of PF-3512676, a novel Toll-like receptor 9-activating oligodeoxynucleotide, to the gemcitabine/cisplatin backbone. In the final analysis, median overall survival and median progression-free survival times were found to be not significantly different between the two groups. The unkindest cut of all was that there was the higher rate of toxic effects,
particularly grade ≥3 hematologic adverse events, injection site reactions, and flu-like symptoms, in the PF-3512676 + chemotherapy arm. Similar negative results were reported in a similarly designed parallel phase III study of carboplatin/paclitaxel with or without PF-3512676 in advanced NSCLC [1].

These phase III trials were enthusiastically developed in large part because of the encouraging results of the randomized phase II trial that preceded it. That predecessor trial was a phase II study that randomized 117 advanced NSCLC patients to platinum–taxane chemotherapy with or without PF-3512676 [2]. The primary end point was tumor response rate. The experimental arm demonstrated a significantly higher response rate (38% versus 19%, \( P = 0.043 \)); interestingly, overall survival also appeared to be improved for the PF-3512676 arm (12.3 versus 6.8 months), as was 1-year survival rate (50% versus 33%), although the survival differences were not statistically significant \( (P = 0.188) \). Toxic effects were also higher in the experimental arm. It is important to note that the doublet (cisplatin/vinorelbine) tested in the Manegold phase III trial was not specifically evaluated in a preceding phase II trial. Nevertheless, the experience with PF-3512676 drug development magnifies the many recent disappointments of negative phase III trials despite a ‘positive’ predecessor phase II trial.

For many drug developers and clinical trialists, the randomized phase II trial design has helped to provide some level of comfort. The resource demands, high expense, and complexity of a phase III trial require some margin of confidence around the efficacy results of a preceding phase II study. This is somewhat borne out of historical experience: the likelihood of a positive phase II combination chemotherapy trial resulting in a subsequently positive phase III trial within 5 years was reported to be just 0.038 [3]. The randomized phase II design was supposed to rescue us from this rather dismal yield. Unfortunately, for all its promises and advantages, we must recognize its many limitations and pitfalls [4, 5].

In question is the value of a randomized phase II (or single-arm phase II) to screen for potential activity of new treatment regimens to determine whether there is evidence to warrant further study. A negative phase III study should not necessarily be viewed as evidence of a failure in the screening phase. Just as diagnostic/screening tests are not perfectly predictive of true disease status, phase II studies are not perfectly predictive of the results in a follow-on phase III study. Similar to diagnostic tools, the properties of a phase II design can be defined in terms of two components: (i) the likelihood that a positive phase II is followed by a positive phase III (positive predictive value) and (ii) the likelihood that a negative phase II would be followed by a negative phase III (negative predictive value). In the ideal phase II, study designs would have both a high positive predictive value and high negative predictive value, as for any screening tool. But in reality, these values will never be 100% and maximizing one usually occurs at the expense of the other.

The move away from single-arm to randomized phase II studies was motivated by concerns that single-arm phase II studies are inherently fraught with bias and it is this bias that has resulted in so many false leads. The root of the concern is that without a concurrent control group, the studied population may differ from historical controls or outcome ascertainment may differ from historical data resulting in a biased evaluation of the treatment effect.

When addressing the question regarding why so many positive phase II studies have resulted in a negative phase III, the primary concern is the positive predictive value of the phase II study. All components of a design can have an impact on the positive predictive value of a phase II study. If the study population of the phase II and the phase III are different, then differences seen in the phase II population may not reflect differences in the phase III. If the differences in the outcome used for the phase II do not reflect differences in the phase III, then this too can result in a negative phase III. If the target effect size and outcome used in the phase II do not translate into a reasonable effect on the phase III outcome, then this may result in a negative phase III.

But what are the ‘costs’ of a randomized phase II? As phase II trials are intended to screen for new treatments, they should be completed in a relatively short time frame. Adding a comparator arm to a study typically requires about four times the number of patients if the design properties are held the same. Alternatively, a compromise has to be made on design parameters and this in turn has an impact on the positive predictive value of the study. For example, most single-arm phase II trials employ a 5% false-positive rate, whereas randomized phase II trials often use a 10% type I error rate that translates into a doubling of the percentage of negative trials which will be defined as promising. In addition, randomized designs may be more reliant on the use of intermediate end points (such as response) than single-arm studies for a couple reasons. Firstly, randomized phase II studies may need to be designed with intermediate end points in order to minimize the sample size. Also, intermediate end points may be used because of concerns regarding the ethics of conducting a randomized phase III study following a statistically significant difference in the phase III end point at the phase II.

After basic considerations of whether or not the trial population is a representative sample of the diseased population, at the root of reliability of trial results is sample size. Sample size is determined by the false-positive (type I error) rate, false-negative (type II error) rate, effect size, and the event rate in the control arm. The number of patients required decreases with increasing error rates and effect sizes. The number of patients or more accurately the number of events needed in a trial determines the precision of estimates and accuracy of conclusions. Small sample sizes can result in extreme values due to chance alone. While randomization serves to balance prognostic factors between groups on average, the success of randomization is also a function of sample size. In smaller trials, the role of stratification to help balance prognostic factors becomes even more crucial.

The sample size for the randomized phase II that preceded the study of Manegold et al. (2) was only 117 patients, with 78 patients assigned to the experimental arm and 39 patients assigned to the control arm (2 : 1 randomization). The primary analyses were based on 74 patients randomized to the experimental arm and 37 patients randomized to the control arm using what they call a ‘modified intent-to-treat’ analysis. Of note, randomization was not stratified on any factors and
the sample size in the control arm was quite small. The median overall survival time in the control arm was only 6.8 months, which is significantly lower than what had been reported in the literature for a platinum–taxane doublet and was significantly lower than the observed median survival time in the subsequent phase III study (10.7 months). In addition, the response rate of 19% in the control arm in the phase II was substantially lower than the rate in the phase III of 31%.

The randomized phase II that preceded the randomized phase III is a perfect example of it and is easy to focus on one part of the picture as valid information while discounting other components that may not justify the conclusions as strongly. On the surface, the phase II results appear very promising: the response rate in the experimental arm was significantly higher than the rate in the control arm with a \( P \) value equal to 0.043 and the median survival time, while not statistically different was numerically larger than the control arm. But again, the median survival in the control arm was significantly lower than expected. We are quick to take a significant \( P \) value as the truth. But \( P \) values are also variable in small sample sizes. Moreover, the migration of the estimates of median survival from the phase II to the phase III demonstrates that the interpretation with more data, a numerical difference will become a statistically significant difference is misguided.

Randomized phase II studies suffer from many of the same limitations as single-arm studies. They are a screening tool and it should not be expected that a positive phase II necessarily means that the following phase III study will be positive. The predictive value of a phase II study is a function of the quality and design of the study and not whether it involved a randomized comparison or not. Moreover, non-small-cell cancer—particularly adenocarcinoma histology—is now known to be composed of many heterogeneous biologic subsets, some of which are characterized by molecular ‘drivers’ such as EGFR-activating mutations or EML4-ALK translocation [6]. Randomized phase II or III trials that fail to account for these driver mutations—some of which are successfully ‘druggable’—may lead to imbalances in the distribution of these molecular subsets, consequently affecting the ultimate trial results and their interpretation.

P. N. Lara\textsuperscript{1,2,*} & M. W. Redman\textsuperscript{1,2}

\textsuperscript{1}Division of Hematology and Oncology, Department of Internal Medicine, University of California Davis School of Medicine, Sacramento, \textsuperscript{2}Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, USA

(*E-mail: primo.lara@ucdmc.ucdavis.edu)

disclosure

PNL is a consultant and received honoraria from Pfizer.

references