



Assessing Treatment Benefit in Immuno-oncology

Marc Buyse^{1,2} · Everardo D. Saad³ · Tomasz Burzykowski^{2,3} · Julien Péron^{4,5}

Received: 23 June 2019 / Revised: 16 December 2019 / Accepted: 31 January 2020 /

Published online: 14 February 2020

© International Chinese Statistical Association 2020

Abstract

Immuno-oncology is a buoyant field of research, with recently developed drugs showing unprecedented response rates and/or a hope for a meaningful prolongation of the overall survival of some patients. These promising clinical developments have also pointed to the need of adapting statistical methods to best describe and test for treatment effects in randomized clinical trials. We review adaptations to tumor response and progression criteria for immune therapies. Survival may be the endpoint of choice for clinical trials in some tumor types, and the search for surrogate endpoints is likely to continue to try and reduce the duration and size of clinical trials. In situations for which hazards are likely to be non-proportional, weighted logrank tests may be preferred as they have substantially more power to detect late separation of survival curves. Alternatively, there is currently much interest in accelerated failure time models, and in capturing treatment effect by the difference in restricted mean survival times between randomized groups. Finally, generalized pairwise comparisons offer much promise in the field of immuno-oncology, both to detect late emerging treatment effects and as a general approach to personalize treatment choices through a benefit/risk approach.

Keywords Immuno-oncology · Immune RECIST · Weighted logrank test · Accelerated failure time model · Restricted mean survival time · Generalized pairwise comparisons

✉ Marc Buyse
marc.buyse@iddi.com

¹ International Drug Development Institute, 775 Francisco Street, San Francisco, CA 94133, USA

² Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, Hasselt, Belgium

³ International Drug Development Institute, Louvain-la-Neuve, Belgium

⁴ Departments of Oncology and Biostatistics, ImmuCare, Hospices Civils de Lyon, Pierre-Benite, France

⁵ University of Lyon 1, CNRS UMR 5558, Biometry and Evolutive Biology Laboratory, Biostatistics-Health Team, Villeurbanne, France

1 Introduction

Immuno-oncology (IO), which embodies the confluence of tumor immunology and medical oncology, is a contemporary approach to cancer treatment using an old idea [1]. Immunotherapy (IT), the attempt to elicit the immune system to fight cancer, dates back at least 120 years, but until very recently, there had been little impact on clinical practice. IT comprises a variety of treatments that have as primary mechanism of action the generation of an immune response against cancer [2–4]. Such treatments include cytokines (e.g., interferons and interleukins), checkpoint inhibitors (CPIs) and other types of antibodies with immunological targets, genetically engineered T-cell therapies and other cell-based products, small molecules, oncolytic viruses, and different types of vaccines [2–5]. Some of these agents have led to unprecedented responses in clinical settings marked by resistance to conventional treatments, and improvements in overall survival (OS) have been very frequently observed in phase 3 trials of CPIs [2]. However, the dynamics of tumor responses, disease progression, and long-term gains of several IT agents, particularly CPIs, have called into question some of the conventional methods of assessing treatment benefit in oncology [2, 6–11]. In this article, we provide an overview of the conventional and novel statistical methods for assessing treatment benefit in IO clinical trials. Our focus is on advanced disease, the setting in which most of the current evidence on IT has been generated. We begin by considering the direct effect of IT on tumors, discuss the translation of these effects into patient benefit, and end by exploring measures of treatment effect that may capture such benefit.

2 Tumor Responses as the First Step Toward Benefit

2.1 Mechanistic Considerations

As a general rule, a measurable antitumor effect is a *sine qua non* of effective treatments in oncology. Unlike chemotherapy and targeted therapy, IT works indirectly, through the generation of an immune response against tumors. The effect of IT comprises a continuum of biological phenomena that involve both innate and adaptive immune mechanisms, as well as cellular and humoral immune responses [2, 12]. Of importance, tumor infiltration by cytotoxic T lymphocytes and other effector immune cells is one of the prerequisites for the antitumor activity of IT [13–15]. This activity is in turn countered by several immune suppression mechanisms acting in the tumor microenvironment [12, 13]. The dynamic interactions between the immune system and the tumor, and the varying nature of such interactions over time, can be described using the concept of immunoediting [15]. According to this view, there are three states of interaction between the immune system and the tumor: elimination, equilibrium, and escape. Interestingly, a parallel has been suggested between these three states and the clinical

observations of response, stable disease and disease progression after IT [2]. Thus, the knowledge about the mechanisms involved in response and resistance to IT can be used to some extent to explain the dynamics of tumor shrinkage and growth during treatment.

2.2 Patterns of Response

The experience to date suggests that tumor responses in most patients follow the usual pattern seen with other treatment modalities, with objective response rates to CPIs that vary between close to zero and slightly over 60% according to disease setting [16]. However, unusual response patterns have emerged with the use of IT, and the mechanistic aspects mentioned above may underlie some of these patterns. For example, the use of CPIs is associated with a response profile that is not adequately captured by conventional response assessment criteria, such as the Response Evaluation Criteria in Solid Tumors (RECIST) [2, 7, 17]. The three unusual patterns of response described to date are mixed responses, pseudoprogression, and hyperprogression [2, 7]. The heterogeneity of metastatic cancers, which also characterizes their immunological landscape, probably underlies cases in which a mixed clinical picture emerges after treatment, with some lesions shrinking while others remain stable or grow [2, 7, 18]. The frequency and prognostic significance of such mixed responses are still unclear. On the other hand, in 2% to 9% of patients treated with a CPI, an initial tumor growth is followed by *bona fide* responses, a phenomenon now termed pseudoprogression [7, 17, 19]. In some of these cases, lymphocytic infiltration of tumors is probably responsible for the initial increase in volume of a lesion destined to shrink, but a delayed action of IT is also postulated as an underlying mechanism in some cases [7, 10, 19]. In advanced melanoma, greater increase in CD8+ cells in serial tumor samples during therapy correlated with a greater tumor size decrease on imaging [20]. The role of these postulated mechanisms is corroborated by several studies showing that pseudoprogression is associated with favorable outcomes when compared with RECIST-defined progressions, especially in advanced melanoma [19, 21, 22]. Both mixed responses and the suspicion of pseudoprogression represent great challenges to patients and physicians, as a decision needs to be made about treatment continuation. This is not the case with the more recently described phenomenon of hyperprogression, whereby some patients display very early signs of unquestionable disease progression after treatment with IT [7, 23]. Although definitions have varied among studies, hyperprogression has been associated with unfavorable outcomes [7, 23–25]. Moreover, it has been postulated that hyperprogression may underlie the early detriment from the use of CPIs in some phase 3 trials [7]. Despite this hypothesis, a putative immunological mechanism for hyperprogression remains to be elucidated, and controversy still exists on whether the phenomenon is particular to IT or reflects the natural history of some tumors.

2.3 Response Criteria

Under RECIST [26] and its predecessor guideline, proposed by the World Health Organization (WHO) in 1979 [27], tumor growth beyond a certain magnitude or the

appearance of a new lesion indicated progressive disease (PD), synonymous with treatment failure in the chemotherapy era. Early in the development of CPIs, the unusual response patterns raised concerns about the adequacy of previous guidelines in this setting [10, 19]. Of particular concern was pseudoprogression, given the observation of prolonged periods of stable disease (SD) or event complete (CR) or partial response (PR) after an initial increase in tumor burden. These concerns led to an international collaboration of experts and the publication, in 2009, of a new set of guidelines for use with immunotherapy, which were developed based on radiographic images from patients with advanced melanoma treated with ipilimumab [19] and were later applied to another CPI used in this setting, pembrolizumab [22]. These immune-related response criteria (irRC), based on the WHO method of bidimensional measurement, introduced the concept of “total tumor burden” and the need to confirm PD. Since 2009, three additional sets of response criteria have been published [21, 28, 29]. The so-called immune-related RECIST combine some features of irRC (total tumor burden and the need to confirm PD) and of RECIST, the latter because only unidimensional measurements are used [28]. The RECIST group developed immune RECIST (iRECIST), which differs from previous guidelines in that (1) PD that is not confirmed leads to “resetting of the bar” for the assessment of progression, and (2) new lesions are not incorporated into the total tumor burden, but rather lead to a new set of lesions assessed in parallel to the original ones [29]. The more recently published immune-modified RECIST have been developed on the basis of imaging studies from patients with non-small-cell lung cancer and urothelial carcinoma treated with atezolizumab, yet another CPI, and is generally similar to iRECIST [21]. The application of these criteria is costly and time-consuming, especially in view of the fact that they increase the final overall response rate by 1% to 2% in many cases, with an additional 10% of patients overall who would have RECIST 1.1-defined PD being characterized as having SD [19, 21, 30, 31]. On the other hand, some retrospective studies have shown higher percentages of patients moving from RECIST 1.1-defined PD to SD or an objective response when treated beyond progression, especially in advanced melanoma and renal-cell carcinoma [32–34]. One important criticism of some of these results is the fact that IT-related response criteria have been developed in the context of clinical trials in which physicians could make a decision to continue IT in patients with an apparent clinical benefit despite evidence of progression. This subjective decision may have introduced bias due to the separation of patients with more aggressive disease from those with more indolent disease [11, 30].

Given the limitations of imaging assessment in IO, an interesting research avenue involves pathological assessment of responses in an attempt to correlate biopsy findings with those from radiographic assessment. In recent studies, pathological findings have shown promise as predictors of objective response, as well as of long-term benefit from IT, both in the neoadjuvant [35] and in the advanced setting [36].

3 Translation of Responses into Prolonged benefit

3.1 Response Duration and Its Assessment

Although objective responses are a desirable first step toward deriving favorable results from treatment, and in some cases the means to obtain improvements in symptoms, ultimately there is an expectation that responses will be durable and will bring long-term benefit. In fact, there is general empirical evidence for that, since anticancer agents receiving accelerated approval based on tumor responses often have their benefit confirmed later on [37]. On the other hand, prolonged disease stabilization can also be seen as an important benefit from IT [2, 3]. Moreover, long-term survivors may have had SD or even PD as their best responses to IT [38, 39], and in some patients responses have improved over time even without subsequent treatment, especially among those with melanoma [39, 40].

Prolonged responses appear to be more specific to IT than to other treatment types. Early in the development of cancer vaccines and CPIs, it became apparent that these agents were associated with responses lasting several weeks or months in settings for which this was not typically the case with chemotherapy [41]. Likewise, prolonged responses often occur when chimeric antigen receptor T cells are used in hematological malignancies [42]. Some IT agents have received a first approval based on responses in early-phase trials, and regulators have expressed interest in expanding our understanding of response-based metrics and their association with clinical benefit [43]. Talimogene laherparepvec, an oncolytic viral therapy, was approved after a phase 3 trial demonstrated improvements in durable response rate, defined as the percentage of patients with CR or PR maintained continuously for at least 6 months [44].

Given the above considerations, the assessment of response duration is a laudable goal toward better understanding the benefit from IT. Such an assessment is straightforward when made descriptively, but problematic when there is a comparative intent. The comparison of treatments in terms of response duration is likely biased because only responding patients are considered, with the groups under comparison being defined by a post-randomization feature. Interestingly, the treatment producing more responses will usually have responding patients of worse prognosis, and the bias may in fact go against the superior treatment [45]. Although modeling approaches have been proposed to avoid this analysis-by-responder bias [46], simpler procedures proposed recently may lead to increased use of analyses of response duration, at least in an exploratory fashion [47, 48].

The first of these procedures, due to Korn and colleagues, consists in generating more comparable patient subsets by removing responding patients with the least tumor shrinkage in the treatment group with more responders or by adding non-responding patients with the most tumor shrinkage to the group with fewer responders, in both cases maintaining similar proportions of responders in both groups [48]. Huang and colleagues have proposed a method that takes advantage of the additive properties of restricted mean survival times (RMSTs), which are discussed in more detail below [47]. The proposed method consists in

ascribing a response duration to each patient in a trial, thus avoiding the exclusion of non-responding patients from analysis. The method entails the construction of Kaplan–Meier curves (for each arm separately) for a composite endpoint defined as the time elapsed between treatment initiation and response, progression, or death, whichever comes first. Kaplan–Meier curves for each arm are also constructed for progression-free survival (PFS) in the usual manner. The RMST for this composite endpoint is then computed for each arm and subtracted from the RMST for the corresponding PFS curve, yielding the restricted mean duration of response for each treatment arm. As a result of this procedure, non-responding patients will have a response duration of zero, because the same event (of progression or death) will be used for these patients to indicate the occurrence of the composite endpoint and of PFS.

3.2 Quantifying the Association Between Responses and Long-Term Benefit

With chemotherapy and targeted therapy, there is often a strong association between objective responses and PFS and between the treatment effect on these endpoints [49, 50]. On the other hand, the association between responses and OS has been more modest [49–51]. Several authors have attempted to quantify the association between response to IT and long-term endpoints [16, 52, 53]. Unfortunately, none of these studies on IT used individual-patient data; nevertheless, a weak association was generally found between objective response rates and both PFS and OS, as well as between the treatment effects on response rates and these long-term endpoints. As a possible exception, a modest association was found between the treatment effects on response rate and on PFS in one study ($R^2 = 0.47$; 95% confidence interval 0.03–0.77) [53]. To our knowledge, no similar evidence has yet been generated for the duration of response as a potential surrogate for PFS or OS. The reason for the weak association between responses and PFS with IT is not clear. In addition to the limitations of analyzing aggregated data, these results may reflect the play of chance, real biological phenomena, and issues related to the assessment of responses and PFS. For example, the application of immune-related criteria to the assessment of PFS has only partially been explored [21], and phase 3 trials of CPIs have based such an assessment on RECIST 1.1 methods. Whether different associations could exist using response and PFS definitions of the immune-related criteria is a matter of speculation. The assessment of PFS is addressed explicitly by iRECIST and imRECIST, both of which specifying the need to confirm progressions [21, 29]. Thus, further work is needed to assess the relationship between rates and duration of responses and long-term outcomes in IO.

4 Assessing the Ultimate Benefit for Patients

4.1 The Revival of Overall Survival as Primary Endpoint

The prospect of curing metastatic cancer has never been better for patients [2], and unprecedented 5-year survival rates have been reported in some settings [38, 39, 54]. Prolongation of OS is a realistic goal in IO. In fact, OS is currently the preferred endpoint for phase 3 trials in IO, a lesson learned during the development of ipilimumab and confirmed in later trials [55, 56]. This is in contrast to chemotherapy and targeted therapy, settings in which a decade-long debate prevailed between proponents of OS and proponents of PFS as the primary endpoint in phase 3 trials. Given the limitations of OS, PFS eventually became the preferred primary endpoint in several settings in the era of chemotherapy and targeted therapy [57]. With these modalities, the effects of treatment coincide with its administration; however, IT behaves differently in that regard, given its putative delayed effects. Confirming the initial impression about a discordance between PFS and OS in IO [58], several phase 3 trials have shown gains in OS without an accompanying significant gain in PFS [59–64], an infrequent observation in the previous era. An initial increase in tumor volume from immune infiltration, delayed antitumor activity, or a sustained antitumor effect beyond progression have been postulated to explain that discordance [64]. Moreover, several meta-analyses based on published data have shown weak associations between PFS and OS in IO [16, 52, 53, 65]. Thus, in the remainder of this article we restrict the discussion to the assessment of OS in IO. The reader should note that we leave aside considerations related to predictive biomarkers, even though they bear implication in some of the design and analysis issues discussed. Moreover, the reader should consider that the development of IO will probably lead to increasing frequency of crossover to IT after disease progression, recapitulating the challenges observed with chemotherapy and targeted therapy.

4.2 Non-proportional Hazards of Survival

An early observation from comparative trials of CPIs has been the unusual behavior of Kaplan–Meier curves, especially with regard to the presence of delayed treatment effects on OS and of an apparent plateau in the tail of the curves. Later on, a third unusual phenomenon became apparent, albeit less frequently: the crossing of survival curves in some trials [60]. The mechanism of action of IT has been summoned as one of the potential explanations for delayed separation of OS curves, a phenomenon that is frequent [55, 58, 61, 64, 66, 67] but not universal [63, 68]. It is conceivable that an early detriment from IT, manifested as crossing of the curves a few months after randomization, also results from delayed effects, although hyperprogression may also play a role [7]. Likewise, it is conceivable that crossing curves reflect the existence of subpopulations with differential effects from treatment, as seen with targeted therapy [69]. Finally, the flattening of OS curves, which can be seen as evidence for a cure fraction [8], may

also indicate the natural history of the disease in patients with indolent tumors [39]. All these observations suggest that design and analysis models assuming the proportionality of hazards are less than optimal for trials in IO, especially when comparisons are made with treatments from other classes [10, 11]. Among other problems, the presence of non-proportional hazards may lead to loss of statistical power [8], incorrect conclusions from interim analyses [70, 71], and difficulties in understanding and communicating treatment benefit [72–74]. Although several solutions to these problems have been proposed in the literature, their uptake appears to have been low in terms of both design and analysis of phase 3 trials in IO [6, 11, 56, 70, 73, 75–77]. Table 1 displays the most frequently proposed solutions to deal with the issue of loss of statistical power and interpretation of treatment benefit. The different methods have advantages and disadvantages summarized in Table 1; all of them share a lack of regulatory precedent as compared with the decades of use of the logrank test. In the following, we briefly discuss methods that appear to have the greatest potential for assessing treatment benefit in a more meaningful way than with conventional methods based on the hazard ratio (HR) estimated from Cox models.

4.3 Weighted Logrank Tests

The proportional hazards assumption is arguably too strong in many practical situations. The violation of this assumption is frequent in oncology, and even more so in phase 3 trials in IO, where up to 50% present evidence of non-proportionality [78]. The omission of prognostic covariates from the proportional hazards model, many of which are often unknown, induces time dependence of the HR for coefficients in the model, making it difficult to distinguish the effect from a true time-dependent coefficient even in randomized trials; moreover, this bias is accentuated by increasing censoring [79].

The estimation and testing of treatment effects in the presence of non-proportional hazards has been a topic of research for a long time, but proportional hazards models have remained the standard approach in oncology because deviations from proportionality were uncommon and/or unknown in advance. With the advent of immunotherapies, the standard approach is increasingly being questioned, and weighted logrank tests have received renewed attention. Harrington and Fleming [80] proposed a two-parameter family of weighted logrank tests that can accommodate a large number of situations, in particular delayed treatment effects. Specifically, for I ordered survival times t_1, t_2, \dots, t_I , the weighted logrank test statistic is

$$Z = \frac{\sum_{i=1}^I w(t_i)(O_i - E_i)}{\sqrt{\sum_{i=1}^I w(t_i)^2 \text{var}(O_i - E_i)}}$$

where O_i and E_i represent, respectively, the observed and expected numbers of deaths at the i th event time, t_i , and $w(t_i)$ a weight at time t_i . The $G^{\rho,\gamma}$ family defines the weight function as

Table 1 Selected proposed statistical methods to deal with non-proportional hazards

Method	Advantages	Disadvantages
Weighted logrank test	Higher statistical power in the same nonparametric framework as the logrank test	Potential bias from weight selection; ethical concern from differential weighing of earlier and later events; loss of power if shape of curves is incorrectly specified
Non-proportional hazard parametric models	Robustness to omission of covariates; good description of data from long-tailed distributions; good compromise between exponential and nonparametric models	Lack of flexibility due to strong assumptions; unsuitable for the extreme cases of non-proportional hazards with initially overlapping or crossing survival functions
Cure-rate models	Correct model if some patients are truly cured	Cure remains elusive for patients with advanced disease; delays in disease progression seem more likely
Semiparametric accelerated failure time models	Robustness to omission of covariates; interpretation in terms of the mean survival time (preferable to median survival time); no parametric distributional assumptions	Unsuitable for extreme cases of non-proportional hazards with initially overlapping or crossing survival functions
Restricted mean survival time	Additive properties; usefulness even when median survival is not reached; applicable even in the extreme cases of non-proportional hazards with initially overlapping or crossing survival functions	Dependence on truncation time; interpretation may not be intuitive when curves do not reach zero; meaning to patient may be questionable
Combination tests	Have high power when shape of treatment effect over time is completely unknown	Interpretation of treatment effect when hazards are not proportional
Generalized pairwise comparisons	Allows different stakeholders to prioritize outcomes and thresholds of benefit; intuitively conveys probabilities on an absolute scale (net benefit) or relative scale (win ratio); allows simultaneous assessment of several endpoints, including safety	Recently proposed, with uncertain acceptability by regulatory agencies; potential for bias when average follow-up is much shorter than the longest event time; properties such as impact of censoring still under study; choice of priorities and clinical thresholds arbitrary

$$w(t_i) = \hat{S}(t_i)^\rho (1 - \hat{S}(t_i))^\gamma$$

where $\hat{S}(t_i)$ is an estimate of the overall survival function at time t_i and ρ and γ are shape parameters for the weight function. The unweighted logrank test is obtained for $\rho = \gamma = 0$ ($G^{0,0}$), the Peto–Prentice test for $\rho = 1$ and $\gamma = 0$ ($G^{1,0}$). The test gives more weight to later time points (and is thus preferable for delayed treatment effects) when $\rho = 0$ and $\gamma = 1$ ($G^{0,1}$) [81]. Zucker and Lakatos proposed weights that achieve maximum efficiency when there is a delayed treatment effect [82]. Yang and Prentice extended these ideas by using adaptive weights that ensure good power over a range of possible alternative hypotheses [83]. More recently, Magirr and Burman proposed “modestly weighted” logrank tests [84]. As an example, among many others, of delayed treatment benefit, the phase 3 trial KEYNOTE-40 investigated pembrolizumab versus chemotherapy in patients with recurrent or metastatic head and neck carcinoma [85]. In this trial, the standard logrank test of the difference in OS, the primary endpoint, produced only a marginally significant one-sided p value of 0.016. However, a delayed separation of the survival curves after 5 months of follow-up was observed. As the delayed treatment effect could have been reasonably expected based on the mechanisms of action of the experimental treatment and its competitor, a weighted logrank test might have been chosen in order to improve the power of the comparison [86].

A weighted logrank statistic may maximize statistical power, but the interpretation of the corresponding treatment effect is far from straightforward—in contrast to the hazard ratio that quantifies a reduction in the instantaneous risk of death at any time; the ease of interpretation of the treatment parameter in a proportional hazards model is no small reason for its enduring success, regardless of deviations from the underlying assumption [87].

4.4 Accelerated Failure Time Models

Interpretation of treatment effects on the hazard scale is not intuitive, as it is not straightforward to translate the information about the mortality hazard reduction, conveyed by the estimated value of HR, into a difference of the survival time. The latter scale is therefore more natural. Accelerated failure–time (AFT) models assume that the effect of treatment manifests itself in shrinking or extending the time scale. The model leads to a simple and natural interpretation of the treatment effect, which can be quantified in terms of the ratio of the mean survival time for the experimental and control treatment.

The AFT model is, essentially, a linear model on the logarithmic time scale, very similar to the classical linear regression model. Symbolically, the model can be expressed as follows:

$$\ln(t_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_k x_{k,i} + \varepsilon_i,$$

where t_i is the observed time to event for the i th patient, $x_{1,i}, \dots, x_{k,i}$ are the values of k explanatory variables describing the patient, and ε_i is the residual random error with mean equal to 0. If one assumes that ε_i is normally distributed, the AFT model

becomes the familiar linear regression model for the logarithm of survival time. Equivalently, the model can be expressed as follows:

$$t_i = e^{\varepsilon_i} e^{\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_k x_{k,i}}$$

According to this formulation, the effect of a unit change in variable x_j amounts to multiplying the time corresponding to the residual error, e^{ε_i} , by a factor equal to e^{β_j} . Hence, e^{β_j} can be naturally interpreted as the ratio of the mean survival time corresponding to $x_j + 1$ and x_j .

The AFT model does not require the proportional hazards assumption and can be used in the case of proportional and non-proportional hazards. Hence, it is an important alternative to the proportional hazards model. It has been observed that, for instance, the log-normal AFT model, in which residual error ε_i is assumed to be normally distributed and which is a non-proportional hazard model, is more suitable than the proportional hazard model for analyzing disease-free survival in colon cancer [88] or disease-free interval and disease-specific survival in breast cancer [89]. In glioblastoma, the log–logistic AFT model, in which residual error ε_i is assumed to have a logistic distribution and which is a non-proportional hazard model, was found to perform best with respect to prediction of patient survival time [90].

The main practical issue, often raised in the context of the use of the AFT model, is that the estimation of the model is usually carried out by assuming a parametric form of the distribution of the survival time, which in most cases is unknown. Parametric models can be used in situations where survival curves are smooth and can be approximated well with models with few parameters [91, 92]. Such approach is implemented in commercial statistical software such as SAS (PROC LIFEREG) and STATA (streg command) and in open-source software such as R (for instance, in function *survreg* in the *survival* package, and in function *psm* in the *rms* package). However, it is possible to estimate the model without the specification of the survival–time distribution. The semiparametric AFT model has been around since the end of 1970s [93, 94]. The main challenge, limiting the use of the model, was the lack of efficient and reliable computing algorithms. However, in the last decade, this has fundamentally changed. While the new computing algorithms have not yet been included in commercial statistical software, they are available in open-source software R [95]. These developments open the door to more widespread application of the semiparametric AFT model.

In the context of IT trials, it is worth noting that the AFT model, as the proportional hazards model, is not valid in the situation of a delayed treatment effect, when the survival functions for the experimental and control treatments initially overlap or cross. In that case, the use of the restricted mean survival time might be considered, as we discuss next.

4.5 Restricted Mean Survival Times

Ideally, one would prefer to express the treatment effect in terms of a difference in the mean survival time. If the survival curve reaches 0 (i.e., if the single longest observed time is an event), the mean survival can be estimated nonparametrically

by computing the area under the survival curve. However, this is almost never the case in practice. It is nevertheless possible to estimate the RMST by restricting (or truncating) the follow-up to a given time t and computing the area under the survival curve only up to that point [96–99]. The landmark time t can be chosen arbitrarily, but it is usually taken equal to the minimum of the largest times in all treatment groups. Once the restricted means in both groups are computed, they may be contrasted by subtraction. Importantly, the use and interpretation of the difference of RMST does not depend on whether hazards are proportional or not [100]. The RMST can be applied even in the extreme cases of non-proportional hazards, when the survival curves initially overlap or when they cross, as can be observed in IT trials [73]. The difference in RMST measures the mean gain in life expectancy through time t associated with the superior treatment. The interpretation of RMST may not be trivial, as it depends on the duration of follow-up, which dictates the choice of the landmark time t . Moreover, a mean survival time may not be meaningful to a patient, in so far as a month of survival gained in the near future (for a patient of poorer prognosis) may be quite different from a month of survival gained in a distant future (for a patient of better prognosis).

The difference in RMST can be tested for significance, and it is worth noting that the power of the test depends on the pattern of the difference and the chosen landmark time, among other factors [76, 101]. Hence, even in the situation of non-proportional hazards, the power of the test may not necessarily be larger than the power of the logrank test. Luo et al discuss the design and monitoring of trials using RMST [102]. Significance tests for RMST are available in the R software packages *survRM2* and *survRM2adapt*. Package *SSRMST* implements a method to compute sample size for a clinical trial with RMST used as an endpoint.

4.6 Combination Tests

Some authors have proposed to combine several tests in order to maximize power. For instance, if one has no idea whether the effect of treatment will be early or late, a combination test can use $Z = \max(|Z^{0,0}|, |Z^{1,0}|, |Z^{0,1}|)$, where $Z^{0,0}$, $Z^{1,0}$, and $Z^{0,1}$ are the statistics obtained from the $G^{0,0}$, $G^{1,0}$, and $G^{0,1}$ weighted logrank tests introduced above [103, 104]. Another combination test uses both weighted logrank tests and weighted Kaplan–Meier tests, which may be more sensitive than rank tests to differences in survival estimates [105, 106]. Yet another combination test uses a logrank test that would perform best under proportional hazards, and a permutation test of the difference in restricted mean survival times that might perform better in other situations [107]. These combination tests require small sample size increases as compared with the logrank, but they protect the power of the test against departures from proportional hazards [108]. Moreover, combination tests do not require pre-specification of a unique test (such as a weighted logrank test) which might or might not turn out to be appropriate for the situation at hand.

An example of this strategy is given by the analysis of the IM211 trial evaluating atezolizumab versus chemotherapy in patients with advanced or metastatic urothelial cancer. The comparison of OS in the PD-L1-positive population reported in the

publication showed a non-significant treatment effect. In this trial, survival curves cross between 4 and 5 months and show a numeric benefit in favor of atezolizumab in the long-term follow-up [109]. Roychoudhury et al. have evaluated the use of the MaxCombo test, a combined test based on multiple Fleming–Harrington weighted logrank tests used adaptively based on underlying data. The MaxCombo chose the $G^{0.1}$ with the minimum p value, and the test was highly statistically significant ($p = 0.005$). It strongly suggests that the use of this test strategy increased significantly the power of the comparison in this scenario of crossing survival [110].

4.7 Generalized Pairwise Comparisons (GPC) for Delayed Treatment Effects

Generalized pairwise comparisons (GPC) have recently been proposed to address situations of non-proportional hazards, in particular when the treatment effect is likely to manifest itself after some time. GPC extend the Wilcoxon–Mann–Whitney test to compare two samples, e.g., two randomized groups in a clinical trial. The outcome of interest is continuous and captured by a variable denoted X (taking values x_1, x_2, \dots, x_n) in the treatment group and denoted Y , taking values y_1, y_2, \dots, y_m in the control group. Consider all possible pairs (x_i, y_j) consisting of one observation from the treatment group and one observation from the control group. The U -statistic for the Wilcoxon–Mann–Whitney test is given by

$$U = \frac{1}{n \cdot m} \sum_{i=1}^n \sum_{j=1}^m u_{ij}$$

where

$$u_{ij} = \begin{cases} +1 & \text{if } x_i > y_j \\ -1 & \text{if } x_i < y_j \\ 0 & \text{if } x_i = y_j \end{cases}$$

The Wilcoxon–Mann–Whitney test was extended by Gehan to potentially censored outcomes. GPC generalize the test further to any situation in which every pair can be classified as a “win” (if the individual in the treated group has a better outcome than the individual in the control group), as a “loss” (if the individual in the treated group has a worse outcome than the individual in the control group), or as a “tie” (if there is no difference in outcome between the two individuals) [111, 112]. Hence, the U -statistic is now calculated using generalized pairwise scores:

$$u_{ij} = \begin{cases} +1 & \text{if pair is a win} \\ -1 & \text{if pair is a loss} \\ 0 & \text{if pair is a tie} \end{cases}$$

This generalized U -statistic, called the net benefit, is the difference between the probability of a win and the probability of a loss. The ratio between the probability of a win and the probability of a loss is called the win ratio [112].

In the analysis of times to event such as survival time, a win (loss) could be declared if the difference in survival exceeded a threshold considered to be clinically meaningful, say m [111, 113]. For treatments that have a short-lived effect, the net benefit will tend to decrease as a function of m , while for treatments that have a delayed effect, the net benefit will tend to remain stable or to increase as a function of m [114]. In fact, for treatments that achieve long-term cure in a given proportion of patients, the net benefit will tend to the cure rate.

The net benefit has been advocated as a patient-relevant measure of treatment benefit, because it is expressed on the time scale and directly answers a question a patient might ask, that is, “What is the net chance, for a patient taken at random, of surviving longer by at least m months on treatment than on control?” In addition, when the treatment benefit is delayed, the GPC test has increasing power when the threshold of clinical relevance increases [74]. Figure 1 shows average results from a large number of simulated trials in which survival in the control arm was assumed to follow an exponential distribution with parameter 0.1. The treatment arm was simulated for two distinct situations: one in which the hazard ratio remained equal to 0.65 over time (Fig. 1, Panel A), and the other in which the hazard ratio was equal to 1 until 4 months, then decreased linearly to 0.4 at 20 months and stayed at 0.4 thereafter, in such a way that the mean hazard ratio was also equal to 0.65 over the follow-up duration (Fig. 1, Panel B). Simulations were performed on complete times to event and also by setting a censoring mechanism corresponding on average to a proportion of 20% of censored observations. The censoring distribution was uniformly distributed, corresponding to an administrative censoring. The shape of the survival curves is not strikingly different between panels A and B, yet the net benefit as a function of time shows a clear difference between them and emphasizes the more substantial long-term net survival benefit from a treatment that has a delayed effect. The power of a GPC test with varying thresholds of clinical relevance is

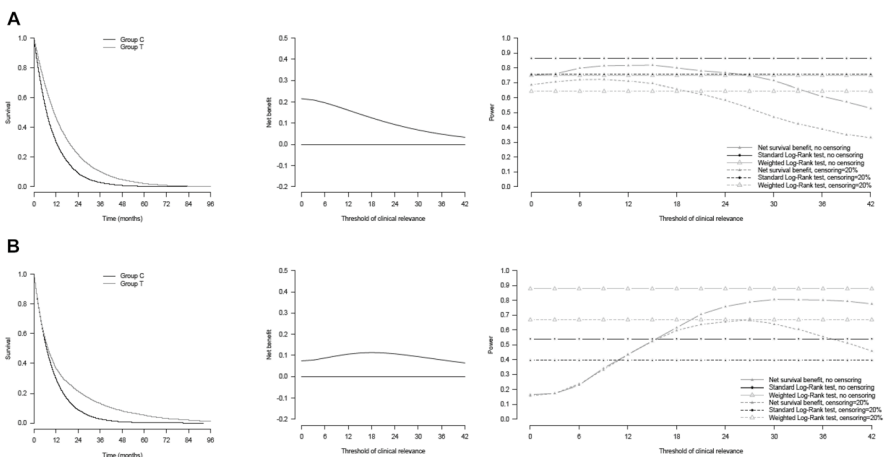


Fig. 1 Survival estimates as functions of time, net survival benefit, and power as functions of threshold of clinical relevance, in situations of proportional hazards (a) and delayed treatment effect (b)

inferior, as expected, to the power of the logrank test when hazards are proportional (Fig. 1, Panel A). However, the power of a GPC test becomes superior to the power of the logrank test, and similar to the power of a $G^{0,1}$ weighted logrank test, when the thresholds of clinical relevance are large (Fig. 1, Panel B).

Of note, however, censoring reduces the power of a GPC test for large thresholds of clinical relevance. This observation underscores the need for the follow-up time to be commensurate with the threshold of clinical relevance of interest.

The CA184-024 trial assessed the combination of ipilimumab plus dacarbazine versus placebo plus dacarbazine in patients with metastatic melanoma [66]. In this trial, PFS curves separated after the median, violating the proportional hazards assumption. When focusing on long-term PFS benefit, corresponding to higher values for the threshold of clinical relevance, the values of the net PFS benefit increased. The elevated and sustained value of the net PFS benefit, even for high threshold values, was a statistically testable measure of the delayed treatment effect [74].

4.8 Generalized Pairwise Comparisons (GPC) for Personalized Medicine

GPC can also prove useful to go beyond the analysis of a single outcome, by defining wins and losses for multiple outcomes. Hence, for instance, if time to death and time to disease progression were both of interest, one could use the composite endpoint of progression-free survival, which is the time to progressive disease or death, whichever comes first. A major objection against using such a composite endpoint is that it focuses on the time to first event, rather than on the time to most relevant endpoint. In other words, the crucially important time to death after progression is ignored. Using GPC, one can instead consider survival to have priority over time to progressive disease. Hence, if variables $\{X_j, Y_j\}$ denote the outcome of first priority (e.g., overall survival), respectively, in the treatment and control group, and $\{X_2, Y_2\}$ denote the outcome of second priority (e.g., time to progressive disease), respectively, in the treatment and control group, the pairwise scores can be generalized as follows (ignoring censoring for notational simplicity):

$$u_{ij} = \begin{cases} +1 & \text{if } X_{1,i} > Y_{1,j} \text{ or } (X_{1,i} = Y_{1,j} \text{ and } X_{2,i} > Y_{2,j}) \\ -1 & \text{if } X_{1,i} < Y_{1,j} \text{ or } (X_{1,i} = Y_{1,j} \text{ and } X_{2,i} < Y_{2,j}) \\ 0 & \text{otherwise} \end{cases}$$

The U -statistic captures the overall treatment effect on any number of prioritized outcomes of any type, including safety outcomes, quality of life, or other patient-relevant outcomes [115]. As such, this approach permits an overall benefit/risk assessment of the treatment effects using direct patient comparisons, rather than marginal treatment effects on the various outcomes considered that ignore the correlation between these outcomes [116]. In cancer, such a benefit/risk assessment is acutely required when treatments induce severe toxicities, some of which have a substantial impact on the patient well-being. Finally, because outcomes can be prioritized,

GPC can conceivably take into account individual patient preferences, thus paving the way to truly personalized medicine.

5 Conclusion

IT is not only revolutionizing the systemic treatment of patients with cancer, but also paving the way to the adoption of novel methodological approaches to trial design, analysis, and interpretation. In fairness, some of the methods now being adopted are not new, but their revival is largely due to the issues that have emerged in IT trials. IT trials have brought increased attention to the need to follow patients for as long as possible, an issue that was largely neglected in situations of proportional hazards. In fact, there is a strong incentive for the sponsor of a trial to terminate the follow-up as soon as a treatment effect is statistically established. This tendency should be actively resisted, and ensuring long-term follow-up of clinical trials should become the norm rather than the exception. The implementation of the methods discussed here in IT trials brings a fresh look to old problems, such as that of non-proportional hazards, or the possibility to tackle emerging questions in drug development and medical practice, such as that of prioritizing outcomes according to individual preferences. It is hoped that further improvements in the ability to deliver more efficacious, and hopefully less toxic, IT modalities to patients with cancer will be made more efficient by the use of improved statistical methodology.

References

1. Hoos A, Britten C (2012) The immuno-oncology framework: enabling a new era of cancer therapy. *Oncoimmunology* 1:334–339
2. Hoos A (2016) Development of immuno-oncology drugs—from CTLA4 to PD1 to the next generations. *Nat Rev Drug Discov* 15:235–247
3. Topalian SL, Weiner GJ, Pardoll DM (2011) Cancer immunotherapy comes of age. *J Clin Oncol* 29:4828–4836
4. Tsimberidou AM, Levit LA, Schilsky RL, Averbuch SD, Chen D, Kirkwood JM, McShane LM, Sharon E, Mileham KF, Postow MA (2019) Trial reporting in immuno-oncology (TRIO): an American Society of clinical oncology-society for immunotherapy of cancer statement. *J Clin Oncol* 37:72–80
5. Tang J, Shalabi A, Hubbard-Lucey VM (2018) Comprehensive analysis of the clinical immuno-oncology landscape. *Ann Oncol* 29:84–91
6. Anagnostou V, Yarchoan M, Hansen AR, Wang H, Verde F, Sharon E, Collyar D, Chow LQM, Forde PM (2017) Immuno-oncology trial endpoints: capturing clinically meaningful activity. *Clin Cancer Res* 23:4959–4969
7. Borcoman E, Kanjanapan Y, Champiat S, Kato S, Servois V, Kurzrock R, Goel S, Bedard P, Le Tourneau C (2019) Novel patterns of response under immunotherapy. *Ann Oncol* 30:385–396
8. Chen TT (2013) Statistical issues and challenges in immuno-oncology. *J Immunother Cancer* 1:18
9. Hales RK, Banchereau J, Ribas A, Tarhini AA, Weber JS, Fox BA, Drake CG (2010) Assessing oncologic benefit in clinical trials of immunotherapy agents. *Ann Oncol* 21:1944–1951
10. Hoos A, Eggermont AM, Janetzki S, Hodi FS, Ibrahim R, Anderson A, Humphrey R, Blumenstein B, Old L, Wolchok J (2010) Improved endpoints for cancer immunotherapy trials. *J Natl Cancer Inst* 102:1388–1397
11. Huang B (2018) Some statistical considerations in the clinical development of cancer immunotherapies. *Pharm Stat* 17:49–60

12. Finn OJ (2012) Immuno-oncology: understanding the function and dysfunction of the immune system in cancer. *Ann Oncol* 23(Suppl 8):viii6–viii9
13. Cogdill AP, Andrews MC, Wargo JA (2017) Hallmarks of response to immune checkpoint blockade. *Br J Cancer* 117:1–7
14. Daud AI, Loo K, Pauli ML, Sanchez-Rodriguez R, Sandoval PM, Taravati K, Tsai K, Nosrati A, Nardo L, Alvarado MD et al (2016) Tumor immune profiling predicts response to anti-PD-1 therapy in human melanoma. *J Clin Invest* 126:3447–3452
15. Mittal D, Gubin MM, Schreiber RD, Smyth MJ (2014) New insights into cancer immunoediting and its three component phases—elimination, equilibrium and escape. *Curr Opin Immunol* 27:16–25
16. Ritchie G, Gasper H, Man J, Lord S, Marschner I, Friedlander M, Lee CK (2018) Defining the most appropriate primary end point in phase 2 trials of immune checkpoint inhibitors for advanced solid cancers: a systematic review and meta-analysis. *JAMA Oncol* 4:522–528
17. Chiou VL, Burotto M (2015) Pseudoprogression and immune-related response in solid tumors. *J Clin Oncol* 33:3541–3543
18. Tazdait M, Mezquita L, Lahmar J, Ferrara R, Bidault F, Ammari S, Balleyguier C, Planchard D, Gazzah A, Soria JC et al (2018) Patterns of responses in metastatic NSCLC during PD-1 or PDL-1 inhibitor therapy: comparison of RECIST 1.1, irRECIST and iRECIST criteria. *Eur J Cancer* 88:38–47
19. Wolchok JD, Hoos A, O’Day S, Weber JS, Hamid O, Lebbe C, Maio M, Binder M, Bohnsack O, Nichol G et al (2009) Guidelines for the evaluation of immune therapy activity in solid tumors: immune-related response criteria. *Clin Cancer Res* 15:7412–7420
20. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJ, Robert L, Chmielowski B, Spasic M, Henry G, Ciobanu V et al (2014) PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* 515:568–571
21. Hodi FS, Ballinger M, Lyons B, Soria JC, Nishino M, Tabernero J, Powles T, Smith D, Hoos A, McKenna C et al (2018) Immune-modified response evaluation criteria in solid tumors (imRECIST): refining guidelines to assess the clinical benefit of cancer immunotherapy. *J Clin Oncol* 36:850–858
22. Hodi FS, Hwu WJ, Kefford R, Weber JS, Daud A, Hamid O, Patnaik A, Ribas A, Robert C, Gangadhar TC et al (2016) Evaluation of immune-related response criteria and RECIST v1.1 in patients with advanced melanoma treated with pembrolizumab. *J Clin Oncol* 34:1510–1517
23. Champiat S, Derclé L, Ammari S, Massard C, Hollebecque A, Postel-Vinay S, Chapat N, Eggermont A, Marabelle A, Soria JC et al (2017) Hyperprogressive disease is a new pattern of progression in cancer patients treated by anti-PD-1/PD-L1. *Clin Cancer Res* 23:1920–1928
24. Ferrara R, Mezquita L, Texier M, Lahmar J, Audigier-Valette C, Tessonnier L, Mazieres J, Zalcman G, Brosseau S, Le Moulec S et al (2018) Hyperprogressive disease in patients with advanced non-small cell lung cancer treated With PD-1/PD-L1 inhibitors or with single-agent chemotherapy. *JAMA Oncol* 4:1543–1552
25. Saada-Bouزيد E, Defaucheux C, Karabajakian A, Coloma VP, Servois V, Paoletti X, Even C, Fayette J, Guigay J, Loirat D et al (2017) Hyperprogression during anti-PD-1/PD-L1 therapy in patients with recurrent and/or metastatic head and neck squamous cell carcinoma. *Ann Oncol* 28:1605–1611
26. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancy J, Arbuck S, Gwyther S, Mooney M et al (2009) New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 45:228–247
27. WHO (1979) WHO handbook for reporting results of cancer treatment. World Health Organization Offset Publication No. 48, Geneva
28. Nishino M, Giobbie-Hurder A, Gargano M, Suda M, Ramaiya NH, Hodi FS (2013) Developing a common language for tumor response to immunotherapy: immune-related response criteria using unidimensional measurements. *Clin Cancer Res* 19:3936–3943
29. Seymour L, Bogaerts J, Perrone A, Ford R, Schwartz LH, Mandrekar S, Lin NU, Litiere S, Dancy J, Chen A et al (2017) iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol* 18:e143–e152
30. Chmielowski B (2018) How should we assess benefit in patients receiving checkpoint inhibitor therapy? *J Clin Oncol* 36:835–836
31. Kazandjian D, Keegan P, Suzman DL, Pazdur R, Blumenthal GM (2017) Characterization of outcomes in patients with metastatic non-small cell lung cancer treated with programmed cell death

- protein 1 inhibitors past RECIST version 1.1-defined disease progression in clinical trials. *Semin Oncol* 44:3–7
32. Escudier B, Motzer RJ, Sharma P, Wagstaff J, Plimack ER, Hammers HJ, Donskov F, Gurney H, Sosman JA, Zaleski PG et al (2017) Treatment beyond progression in patients with advanced renal cell carcinoma treated with nivolumab in CheckMate 025. *Eur Urol* 72:368–376
 33. George S, Motzer RJ, Hammers HJ, Redman BG, Kuzel TM, Tykodi SS, Plimack ER, Jiang J, Waxman IM, Rini BI (2016) Safety and efficacy of nivolumab in patients with metastatic renal cell carcinoma treated beyond progression: a subgroup analysis of a randomized clinical trial. *JAMA Oncol* 2:1179–1186
 34. Long GV, Weber JS, Larkin J, Atkinson V, Grob JJ, Schadendorf D, Dummer R, Robert C, Marquez-Rodas I, McNeil C et al (2017) Nivolumab for patients with advanced melanoma treated beyond progression: analysis of 2 phase 3 clinical trials. *JAMA Oncol* 3:1511–1519
 35. Cottrell TR, Thompson ED, Forde PM, Stein JE, Duffield AS, Anagnostou V, Rekhman N, Anders RA, Cuda JD, Illei PB et al (2018) Pathologic features of response to neoadjuvant anti-PD-1 in resected non-small-cell lung carcinoma: a proposal for quantitative immune-related pathologic response criteria (irPRC). *Ann Oncol* 29:1853–1860
 36. Stein JE, Soni A, Danilova L, Cottrell TR, Gajewski TF, Hodi FS, Bhatia S, Urba WJ, Sharfman WH, Wind-Rotolo M et al (2019) Major pathologic response on biopsy (MPRbx) in patients with advanced melanoma treated with anti-PD-1: evidence for an early, on-therapy biomarker of response. *Ann Oncol* 30:589–596
 37. Beaver JA, Howie LJ, Pelosof L, Kim T, Liu J, Goldberg KB, Sridhara R, Blumenthal GM, Farrell AT, Keegan P et al (2018) A 25-year experience of US Food and Drug Administration accelerated approval of malignant hematology and oncology drugs and biologics: a review. *JAMA Oncol* 4:849–856
 38. Gettinger S, Horn L, Jackman D, Spigel D, Antonia S, Hellmann M, Powderly J, Heist R, Sequist LV, Smith DC et al (2018) Five-year follow-up of nivolumab in previously treated advanced non-small-cell lung cancer: results from the CA209-003 study. *J Clin Oncol* 36:1675–1684
 39. Maio M, Grob JJ, Aamdal S, Bondarenko I, Robert C, Thomas L, Garbe C, Chiarion-Sileni V, Testori A, Chen TT et al (2015) Five-year survival rates for treatment-naïve patients with advanced melanoma who received ipilimumab plus dacarbazine in a phase III trial. *J Clin Oncol* 33:1191–1196
 40. Topalian SL, Sznol M, McDermott DF, Kluger HM, Carvajal RD, Sharfman WH, Brahmer JR, Lawrence DP, Atkins MB, Powderly JD et al (2014) Survival, durable tumor remission, and long-term safety in patients with advanced melanoma receiving nivolumab. *J Clin Oncol* 32:1020–1030
 41. Attia P, Phan GQ, Maker AV, Robinson MR, Quezado MM, Yang JC, Sherry RM, Topalian SL, Kammula US, Royal RE et al (2005) Autoimmunity correlates with tumor regression in patients with metastatic melanoma treated with anti-cytotoxic T-lymphocyte antigen-4. *J Clin Oncol* 23:6043–6053
 42. Maude SL, Frey N, Shaw PA, Aplenc R, Barrett DM, Bunin NJ, Chew A, Gonzalez VE, Zheng Z, Lacey SF et al (2014) Chimeric antigen receptor T cells for sustained remissions in leukemia. *N Engl J Med* 371:1507–1517
 43. Blumenthal GM, Pazdur R (2016) Response rate as an approval end point in oncology: back to the future. *JAMA Oncol* 2:780–781
 44. Andtbacka RH, Kaufman HL, Collichio F, Amatruda T, Senzer N, Chesney J, Delman KA, Spitler LE, Puzanov I, Agarwala SS et al (2015) Talimogene laherparepvec improves durable response rate in patients with advanced melanoma. *J Clin Oncol* 33:2780–2788
 45. Morgan TM (1988) Analysis of duration of response: a problem of oncology trials. *Control Clin Trials* 9:11–18
 46. Ellis S, Carroll KJ, Pemberton K (2008) Analysis of duration of response in oncology trials. *Contemp Clin Trials* 29:456–465
 47. Huang B, Tian L, Talukder E, Rothenberg M, Kim DH, Wei LJ (2018) Evaluating treatment effect based on duration of response for a comparative oncology study. *JAMA Oncol* 4:874–876
 48. Korn EL, Othus M, Chen T, Freidlin B (2017) Assessing treatment efficacy in the subset of responders in a randomized clinical trial. *Ann Oncol* 28:1640–1647
 49. Blumenthal GM, Karuri SW, Zhang H, Zhang L, Khozin S, Kazandjian D, Tang S, Sridhara R, Keegan P, Pazdur R (2015) Overall response rate, progression-free survival, and overall survival with targeted and standard therapies in advanced non-small-cell lung cancer: US Food and Drug Administration trial-level and patient-level analyses. *J Clin Oncol* 33:1008–1014

50. Burzykowski T, Buyse M, Piccart-Gebhart MJ, Sledge G, Carmichael J, Luck HJ, Mackey JR, Nabholz JM, Paridaens R, Biganzoli L et al (2008) Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer. *J Clin Oncol* 26:1987–1992
51. Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P (2000) Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. *Meta-analysis Group in Cancer. Lancet* 356:373–378
52. Kaufman HL, Schwartz LH, William WN Jr, Sznol M, Fahrback K, Xu Y, Masson E, Vergara-Silva A (2018) Evaluation of classical clinical endpoints as surrogates for overall survival in patients treated with immune checkpoint blockers: a systematic review and meta-analysis. *J Cancer Res Clin Oncol* 144:2245–2261
53. Roviello G, Andre F, Venturini S, Pistilli B, Curigliano G, Cristofanilli M, Rosellini P, Generali D (2017) Response rate as a potential surrogate for survival and efficacy in patients treated with novel immune checkpoint inhibitors: a meta-regression of randomised prospective studies. *Eur J Cancer* 86:257–265
54. Hamid O, Robert C, Daud A, Hodi FS, Hwu WJ, Kefford R, Wolchok JD, Hersey P, Joseph R, Weber JS et al (2019) Five-year survival outcomes for patients with advanced melanoma treated with pembrolizumab in KEYNOTE-001. *Ann Oncol* 30:501–503
55. Hodi FS, O’Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, Gonzalez R, Robert C, Schadendorf D, Hassel JC et al (2010) Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med* 363:711–723
56. Mick R, Chen TT (2015) Statistical challenges in the design of late-stage cancer immunotherapy studies. *Cancer Immunol Res* 3:1292–1298
57. Saad ED, Buyse M (2016) Statistical controversies in clinical research: end points other than overall survival are vital for regulatory approval of anticancer agents. *Ann Oncol* 27:373–378
58. Kantoff PW, Higano CS, Shore ND, Berger ER, Small EJ, Penson DF, Redfern CH, Ferrari AC, Dreicer R, Sims RB et al (2010) Sipuleucel-T immunotherapy for castration-resistant prostate cancer. *N Engl J Med* 363:411–422
59. Bellmunt J, de Wit R, Vaughn DJ, Fradet Y, Lee JL, Fong L, Vogelzang NJ, Climent MA, Petrylak DP, Choueiri TK et al (2017) Pembrolizumab as second-line therapy for advanced urothelial carcinoma. *N Engl J Med* 376:1015–1026
60. Borghaei H, Paz-Ares L, Horn L, Spigel DR, Steins M, Ready NE, Chow LQ, Vokes EE, Felip E, Holgado E et al (2015) Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *N Engl J Med* 373:1627–1639
61. Ferris RL, Blumenschein G Jr, Fayette J, Guigay J, Colevas AD, Licitra L, Harrington K, Kasper S, Vokes EE, Even C et al (2016) Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *N Engl J Med* 375:1856–1867
62. Herbst RS, Baas P, Kim DW, Felip E, Perez-Gracia JL, Han JY, Molina J, Kim JH, Arvis CD, Ahn MJ et al (2016) Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet* 387:1540–1550
63. Motzer RJ, Escudier B, McDermott DF, George S, Hammers HJ, Srinivas S, Tykodi SS, Sosman JA, Procopio G, Plimack ER et al (2015) Nivolumab versus everolimus in advanced renal-cell carcinoma. *N Engl J Med* 373:1803–1813
64. Rittmeyer A, Barlesi F, Waterkamp D, Park K, Ciardiello F, von Pawel J, Gadgeel SM, Hida T, Kowalski DM, Dols MC et al (2017) Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet* 389:255–265
65. Petrelli F, Coimu A, Cabiddu M, Borgonovo K, Ghilardi M, Lonati V, Barni S (2016) Early analysis of surrogate endpoints for metastatic melanoma in immune checkpoint inhibitor trials. *Medicine (Baltimore)* 95:e3997
66. Robert C, Thomas L, Bondarenko I, O’Day S, Weber J, Garbe C, Lebbe C, Baurain JF, Testori A, Grob JJ et al (2011) Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. *N Engl J Med* 364:2517–2526
67. Robert C, Long GV, Brady B, Dutriaux C, Maio M, Mortier L, Hassel JC, Rutkowski P, McNeil C, Kalinka-Warzocho E et al (2015) Nivolumab in previously untreated melanoma without BRAF mutation. *N Engl J Med* 372:320–330

68. Reck M, Rodriguez-Abreu D, Robinson AG, Hui R, Czoszi T, Fulop A, Gottfried M, Peled N, Tafreshi A, Cuffe S et al (2016) Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. *N Engl J Med* 375:1823–1833
69. Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, Sunpaweravong P, Han B, Margono B, Ichinose Y et al (2009) Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* 361:947–957
70. Chen TT (2015) Milestone survival: a potential intermediate endpoint for immune checkpoint inhibitors. *J Natl Cancer Inst* 107:156
71. Korn EL, Freidlin B (2018) Interim futility monitoring assessing immune therapies with a potentially delayed treatment effect. *J Clin Oncol* 36:2444–2449
72. Liang F, Zhang S, Wang Q, Li W (2018) Treatment effects measured by restricted mean survival time in trials of immune checkpoint inhibitors for cancer. *Ann Oncol* 29:1320–1324
73. Pak K, Uno H, Kim DH, Tian L, Kane RC, Takeuchi M, Fu H, Claggett B, Wei LJ (2017) Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. *JAMA Oncol* 3:1692–1696
74. Peron J, Lambert A, Munier S, Ozenne B, Giai J, Roy P, Dalle S, Machingura A, Maucort-Boulch D, Buyse M (2019) Assessing long-term survival benefits of immune checkpoint inhibitors using the net survival benefit. *J Natl Cancer Inst* 111:1186–1191
75. Hoering A, Durie B, Wang H, Crowley J (2017) End points and statistical considerations in immuno-oncology trials: impact on multiple myeloma. *Fut Oncol* 13:1181–1193
76. Huang B, Kuan PF (2018) Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point. *Pharm Stat* 17:202–213
77. Xu Z, Zhen B, Park Y, Zhu B (2017) Designing therapeutic cancer vaccine trials with delayed treatment effect. *Stat Med* 36:592–605
78. Rahman R, Fell G, Trippa L, Alexander BM (2018) Violations of the proportional hazards assumption in randomized phase III oncology clinical trials. *J Clin Oncol* 36 (15 Suppl):abstract 2543
79. Lin NX, Logan S, Henley WE (2013) Bias and sensitivity analysis when estimating treatment effects from the cox model with omitted covariates. *Biometrics* 69:850–860
80. Harrington DP, Fleming TR (1982) A class of rank test procedures for censored survival data. *Biometrika* 69:133–143
81. Lin RS, Leon LF (2017) Estimation of treatment effects in weighted log-rank tests. *Contemp Clin Trials Commun* 8:147–155
82. Zucker M, Lakatos E (1990) Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* 77:853–864
83. Yang S, Prentice R (2010) Improved logrank-type tests for survival data using adaptive weights. *Biometrics* 66:30–38
84. Magirr D, Burman CF (2019) Modestly weighted logrank tests. *Stat Med* 38:3782–3790
85. Cohen EEW, Soulieres D, Le Tourneau C, Dinis J, Licitra L, Ahn MJ, Soria A, Machiels JP, Mach N, Mehra R et al (2019) Pembrolizumab versus methotrexate, docetaxel, or cetuximab for recurrent or metastatic head-and-neck squamous cell carcinoma (KEYNOTE-040): a randomised, open-label, phase 3 study. *Lancet* 393:156–167
86. Su Z, Zhu M (2018) Is it time for the weighted log-rank test to play a more important role in confirmatory trials? *Contemp Clin Trials Commun* 10:A1–A2
87. Freidlin B, Korn EL (2019) Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis? *J Clin Oncol* 37:3455–3459
88. Chapman JW, O’Callaghan CJ, Hu N, Ding K, Yothers GA, Catalano PJ, Shi Q, Gray RG, O’Connell MJ, Sargent DJ (2013) Innovative estimation of survival using log-normal survival modelling on ACCENT database. *Br J Cancer* 108:784–790
89. Chapman JA, Lickley HL, Trudeau ME, Hanna WM, Kahn HJ, Murray D, Sawka CA, Mobbs BG, McCready DR, Pritchard KI (2006) Ascertain prognosis for breast cancer in node-negative patients with innovative survival analysis. *Breast J* 12:37–47
90. Senders JT, Staples P, Mehrtash A, Cote DJ, Taphoorn MJB, Reardon DA, Gormley WB, Smith TR, Broekman ML, Arnaout O (2019) An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning. *Neurosurgery* 86:184–192
91. Anderson KM (1991) A nonproportional hazards Weibull accelerated failure time regression model. *Biometrics* 47:281–288
92. Odell PM, Anderson KM, Kannel WB (1994) New models for predicting cardiovascular events. *J Clin Epidemiol* 47:583–592

93. Buckley J, James I (1979) Linear regression with censored data. *Biometrika* 66:429–436
94. Prentice RL (1978) Linear rank tests with censored data. *Biometrika* 65:167–179
95. Chiou SH, Kang S, Yan J (2014) Fitting accelerated failure time models in routine survival analysis with R package aftgee. *J Stat Softw* 61:1–23
96. Royston P, Parmar MK (2011) The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med* 30:2409–2421
97. Royston P, Parmar MK (2013) Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol* 13:152
98. Seruga B, Pond GR, Hertz PC, Amir E, Ocana A, Tannock IF (2012) Comparison of absolute benefits of anticancer therapies determined by snapshot and area methods. *Ann Oncol* 23:2977–2982
99. Trinquart L, Jacot J, Conner SC, Porcher R (2016) Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *J Clin Oncol* 34:1813–1819
100. A'Hern RP (2016) Restricted mean survival time: an obligatory end point for time-to-event analysis in cancer trials? *J Clin Oncol* 34:3474–3476
101. Tian L, Fu H, Ruberg SJ, Uno H, Wei LJ (2018) Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics* 74:694–702
102. Luo X, Huang B, Quan H (2019) Design and monitoring of survival trials based on restricted mean survival times. *Clin Trials* 16:616–625
103. Karrison T (2016) Versatile tests for comparing survival curves based on weighted log-rank statistics. *Stata J* 16:678–690
104. Lee JW (1996) Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics* 52:721–725
105. Chi Y, Tsai MH (2001) Some versatile tests based on the simultaneous use of weighted logrank and weighted Kaplan-Meier statistics. *Commun Stat Simulat* 30:743–759
106. Pepe MS, Fleming TR (1989) Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics* 45:497–507
107. Royston P, Parmar MK (2016) Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Med Res Methodol* 16:16
108. Royston P, Choodari-Oskooei B, Parmar MKB, Rogers JK (2019) Combined test versus logrank/Cox test in 50 randomised trials. *Trials* 20:172
109. Powles T, Duran I, van der Heijden MS, Loriot Y, Vogelzang NJ, De Giorgi U, Oudard S, Retz MM, Castellano D, Bamias A et al (2018) Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (IMvigor211): a multicentre, open-label, phase 3 randomised controlled trial. *Lancet* 391:748–757
110. Roychoudhury S, Anderson KM, Ye J, Mukhopadhyay P (2019) Robust design and analysis of clinical trials with non-proportional hazards: a straw man guidance from a cross-pharma Working Group. <https://arxiv.org/abs/1908.07112>
111. Buyse M (2010) Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med* 29:3245–3257
112. Pocock SJ, Ariti CA, Collier TJ, Wang D (2012) The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J* 33:176–182
113. Peron J, Buyse M, Ozenne B, Roche L, Roy P (2018) An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. *Stat Methods Med Res* 27:1230–1239
114. Peron J, Roy P, Ozenne B, Roche L, Buyse M (2016) The net chance of a longer survival as a patient-oriented measure of treatment benefit in randomized clinical trials. *JAMA Oncol* 2:901–905
115. Buyse M (2019) Multiple prioritized outcomes. *Wiley StatsRef: Statistics Reference Online*. <https://doi.org/10.1002/9781118445112.stat08158>
116. Evans SR, Follmann D (2016) Using outcomes to analyze patients rather than patients to analyze outcomes: a step toward pragmatism in benefit: risk evaluation. *Stat Biopharm Res* 8:386–393