Statistical controversies in clinical research: end points other than overall survival are vital for regulatory approval of anticancer agents

E. D. Saad1,2* & M. Buyse1,3

1International Drug Development Institute, Louvain-la-Neuve, Belgium; 2Department of Medical Oncology, Dendrix Research, Sao Paulo, Brazil; 3Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Hasselt University, Diepenbeek, Belgium

Received 1 September 2015; revised 27 October 2015; accepted 2 November 2015

There is an ongoing debate about the relative merits of overall survival (OS) and other metrics that can be used as primary end points in cancer clinical trials. Although survival time is arguably the most objective metric for assessing the efficacy of anticancer treatment, OS as a clinical-trial end point needs to be conceptually distinguished from increased survival time as a goal desired by patients, clinicians and public-health policy makers. OS presents several drawbacks as a primary end point that threatens to hamper further drug development, including the increase in the number of patients and the much longer follow-up required in a clinical trial. In many settings of first-line therapy for metastatic disease, median OS is currently two to four times longer than median progression-free survival. As a result, the analysis of OS may be increasingly confounded by the effect of salvage therapies used after disease progression. In this review, we use straightforward statistical reasoning and examples from the oncology literature to argue that OS should no longer be the primary end point of most future phase III cancer clinical trials that aim at assessing the efficacy of novel therapies in the setting of metastatic disease.

Key words: disease-free survival, end point determination, neoplasms, survival, survival analysis

introduction

There is an ongoing debate about the appropriateness of using overall survival (OS) as the primary end point in cancer clinical trials. Proponents of OS as primary end point argue that increased survival is the most relevant outcome for patients with incurable tumors, and for a long time OS has been considered the gold standard end point in oncology [1]. Survival time is arguably the most objective metric for assessing the efficacy of anticancer treatment, and—along with quality of life—the most relevant measure of patient benefit. However, as we have argued elsewhere, OS as a clinical-trial end point needs to be conceptually distinguished from increased survival time as a goal desired by patients, clinicians and public-health policy makers [2]. The fact that increasing the length of survival is the ultimate objective of anticancer treatment does not automatically make OS the preferred end point to assess treatment efficacy in metastatic disease, for reasons that will be outlined and illustrated below. By using straightforward statistical reasoning and examples from the oncology literature, we will argue that OS should no longer be the primary end point of most future cancer clinical trials. Although the focus of our reasoning will be pivotal phase III trials that aim at regulatory approval, the same arguments can be used for cooperative group and academic trials, but not necessarily to more pragmatic trials that assess different therapeutic strategies along a continuum of therapy (e.g. predetermined interventions for sequential lines). For the latter types of trials, OS may continue to be the preferred primary end point.

statistical drawbacks of using OS as primary end point

smaller number of events

Primary end points serve the dual function of providing the basis for sample-size calculation and for deciding whether a study is statistically positive at its completion. Despite its objectivity, several difficulties arise when OS is used as primary end point. First, for any given number of patients enrolled in a clinical trial, there are by definition fewer deaths than cases of disease progression at any time point (except at the very end of the follow-up, if all patients have eventually died); as a result, the analysis of OS has lower power than the analysis of the other relevant time-to-event end points, such as disease-free survival in the adjuvant setting and progression-free survival (PFS) in advanced disease.
longer median time to event

The second drawback of using OS as primary end point stems from the simple fact that median OS is longer than median PFS. In recent years, median OS time has been two to four times longer than median PFS in first-line, phase III trials of common solid tumors, such as breast cancer, nonsmall-cell lung cancer (NSCLC), and colorectal cancer [3–5]. As a result, a much longer follow-up time is required if OS is used as primary end point than if PFS is used. For example, 120 events are required to demonstrate the superiority of an experimental treatment, assuming the same hazard ratio (HR) of 0.60 for either end point, a two-sided type I error rate of 5%, and a power of 80%. For a fixed accrual rate of 10 patients per month, 120 PFS events will be observed after ~24 months if median PFS is ~6 months. In contrast, 120 deaths will be observed after ~43 months if median OS is ~24 months.

Table 1 illustrates this point using the concrete example of pivotal phase III trials that aimed to demonstrate a PFS benefit for tyrosine kinase inhibitors (TKIs) currently approved in the United States for treatment of patients with advanced NSCLC [6–8] (for afatinib, OS data were obtained from [9]). For afatinib and erlotinib, only approval for first line is illustrated, whereas for crizotinib, the approval was not based on treatment line. Although gefitinib has recently received approval in the United States for the treatment of epidermal growth factor receptor (EGFR)-mutated, metastatic NSCLC, such approval was based on a nonrandomized trial. Table 1 shows the statistical power of the trial to detect the same HR on OS as the HR assumed for PFS, at the time of the PFS analysis, and 2 years later. Clearly, these three trials would have had good power to detect the same HR for OS than that observed for PFS, albeit after 2 additional years of follow-up. In the next section, we turn to situations in which the HR for OS is closer to 1.00 than the HR for PFS.

smaller relative effect

While there are some instances of identical [10] or very close [11] HRs for PFS and OS in the oncology literature, there is a general trend for the HR for OS to be closer to 1.00 than the HR for PFS (i.e., smaller effects on OS than on PFS). We discuss some reasons why this may be the case in the next section. For now, let us assume identical differences in medians (i.e., the same absolute magnitude of effect on the time scale). For example, a 4-month difference in median PFS between experimental and control arm translates into a HR of 0.60 if median PFS is 6 months in the control arm; assuming no effect from subsequent treatments, the same hypothetical 4-month difference could carry over to OS and translate into an OS HR of 0.86, assuming a median OS of 24 months in the control arm. In this hypothetical example, we have seen above that 120 events are required to detect a HR of 0.60 for PFS, whereas 1380 deaths are required to detect a HR of 0.86 for OS. Table 1 shows the statistical power of the three TKI trials in NSCLC to detect the same difference in median OS as that observed in median PFS. Clearly, none of these trials has reasonable power to detect such differences, even after 2 additional years of follow-up. As it turns out, Table 1 shows that reality was even more daunting, since the observed median OS was almost identical in the control and experimental arms of all three TKI trials. We now turn to potential reasons for smaller treatment effects on OS than on PFS.

Table 1. Statistical power to demonstrate a difference in overall survival (OS) in recent positive trials that had PFS as primary end point, using the same treatment effect for OS as for PFS (same assumed hazard ratio (HR) or same observed median difference)
competing risks and cross-overs

Patient survival times are influenced by competing risks of death and the effect of cross-over therapy. Competing risks are especially evident in the adjuvant setting, as potentially cured patients continue to die of other causes. This seems particularly evident in early-stage disease and among elderly patients [12, 13]. As cancer treatments become more effective and less toxic, patients entered in clinical trials will increasingly be drawn from older age groups with higher risks of dying from competing causes unrelated to their cancer. The effect of post-trial cross-over is evident in advanced disease, especially when the trial drug or agents from the same or a closely related class are commercially available. Probably, the most cogent example of the effect of cross-over is provided by first-generation TKIs of the EGFR, such as erlotinib and gefitinib. These agents have radically changed the treatment of patients with EGFR-mutated, advanced NSCLC [14]. Both agents promote dramatic improvements in response rates and PFS among such patients, in comparison with chemotherapy, but gain in OS has not been demonstrated in phase III trials, possibly as a result of cross-comparison with chemotherapy, but gain in OS has not been demonstrated in phase III trials, possibly as a result of cross-over to a TKI in 64% to 95% of patients [6, 15, 16]. A similar situation will likely be encountered with another very active TKI, crizotinib. This agent targets anaplastic lymphoma kinase (ALK) abnormalities and has been shown to improve response rates and PFS in ALK-positive NSCLC, in comparison with chemotherapy, but a significant gain in OS is unlikely to be found in these trials, given the fact that 64% to 70% of patients initially treated with chemotherapy crossed over to crizotinib after disease progression [7, 17]. A striking feature about the effect of cross-over is that it is more likely to influence OS results when treatment effects on response rate or PFS are larger. This creates an apparent paradox: the more efficacious a new agent, the less likely it is to impact on OS, since the perceived improved efficacy will lead to a higher percentage of patients crossing over to the new agent or one from the same class.

role of OS, current, and future

use of OS in regulatory approvals

Table 2 shows the end points used by the United States Food and Drug Administration to approve oncology drugs between 2005 and 2007 [19]. PFS and TTP have been used more often than OS, but response rate has been the end point more often used in this series of regular or accelerated approvals. Interestingly, novel end points, which have been used for specific conditions, were reduction in hepatic iron content, asparagine depletions, red-blood-cell-transfusion independence, stabilization of hemoglobin levels, rate of venous thromboembolism, and reduction in tissue injury from chemotherapy extravasation. More recently, the proportion of patients who experienced a reduction in spleen volume of ≥35% has been used as primary end point for approval of ruxolitinib in myelofibrosis, further exemplifying the view that novel end points are acceptable from a regulatory viewpoint [20].

OS as primary end point

There are situations for which OS arguably constitutes the most relevant primary end point. In diseases such as advanced pancreatic cancer, differences in median PFS appear to translate into differences in median OS after first-line chemotherapy, probably because of the lack of effective salvage therapy [21, 22]. Likewise, it has been possible to demonstrate gain in OS in disease settings for which cross-over was not allowed and the same trial therapy was not commercially available [23–25]. However, greater availability of active drugs in these tumors will likely make gain in OS a much rarer finding in the future. This is currently a concern in many disease settings, with metastatic castration-resistant prostate cancer being an example [26].

In settings for which effective post-trial therapy is available either in the form of cross-over or use of treatment from the same class after disease progression, there is considerable uncertainty as to the reliability of OS as an indicator of treatment efficacy. A significant gain in OS has been relatively infrequent in these settings [6, 7, 15, 16, 27–29]. This finding should not be taken to indicate that such therapies are ineffective, but rather that OS does not reliably capture the efficacy of these treatments. Negating this conclusion would be akin to withholding EGFR TKIs from patients with NSCLC and activating EGFR mutations in the clinic just because clinical trials have failed to show significant improvements in OS.

To be sure, there are also instances of significant gains in OS despite the existence of salvage therapy [10] or the absence of preceding gain in PFS [30–32]. The extent to which gain in survival in these latter instances is indeed due to the trial drug, to the unreliability of PFS, to post-trial therapy, or to the play of chance, is presently unknown. Finally, some classes of agents such as immunotherapies may arguably have delayed effects that manifest themselves much later in the course of the disease, possibly after the tumor had progressed. A proper evaluation of these drugs would require either using OS as the primary end point, or validating early biomarkers of immune or tumor response that would reliably predict long-term effects on clinical end points. We return to this important issue later.

OS as essential secondary end point

The proposal that OS should no longer be chosen as the primary end point in most phase III trials does not mean that OS should not be used as a secondary end point. Indeed, in some instances, OS may be a very important indicator of

Table 2. End points used for regular or accelerated approvals by the United States Food and Drug Administration between July 2005 and December 2007 [19]

<table>
<thead>
<tr>
<th>Primary end point</th>
<th>Number (%) of approvals (N = 53)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall survival</td>
<td>10 (19%)</td>
</tr>
<tr>
<td>Disease-free survival*</td>
<td>5 (9%)</td>
</tr>
<tr>
<td>Progression-free survival and overall survivalb</td>
<td>3 (6%)</td>
</tr>
<tr>
<td>Progression-free survival or time to progression</td>
<td>12 (23%)</td>
</tr>
<tr>
<td>Overall response rate</td>
<td>16 (30%)</td>
</tr>
<tr>
<td>Other, novel end point (see text for description)</td>
<td>7 (13%)</td>
</tr>
</tbody>
</table>

*aEnd point used for all approvals in the adjuvant setting.
*bCo-primary end points.
treatment safety or even serve as a quality benchmark in a clinical trial. At least in theory, analyzing OS may capture long-term toxicity that may be missed by PFS if most patients have documented disease progression before death, which is the case in the vast majority of patients participating in contemporary clinical trials. This would be even more important in trials where time to progression (TTP) is used as primary end point in its proper sense (i.e. when deaths without prior documentation of disease progression are censored for the analysis of TTP). Finally, the analysis of OS may point to issues related to treatment strategies that might confound the cause-and-effect relationship between trial therapy and long-term outcomes, a problem recently illustrated by the nonsignificant decrease in OS (despite a significant improvement in PFS) when tivozanib was compared with sorafenib in metastatic renal-cell carcinoma [28]. In this case, discordant results between PFS and OS were attributed to the fact that patients randomized to tivozanib were allowed access to second-line therapy at time of progression, but such therapy was rarely available in the Eastern European countries where many of the patients were enrolled. On the other hand, patients in the sorafenib arm were allowed to cross-over to tivozanib at progression. In addition to blaming a design that in fact compared sequential therapy with monotherapy, other plausible explanations exist for those findings, including that sorafenib may be more effective than tivozanib in terms of OS, and that tivozanib has delayed toxicity; as pointed out recently, however, ‘we will never know for sure’ [33].

are there satisfactory alternative primary end points?

The flip side of the controversy about the appropriateness of using OS as primary end point is that end points based on the tumor itself may be more reliable indicators of the true treatment effect. Among such alternative end points, which are few in number, PFS has been most often defended as a candidate to substitute OS [34, 35]. Thus, an open debate currently pervades the oncology literature regarding the advantages and disadvantages of either OS or PFS as the most appropriate primary end point [2, 36–42]. Both OS and PFS have well-known pros and cons. PFS is attractive because it is available earlier than OS, is not affected by postprogression treatments, and is less vulnerable to competing causes of death than OS. However, PFS has important limitations: the definition of PFS, e.g. per the Response Evaluation Criteria in Solid Tumors (RECIST), is an arbitrary unidimensional increase in tumor size which is at best a rough indicator of the tumor behavior, and which may or may not have biological relevance. Functional imaging might help refine the definition of tumor progression, but no universally accepted standard criteria have yet been proposed for such techniques to be incorporated in the definitions of end points in clinical trials. As it is currently defined, ascertainment of disease progression is subject to considerable measurement error and potential bias, and both of these factors contribute to making PFS an end point that is far from ideal.

In addition to the methodological and biological limitations of PFS, many authors have argued that PFS does not capture the treatment benefit that is most desirable to patients and clinicians. Sometimes progressive disease is accompanied by symptoms that make it a patient-relevant treatment failure in and of itself, but such is not the case when progression is based solely on imaging without any impact on ‘how the patient feels, functions or survives’, which is often used to define a clinical end point (as opposed to a biomarker) [43]. However, the ongoing debate is at least in part due to the failure to distinguish the role of OS as primary end point in trials from its role as the most important therapeutic objective to pursue in incurable settings [2]. Likewise, the controversy may be contaminated by results of surrogacy analyses that have often failed to demonstrate that PFS is a valid surrogate end point for OS. From a statistical viewpoint, surrogacy entails the verification that a potential surrogate end point (e.g. PFS) displays strong associations with the final end point of interest (e.g. OS) both at the individual level (i.e. patients that have a longer PFS tend to have a longer OS) and at the trial level (i.e. treatments that improve PFS tend to improve OS) [44]. Over the past few years, PFS has been found to be a valid surrogate for OS in some tumor types and treatment settings [45, 46], but not in others [47]. One possible explanation for these discordant findings may be the fact that, in settings for which effective post-trial therapy is available, the trial-level association between PFS and OS is blunted [37, 38]. For example, PFS was considered a valid surrogate for OS in metastatic colorectal cancer treated with first-line chemotherapy at a time when targeted agents were not widely available [45]. Subsequent work in the same setting, now including more recent trials, has shown an attenuation of the association between PFS and OS at the trial level [48]. As a result, clinical settings for which surrogates for OS are most urgently needed—because OS is becoming long enough to preclude the feasibility of many trials—present the greatest challenges in terms of surrogate validation [47, 49]. We therefore believe the issue of surrogacy is a parallel one, and should not influence the decision of whether to use OS or PFS as the primary end point of a particular trial.

Given the limitations of both PFS and OS, the future of cancer evaluation probably rests on the discovery of biomarkers that not only capture the effects of treatment on the tumor, but are also able to predict the effects of treatment on long-term clinical end points. This is a tall order, as illustrated by the recent Food and Drug Administration-led meta-analysis of trials testing neoadjuvant treatments for patients with operable breast cancer. This meta-analysis showed pathological complete response (pCR) after neoadjuvant therapy to be a very potent predictor of OS in these patients, yet it failed to validate pCR as a reliable surrogate for OS, because in the set of trials analyzed, the treatment effects on pCR rates did not correlate with the treatment effects on OS [50]. Much work remains to be done to identify biomarker-based surrogates in clinical trials, but as we wait for these, imposing OS as the primary end point in clinical trials may do more harm than good and delay the approval of promising new drugs.

In addition to end points based on tumor morphology, there is an urgent need to develop biomarker-based end points such as pharmacodynamic end points, circulating tumor cells or DNA, functional imaging, and others that attempt to quantify the effect of treatment on salient biological features of the tumors. The next challenge will be the statistical validation of
such biomarker-based end points as surrogates for clinically relevant outcomes, such as survival and quality of life, that remain the ultimate goal of all cancer treatments [18, 51].

**Conclusion**

Although a clear path forward is not yet available, we believe that efficacy end points for clinical trials aiming at drug approval in the metastatic setting will shift from time-to-event end points such as PFS and OS and response rates towards metrics that truly capture the direct mechanism of action of treatments. Until pharmacodynamic biomarkers that are valid surrogates for clinically available outcomes are achievable, tentative recommendations can be made regarding the choice of the primary efficacy end point in the design of current cancer trials. In settings for which there is effective post-trial therapy or when cross-over within the trial is allowed, PFS could generally be used as the primary end point, especially if postprogression survival is expected to be long. When there is limited or no effective post-trial therapy and cross-over is not allowed, OS could be used as the primary end point, especially if postprogression survival is short relative to PFS. Unfortunately, there is little empirical evidence of what ratio of postprogression survival to PFS should be used when choosing an appropriate primary end point. In all cases, consideration should be given to the use of novel end points that may be appropriate in particular indications, as discussed above. Improved knowledge about tumor biology has already resulted in much faster new drug development. Hopefully, this knowledge will also suggest biomarker-based end points that can be validated as surrogates for the long-term clinical end points of primary interest to patients and oncologists.

**Disclosure**

EDS is an employee at the International Drug Development Institute, Louvain-la-Neuve, Belgium. MB is an employee and holds stock at the International Drug Development Institute, Louvain-la-Neuve, Belgium.

**References**


**Disclosure**

EDS is an employee at the International Drug Development Institute, Louvain-la-Neuve, Belgium. MB is an employee and holds stock at the International Drug Development Institute, Louvain-la-Neuve, Belgium.
34. Saad ED, Katz A, Hoff PM, Buyse M. Progression-free survival as surrogate and as true end point: insights from the breast and colorectal cancer literature. Ann Oncol 2010; 21: 7–12.
42. Venook AP, Tabernero J. Progression-free survival: helpful biomarker or clinically meaningless end point? J Clin Oncol 2015; 33: 4–6.